UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Knowledge-based and data-driven approaches for geographical information access

Daniel Ferrés Domènech

Knowledge-Based and Data-Driven Approaches for Geographical Information Access

Ph.D. Thesis

This thesis is presented for the degree of Doctor
of the Universitat Politècnica de Catalunya
by

Daniel Ferrés Domènech

advisor

Dr. Horacio Rodríguez Hontoria

Ph.D. Program in Artificial Intelligence
Department of Computer Science
Universitat Politècnica de Catalunya

September 2017

Abstract

Geographical Information Access (GeoIA) can be defined as a way of retrieving information from textual collections that includes the automatic analysis and interpretation of the geographical constraints and terms present in queries and documents. This PhD thesis presents, describes and evaluates several heterogeneous approaches for the following three GeoIA tasks: Geographical Information Retrieval (GIR), Geographical Question Answering (GeoQA), and Textual Georeferencing (TG). The GIR task deals with user queries that search over documents (e.g. “vineyards in California”) and the GeoQA task treats questions that retrieve answers (e.g. “What is the capital of France?”). On the other hand, TG is the task of associate one or more georeferences (such as polygons or coordinates in a geodetic reference system) to electronic documents.

Current state-of-the-art AI algorithms are not yet fully understanding the semantic meaning and the geographical constraints and terms present in queries and document collections. This thesis attempts to improve the effectiveness results of GeoIA tasks by: 1) improving the detection, understanding, and use of a part of the geographical and the thematic content of queries and documents with Toponym Recognition, Toponym Disambiguation and Natural Language Processing (NLP) techniques, and 2) combining Geographical Knowledge-Based Heuristics based on common sense with Data-Driven IR algorithms.

The main contributions of this thesis to the state-of-the-art of GeoIA tasks are:

1) The presentation of 10 novel approaches for GeoIA tasks: 3 approaches for GIR, 3 for GeoQA, and 4 for Textual Georeferencing (TG).

2) The evaluation of these novel approaches in these contexts: within official evaluation benchmarks, after evaluation benchmarks with the test collections, and with other specific datasets. Most of these algorithms have been evaluated in international evaluations and some of them achieved top-ranked state-of-the-art results, including top-performing results in GIR (GeoCLEF 2007) and TG (MediaEval 2014) benchmarks.

3) The experiments reported in this PhD thesis show that the approaches can combine effectively Geographical Knowledge and NLP with Data-Driven techniques to improve the efectiveness measures of the three Geographical Information Access tasks investigated.

4) TALPGeoIR: a novel GIR approach that combines Geographical Knowledge Re-Ranking (GeoKR), NLP and Relevance Feedback (RF) that achieved state-of-the-art results in official GeoCLEF benchmarks (Ferrés and Rodríguez, 2008a; Mandl et al., 2008) and posterior experiments (Ferrés and Rodríguez, 2015a). This approach has been evaluated with the full GeoCLEF corpus (100 topics) and showed that GeoKR, NLP, and RF techniques evaluated separately or in combination improve the results in MAP and R-Precision effectiveness measures of the state-of-the-art IR algorithms TF-IDF, BM25 and InL2 and show statistical significance in most of the experiments.

5) GeoTALP-QA: a scope-based GeoQA approach for Spanish and English and its evaluation with a set of questions of the Spanish geography (Ferrés and Rodríguez, 2006a).

6) Four state-of-the-art Textual Georeferencing approaches for informal and formal documents that achieved state-of-the-art results in evaluation benchmarks (Ferrés and Rodríguez, 2014) and posterior experiments (Ferrés and Rodríguez, 2011a; Ferrés and Rodríguez, 2015b).

Resum

L'Accés a la Informació Geogràfica (GeoAI) pot ser definit com una forma de recuperar informació de col·lecions textuals que inclou l'anàlisi automàtic i la interpretació dels termes i restriccions geogràfiques que apareixen en consultes i documents. Aquesta tesi doctoral presenta, descriu i avalua varies aproximacions heterogènies a les seguents tasques de GeoAI: Recuperació de la Informació Geogràfica (RIG), Cerca de la Resposta Geogràfica (GeoCR), i Georeferenciament Textual (GT). La tasca de RIG tracta amb consultes d'usuari que cerquen documents (e.g. "vinyes a California") i la tasca GeoCR tracta de recuperar respostes concretes a preguntes (e.g. "Quina és la capital de França"). D'altra banda, GT es la tasca de relacionar una o més referències geogràfiques (com polígons o coordenades en un sistema de referència geodètic) a documents electrònics. Els algoritmes de l'estat de l'art actual en Intel·ligència Artificial encara no comprenen completament el significat semàntic i els termes i les restriccions geogràfiques presents en consultes i col·leccions de documents.

Aquesta tesi intenta millorar els resultats en efectivitat de les tasques de GeoAI de la seguent manera: 1) millorant la detecció, comprensió, i l'utilització d'una part del contingut geogràfic i temàtic de les consultes i documents amb tècniques de reconeixement de topònims, desambiguació de topònims, i Processament del Llenguatge Natural (PLN), i 2) combinant heuristics basats en Coneixement Geogràfic i en el sentit comú humà amb algoritmes de Recuperació de la Informació basats en dades.

Les principals contribucions d'aquesta tesi a l'estat de l'art de les tasques de GeoAI són:

1) La presentació de 10 noves aproximacions a les tasques de GeoAI: 3 aproximacions per RIG, 3 per GeoCR, i 4 per Georeferenciament Textual (GT).

2) L'avaluació d'aquestes noves aproximacions en aquests contexts: en el marc d'avaluacions comparatives internacionals, posteriorment a avaluacions comparatives internacionals amb les col·lections de test, i amb altres conjunts de dades específics. La majoria d'aquests algoritmes han estat avaluats en avaluacions comparatives internacionals i alguns d'ells aconseguiren alguns dels millors resultats en l'estat de l'art, com per exemple els resultats en comparatives de RIG (GeoCLEF 2007) i GT (MediaEval 2014).

3) Els experiments descrits en aquesta tesi mostren que les aproximacions poden combinar coneixement geogràfic i PLN amb tècniques basades en dades per millorar les mesures d'efectivitat en les tres tasques de l'Accés a la Informació Geogràfica investigades.

4) TALPGeoIR: una nova aproximació a la RIG que combina Re-Ranking amb Coneixement Geogràfic (GeoKR), PLN i Retroalimentació de Rellevancia (RR) que aconseguí resultats en l'estat de l'art en comparatives oficials GeoCLEF (Ferrés and Rodríguez, 2008a; Mandl et al., 2008) i en experiments posteriors (Ferrés and Rodríguez, 2015a). Aquesta aproximació ha estat avaluada amb el conjunt complert del corpus GeoCLEF (100 topics) i ha mostrat que les tècniques GeoKR, PLN i RR avaluades separadament o en combinació milloren els resultats en les mesures efectivitat MAP i R-Precision dels algoritmes de l'estat de l'art en Recuperació de la Infomació TF-IDF, BM25 i InL2 i a més mostren significació estadística en la majoria dels experiments.

5) GeoTALP-QA: una aproximació a GeoCR per espanyol i anglès i la seva avaluació amb un conjunt de preguntes de la geografía espanyola (Ferrés and Rodríguez, 2006a).

6) Quatre aproximacions per al georeferenciament de documents formals i informals que obtingueren resultats en l'estat de l'art en avaluacions comparatives internacionals (Ferrés and Rodríguez, 2014) i en experiments posteriors (Ferrés and Rodríguez, 2011a; Ferrés and Rodríguez, 2015b).

Acknowledgments

First of all I want to thank my PhD thesis advisor Dr. Horacio Rodríguez for his advicement and guidance. I want to thank my colleagues and ex-colleagues at UPC from the Natural Language Processing Group and the TALP Research Center for their help. I want to thank the coordination and management team of the Ph.D. program in Artificial Intelligence at UPC and the anonymous thesis reviewers. I want to thank the professors that helped my during my research stages abroad:

- Dr. Eduard Hovy (ISI-USC).
- Dr. Flavia Barros (CIn-UFPE).

I also want to thank the help of the administration and technical staff of the UDGSI unit of UPC.

This research has been realized in the framework of the following grants and research projects:

- grants: UPC-Recerca, BE (AGAUR), EBW II Project (European Comission).
- research projects (Spanish and European): ALIADO (TIC2002-04447-C02), TEXT-MESS (TIN2006-15265-C06-05), KNOW 2 (TIN2009-14715-C04-04), SKATER (TIN2012-38584-C06-01), and CHIL (IST-2004-506909).

Finally I would like to thank my family, friends, and all the other people that helped me.

Agraïments

Primerament vull agrair al meu director de tesi Horacio Rodríguez la seva tutoria i suport. Vull agrair també a tots els meus companys i ex-companys del Grup de Processament del Llenguatge Natural i del centre de recerca TALP de la Universitat Politècnica de Catalunya per la seva ajuda i suport. Vull donar gràcies a l'equip de coordinació i gestió del programa de doctorat en Intel·ligència Artificial de la UPC i als revisors anònims de la tesi.

Estic molt agrait als professors que m'ajudaren durant les meves estades a l'estranger:

- Dr. Eduard Hovy (ISI-USC).
- Dr. Flavia Barros (CIn-UFPE).

També vull agrair la ajuda i suport del l'administració i del laboratori de càlcul de la Unitat de Gestió dels Departaments CS i ESSI.

El treball de tesi ha estat realitzat en el marc de les seguents beques i projectes de recerca:

- beques: UPC-Recerca, BE (AGAUR), EBW II Project (European Comission).
- projectes de recerca españols i europeus: ALIADO (TIC2002-04447-C02), TEXT-MESS (TIN2006-15265-C06-05), KNOW 2 (TIN2009-14715-C04-04), SKATER (TIN2012-38584-C06-01) i CHIL, IST-2004-506909.

Finalment, vull donar gràcies a la meva família, amics i a totes les altres persones que m'han ajudat.

Contents

CHAPTER 1

Introduction

Nowadays, there is a growing need for humans for more intelligent accessing and searching the electronic textual information stored in computers, locally (PCs, laptops, tablets, cell phones,...) or remotely (e.g. Internet, Local area networks (LANs), servers,...). In this context, academic and industrial efforts try to research and develop new techniques that facilitate this intelligent access to the information. Most of this work is classified under the field of Artificial Intelligence (AI).

This PhD thesis presents geographically aware methods to access to the electronic textual information by natural language (i.e. user defined keywords and query searching) using AI techniques. The research presented here involves the use of several concepts and techniques developed in the AI sub-fields of Natural Language Processing (NLP), Information Retrieval (IR) and Knowledge Bases (KB) and its application to some Geographical Information Access tasks. Geographical Information Access (GeoIA) can be defined as a way of retrieve information through the automatic analysis of queries that include geographical constraints.

The GeoIA tasks investigated in this dissertation are the following: Geographical Information Retrieval (GIR), Geographical Question Answering (GeoQA), and Textual Georeferencing (TG). IR and Question Answering techniques can be defined as algorithms that help the user to satisfy an information need. Question Answering (QA) is the task of, given a question expressed in Natural Language (NL), retrieving its correct answer (a single item, a text snippet, a list of items,...) from closed collections or the Web. This task is considered a step beyond IR which consists in searching information in documents, documents themselves, or metadata which describe documents (Baeza-Yates and Ribeiro-Neto, 1999). Therefore, GIR and GeoQA can be defined as algorithms that help the user to satisfy an information need that includes a geographical restriction. GIR deals with user queries that search over documents (e.g. “vineyards in California”), and GeoQA treats questions that retrieve answers (e.g. ”What is the capital of France?). On the other hand, TG is the task of associating one or more georeferences (such as polygons or unique coordinates in the World Geodetic System[1] (WGS)) to an electronic document or text. As an example, the sentence "Gustave Eiffel built the Eiffel Tower in Paris in 1889" can undoubtfully be associated to the location of Paris in France (and even with the Eiffel Tower monument) with precise WSG84[2] coordinates or bounding boxes. TG can be applied to improve both GIR and GQA tasks.

In the last years has emerged a growing community of researchers that explore Information Access tasks on Restricted-Domains (RDs), including the Geographical Domain. The geographical domain has been investigated in several GIR workshops and evaluation benchmarks such as GIR international workshops[3] since 2004, GeoCLEF at CLEF workshops from 2005 to 2008 (F. Gey et al., 2005; F. Gey et al., 2006; Mandl et al., 2008; Mandl et al., 2008), GeoQuery at CLEF 2007 (Z. Li et al., 2007b), GikiCLEF at CLEF 2008 and 2009 (D. Santos et al., 2008; D. Santos and L. M. Cabral, 2009; D. Santos and L. Cabral, 2010), GeoTime in 2010 and 2011 (F. C. Gey et al., 2010; F. C. Gey et al., 2011), and Placing Task at MediaEval workshops from 2010 to 2016 (Choi et al., 2014; M. Larson et al., 2015; Choi et al., 2015; Choi et al., 2016). Building RDs applications implies the need of more precision and the use of specific knowledge of the domain (e.g. lexicons, dictionaries, corpora, axioms, etc.). Data-driven methods based on exploiting redundancy in big data collections are not always useful in these contexts. It must also be taken into consideration that the geographical domain can be considered a special case of RDs because many open domain texts contain a high density of geographical terms (Benamara, 2004).

This thesis investigates Geographical Information Access techniques for the GIR, GeoQA and TG tasks and proposes and evaluates several approaches to deal with these GeoIA tasks. Current state-of-the-art techniques for GeoIA use generally Data-Driven or Knowledge-Based approaches based on Geographical Knowledge Bases. This thesis applied these two major approaches and the combination of both.

1.1 Geographical Information Access

Information Access (IA) is an area of research concerned with technologies that satisfy user's information needs about the information contained in electronic text collections (or single documents). According to Gaussier and Yvon (2012) the following applications facilitate information access: information extraction and retrieval; text classification and clustering; opinion mining; comprehension aids (automatic summarization, machine translation, visualization). Currently Information Retrieval is the dominant form of information access (Manning et al., 2008), but in recent years Question Answering emerged as a new research task for Information Access. Information Access tasks deal with the problem of Natural Language Understanding; that means to interpret the semantic meaning of the texts. Whereas Geographical Information Access not only deals with the semantic meaning but also with the geographical meaning of the texts. Geographic terms are commonly used in Information Access applications such as users queries to the web: 1) in 2004 a study by Sanderson and Kohler (2004a) over a random sample of 2,500 queries of the 2001 Excite query log showed that a 18,6% of the queries contained a geographic term and 14.8% contained a place name, 2) according to Asadi et al. (2005) it was estimated that 22 percent of

[1] The World Geodetic System is a standard for use in cartography, geodesy, and navigation.

[2] WSG84 is the latest revision of the World Geodetic System. WGS84 is an Earth-centered, Earth-fixed terrestrial reference system and geodetic datum that was established in 1984.

[3] `http://www.geo.uzh.ch/~rsp/gir14/index.html`

web searches had a geospatial dimension, 3) in 2008 Gan et al. (2008) reported 12,7 % of query rewrites in Yahoo! query logs add a geographical term and 4) R. Jones et al. (2008) found geographical terms in 12.7% of user query rewrites in Yahoo! search sessions.

Current web search engines (e.g. google, yahoo, bing,...) and IR engines are not fully “understanding” the geographical terms in the queries and in their retrieved documents because these terms are usually processed as normal keywords. There are no AI techniques that perform geographical reasoning in their results. Usually these systems treat geographical terms from queries as simple textual tokens without having into account its geographical meaning and the possible geographical restrictions that these terms can imply. As an example, the previous example of geographical query could led to find documents that mention “vineyards” in California by matching only both tokens “vineyards” and the geographical token “California” with all the indexed documents. In this way the IR system will not return or will return without appropiate ranking positions documents that could report “vineyards” in places of California but not mentioning California (e.g. “vineyards in Santa Barbara County”). On the other hand the IR system should also understand the semantic meaning of the thematic part (“vineyards”). For this reason the system will have to recognize documents that do not mention explicitly “vineyards” but synonyms or sets of words with the same or similar meaning such as: ”farm of grapevines” or ”plantation of grapevines”.

Theoretically, the treatment and automatic understanding of geographical terms appearing in user queries and indexed documents from IR systems (and major search engines) should provide an improvement of the results by retrieving documents that match the geographical restrictions in the query. There is evidence for a need for a more intelligent access to the information that could led to index more detailed informations for each geographical token.

In this context, Geographical Information Access (GIA) can be defined as a way of retrieving information that includes the automatic analysis and interpretation of queries with geographical constraints and geographical terms in document collections.

1.2 Researched Areas and Tasks Investigated

The research areas investigated in this thesis are both Information Extraction and Geographical Information Access. The following tasks related with Geographical Information Access were investigated: Geographical Information Retrieval, Geographical Question Answering, and Textual Georeferencing. Moreover GeoIA tasks require the research of Information Extraction methods to automatically understand geographical terms in electronic texts. So the areas of Toponym Recognition and Toponym Disambiguation were also researched. The Table 1.1 contains a description of the research areas investigated.

1.3 Toponym Recognition and Disambiguation

From the need of this geographically aware access to the information stored locally or on the web emerged the research task of understanding the geographical terms and expressions appearing in digital information. The tasks that allow this understanding are Toponym Recognition and Toponym Disambiguation (a detailed explanation of these tasks and its state-of-the-are is presented at Chapter 2).

Research Area	Task	Input	Output
Information Extraction	Toponym Recognition (TR)	text	toponyms
	Toponym Disambiguation (TD)	text with recognized toponyms	disambiguated unique referents of the toponyms
Information Access	Geographical Information Retrieval(GIR)	query (set of keywords) with with geographical constraints	set of Relevant documents
	Geographical Question Answering[4] (GeoQA)	question with geographical terms	answer or set of answers
	Textual Georeferencing (TG)	formal or informal text	Pair of coordinates[5] (where the text refers to)

Table 1.1: Set of Information Extraction and Information Access tasks investigated in this thesis.

Toponym Recognition (TR) is the task of automatically recognize geographical place names (toponyms) appearing in electronic texts. This task can be considered a sub-task of the more general problem of Named Entity Recognition and Classification (NERC). NERC is is the task of recognizing and properly classifying named Noun Phrases (Named Entities) in a set of predefined categories (Marrero et al., 2013). Most NEC systems reduce this set to the basic 7 MUC (Chinchor and Robinson, 1997) classes: LOCATION (e.g. "New York", "France"), PERSON (e.g. "Isaac Newton") , ORGANIZATION (e.g. "IBM") , MONEY (e.g. "$100"), PERCENT (e.g. "100%", DATE (e.g. "May 2015") ,TIME (e.g. "10 a.m.").

The TR task recognises the toponyms but a further task is necessary to establish which is the appropriate referent according to the context. As an example, the toponym "Paris" recognized in text can refer to multiple places around the world, (over 140 places (M. Lieberman et al., 2010a)) and some referents can have a feature type different from a city (e.g. "Paris" as a region or a river).

The Toponym Recognition task is generally solved by using the following methods (explained above) alone or in combination: 1) Geographical Gazetteers, and 2) Named Entity Recognition and Classification. Geographical Gazetteers can be defined as geospatial dictionaries of geographic names. Normally these geographical names are political and administrative areas, natural features, and man-made structures.

On the other hand, Toponym Disambiguation (TD) implies that (whenever it is possible) every geographical concept in the electronic documents and texts must be recognized, classified in a fine geographical ontology and disambiguated into its geographical world referent (georeferencing) (see examples in Figure 1.1). To apply a TD algorithm firstly a recognition step has to be performed with a Toponym Recognition (TR) algorithm to detect the following geographical concepts:

1. geographical names (toponyms): such as cities (e.g. "Barcelona", "Paris",...), countries (e.g. "Spain", "USA"), rivers (e.g. "Nile", "Amazonas",...), monuments (e.g. "Eiffel Tower",),...

[4] Includes the Query Parsing sub-task.

[5] This thesis presents systems that make the assumption that the text only refers to one place ("only one georeference predicted per text").

2. geographical feature types (e.g. “city”, “cities”, “country”, “village”, “dam”, “dock”, “fabrics”,..),

3. geographical coordinates (e.g. latitude and longitude of geographical places,...)

4. addresses (postal codes,...)

5. geographical expressions (involving spatial prepositions and toponyms): such as “South of France”, “North of Germany”,....

TD algorithms have to face the following ambiguity problems:

1. *Referent ambiguity problem.* This problem occurs when the same name is used for several locations (of the same or different class). The toponym “Paris” recognized in text can refer to multiple places around the world (over 140 places (M. Lieberman et al., 2010a)) such as such as Paris, Texas (USA) and Paris (France) and some referents can have a feature type different from a city (e.g. “Paris” as a region or a river).

2. *Reference ambiguity problem.* This problem occurs when the same location can have more than one name in the same language or in other languages (e.g. Wien, Vienne, or Viena as place names to refer to the city of Vienna (Austria) in German, French, and Spanish respectively).

3. *Referent class ambiguity problem.* The same name can be used for locations and also for other classes of Named Entities like persons or organizations (e.g. “Washington” and “Paris” as the US and French governments in a sentence like “Washington and Paris vetoed the candidate.”). This problem also happens when a common noun and a toponym are homonyms (e.g. “Aurora” (noun) vs “Aurora” (city)).

Type of Toponym	News Sample Extract
Paris as a geopolitical Named Entity meaning the French Government	*Los Angeles Times* 05/13/1994
	At the same time, Juppe dismissed reports that Washington and **Paris** were split over what should be done in Bosnia, contending that the two governments were essentially in agreement on broad policy and “both think the time has come for a political settlement.”
Paris as a part of the “Paris Match” French Magazine	*Los Angeles Times* 05/08/1994
	What were they going to do now? Make French travelers abroad wear earmuffs? Change the name of **Paris** Match to Paris Mzprfz? Seal off the Channel tunnel?
Paris, France as a geographical name and metonimic name (“City of Light”)	*Los Angeles Times* 05/08/1994
	In the long drive through France, Joe Miller’s platoon got just one break, a memorable one – two August days in **Paris**, after the 4th Division helped liberate the **City of Light**.
Paris, Texas a geographical name in Texas, USA	*Los Angeles Times* 02/11/1994
	There are 37 disaster locations on the Texas truck alone – from the **Paris**, Tex., tornado in 1984 to the Mexico City earthquake of 1985.

Figure 1.1: Example of some context-dependent geographical ambiguities of the term “Paris” from the *Los Angeles Times* newspaper (1994).

1.4 Geographical Information Retrieval

Geographical Information Retrieval (GIR) consists in searching documents with geographically restricted queries: the ones that involve both thematic and geographic search (e.g. "rice exportation in Japan"). Geographical queries are normally represented by a triplet < theme, spatial relationship, location > (C. B. Jones and R. S. Purves, 2008) (see some geographical queries in Figure 1.2).

Regarding the properties of geographical queries, the study of Sanderson and Kohler (2004b) provided useful data. In this study they manually analyzed a random sample of 2,500 queries extracted from a log of about 1 million queries from the Excite search engine log. For this analysis they defined a geographical query as a query which included at least one of the following types of geographic terms: place names (e.g. Houston, Texas, US), other locators (e.g. postcode, ZIP code), adjectives of place (e.g. American, international, western), terms descriptive of location (e.g. state, county, city, site, street), geographic features (e.g. island, lake), and directions (e.g. north, south.) Of the 2,500 queries, 18.6% contained a geographic term and and 14.8% held a place name. The one million queries were searched for terms indicating a spatial relationship (e.g. "in", "at", "from",...). About 9,960 queries (0.96% of the total data set) contained the word "in", 5,725 also contained a place name. In most of the queries "in" directly preceded (modified) a place name. There were 821 queries containing "at", of which 274 modified a place name. The spatial term "from" occurred 217 (out of 749) with a place name: generally used in the sense of something originating from, e.g. "famous people from philadelphia" or "flights from denver".

cabins to rent at lake tahoe.
cannon mountain.
fort pulaski national monument.
lakeside mall in michigan.
List of Restaurants in Ottawa.
macdougall dr in atlanta.
law blog in singapore.
plumbers in manhattan ny new york.
which airport is near to st julians in malta.
travel tips to the northwest usa.
bus trips from columbia.

Figure 1.2: Some geographical queries from the Windows Live Search log queries example from GeoQuery 2007 benchmark.

Current existing IR systems, based mostly on simple keyword search without the use of semantics, are yet not suitable to index and search geographical structures.

GIR systems thus have to deal with the following issues:

- Recognition and disambiguation of toponyms in documents and queries.
- Indexing and searching thematic and spatial information.
- Spatial relevance measures. (e.g taking into account hierarchical containment, adjacency of places, connectivity, proximity,....)
- Ranking of documents using both thematic and spatial relevance.

1.5 Geographical Question Answering

GeoQA is a more complex process than GIR that involves the use of Question Answering (QA) techniques to deal with geographical questions (see Figure 1.3 for some examples of geographical questions). In generic QA the user request is a single piece of information instead of an entire document, and the input is a question expressed in Natural Language instead of a an IR query (set of keywords). GeoQA involves the technologies that must deal with questions that involve any kind of geographical terms and reasoning.

GeoQA needs a set of NLP algorithms to perform a comprehension of the user textual request and the textual documents involved in the search. NLP techniques process electronic texts and analyze them in order to provide lexical, syntactic, semantic, and/or discourse information about the text. In the last years NLP tools such as part-of-speech taggers, Named Entity taggers, syntactic parsers and semantic taggers (e.g. using WordNet) have become widely used in several approaches for QA. Moreover specific geographical knowledge is required for the GeoQA task. Thus Geographical Gazetteers can be employed.

what are the top 10 places to live in America
what conventions are in las vegas this week
what countries are located in asia pacific
what is the climate like in spain
what is the closest airport to fortwalton Florida
what is the currency in egypt
what state is philadelphia
what state is washington dc in
what to do on long island
What is on in Wellington New Zealand during September 2006
What is the highest mountain in the Alps
What is the name of the first and oldest national park in the United States
What is the name of the first European to explore the coasts of New Zealand and Australia
What is the population of Switzerland
What Papers Are Needed To Travel To Mexico
where can i find travel information to las vegas
where can i get a flight to florida
where in the world is kota kinabalu
where is costa rica
where to get concert tickets in san diego
where to travel in november in europe
which airport is near to st julians in malta

Figure 1.3: Some extracted geographical questions from the Windows Live Search log 2007.

Current QA systems use a combination of Natural Language Processing and Information Retrieval Techniques. Generic Question Answering systems can be classified from different points of view. Discussed in Carbonell et al., 2000 and Burger et al., 2000, a set of 4 types of questioners could determine the type of QA system by means of the questioner type:

- **Level 1. Casual Questioner**. The Casual Question is the type of QA questioner who asks simple factual questions, which could be answered in a single short phrase. For Example: “Where is New York City?”, “What is the currency unit of India?” “Who was the first man in the moon?”, etc.

- **Level 2. Template Questioner**. This type of user requires a QA system that retrieves multiple documents and combine portions of answers into a single response.

The questions are basically factual but is required more information than a single phrase (e.g. “What do we know about X?”, ”What are all of the countries that border Brazil?”).

- **Level 3. Questioner as a Reporter.** The QA questioner focus on factual questions that need to pull together information from a variety of sources including multiple medias and multiple foreign languages.
- **Level 4. Professional Information Analyst.** This profile requires analytic tools capable of providing answers to complex, multi-faceted questions involving judgement terms that analysts might wish to pose to multiple, very large, very heterogeneous data sources, media types, multiple languages, multiple styles, formats, etc., ”

Moldovan et al., 1999 provided a taxonomy of QA based on the necessary knowledge to resolve the questions. They considered important the three following criteria: Knowledge Bases (KB), Reasoning, and Natural Language Processing (NLP) indexing techniques. Knowledge bases and reasoning provide the medium for building question contexts and matching them against text documents. Indexing identifies the text passages where answers may lie, and natural language processing provides a framework for answer extraction. See more details of these levels in Table 1.2.

Table 1.2: QA taxonomy based on Knowledge bases, reasoning and NLP techniques Moldovan et al., 1999

Class	KB	Reasoning	NLP/Indexing	Examples/Comments
1	dictionaries	simple heuristics, pattern matching	complex noun, apposition, simple semantics, keyword indexing	Q33: What is the largest city in Germany? A: .. Berlin, the largest city in Germany.. Answer is: simple datum or list of items found verbatim in a sentence keyword or paragraph.
2	ontologies	low level	verb nominalization, semantics, coherence, discourse	Q198: How did Socrates die? A: .. Socrates poisoned himself.. Answer is contained in multiple sentences, scattered throughout discourse a document.
3	very large KB	medium level	advanced nlp, semantic indexing	Q: What are the arguments for and against prayer in school? Answer across several texts.
4	Domain KA and Classification, HPKB	high level		Q: Should Fed raise interest rates at their next meeting? Answer across large number of specific documents, domain knowledge acquired automatically.
5	World knowledge	very high level, special purpose		Q: What should be the US foreign policy in the Balkans now? Answer is a solution to a complex, possible developing scenario.

1.6 Textual Georeferencing

Textual Georeferencing consists of extending the information of texts by predicting an explicit location in space and time[6] where and when the text refers. According to Hill (Hill, 2006) *”The application of georeferencing extends to almost all fields of academic and applied study, including the arts and humanities; social , physical, and life sciences; medicine; government administration; petroleum and mineral exploration; message understanding (text analysis); historical and genealogical research; and the documentation of personal histories”.*

Most of the textual and media content in the web is not georeferenced and this means that TG can be used in applications that need to know exactly or at least predict with some confidence geographical information related to these texts such as: the author’s location, the place where the textual content refers or both informations.

Current state-of-the-art Textual Georeferencing (TG) approaches use Knowledge-Based algorithms based on Geographical Gazetteers (Ferrés and Rodríguez, 2011a) or Data-Driven algorithms based on models learnt from huge data collections (Van Laere, 2013).

Textual georeferencing does automatic understanding of the geographical content in texts and involves the need (whenever is possible) for its recognition, disambiguation and grounding.

On the other hand, it must be taken into account that Georeferencing is a prediction task that in some cases it is not possible (or very difficult) to perform in some texts. See in Figure 1.4 a set of keywords associated to georeferenced photos.

violet, video, home, diego, book, dancing, kiss, 2009, august2009, queens, justviolet
canonsd870is, dubocepark, dog, dogs, chihuahua, bug, chieka, rolling, stinky.
crucible, oakland, fire, december132008, openhouse, art, craft, california, unitedstates, usa, pleaseaddtags.
egypt, scuba, diving, saabsehr, underwater, shaabshear, shabsheer, redsea.
cavern, airpocket, vortexspring, vortexsprings, vortex, florida, diving, scubadiving, scuba, underwater.
maddy, beagle, pool, phoenix, arizona, bob, beer, corona, whitley, gary.
estonia, tallinn, oldtown.
vancouver.
manifestation, crise, montpellier.

Figure 1.4: Sample of keywords (tags) associated to georeferenced Flickr photos extracted from MediaEval Placing Task 2010 Development Dataset. (Note: each line contains a set of tags associated to a photo.)

1.7 Objectives

The main objective of this thesis is the study, implementation, and evaluation of approaches that can improve effectiveness measures of the following Geographical Information Access tasks: GIR, GeoQA and TG.

[6] This thesis treats georeferencing in space.

1.8 Research Methodology

The research methodology followed involve these steps:

1. A definition of the Work Hypotheses: the first step of the research methodology is to establish the work hypotheses.
2. An initial research study: a study of the existing state-of-the-art techniques and systems.
3. An incremental and iterative Design-Test-Evaluation-Improvement research cycle:
    - Design and implementation of algorithms that theoretically can solve the tasks.
    - Evaluation and comparative research. Design of appropriate experiments and evaluation of the proposed algorithms with datasets that can serve as a good analysis of the results.
    - Error analysis.
    - Perform improvements of the algorithms by using state-of-the art algorithms and/or propose and implement novel solutions.

1.9 International Benchmarking

Most of the datasets and experiments of this thesis have been performed during evaluation benchmarks or after them with its provided datasets and judgements. An Evaluation Benchmark is a task or set of tasks in which the participants submit manual or automatic predictions to be assessed by human experts or automatically. Benchmarking is useful in research for these reasons: 1) allows the organization of workshops, collaborations and discussion among research groups, 2) provides datasets for experimentation to the participants and further researchers improve the state of the art, 3) provides high-quality evaluation platforms for the comparison state-of-the-art algorithms (M. Larson et al., 2015).

1.10 Work Hypotheses

The Work Hypotheses of this thesis are two following ones:

- **Hypothesis 1**. *Existing Geographical Knowledge Bases and Natural Language Processing techniques can help to face the problems related with Geographical Information Access.* This hypothesis emerge from the common-sense reasoning that both World Knowledge (geography) and Semantic Knowledge help us humans to classify, discover, and extract information from written text, and therefore it should have to be useful in computational ways to access the information.

- **Hypothesis 2**. *Common-Sense Heuristics that use existing Geographical Knowledge Bases and Natural Language Processing techniques can improve state-of-the-art Data-Driven algorithms for Geographical Information Access.* From the first hypothesis emerged this one: if the Geographical Knowledge and the Semantic Knowledge can help the automatic computational access to the information an easy way to prove it is to do it with heuristics derived from human common-sense.

1.11 Scope and Focus of the Thesis

This section describes the Scope of the Geographical Information Access approaches researched in this thesis.

1.11.1 Geographical Information Retrieval Scope

The scope of the GIR approaches implemented in this thesis is the following:

1. *Part-of-relationship queries.* In this thesis the main type of GIR queries treated are the part-of-relationship type of queries (i.e. those that contain the spatial relationship “in” (including ”at” and ”from”), such as “Tornados in Texas”). The other types of queries (e.g. ”near”, ”close”, ...) are treated like “in” queries. In few experiments the spatial relationship operators “near” and “close” have been taken into account.

2. *English Test Collections in the journalistic domain.* The document collections used to perform the GIR experiments described in this thesis come from the journalistic domain. The document collection consists of 169,477 stories from the British newspaper *The Glasgow Herald* (1995) and the American newspaper *Los Angeles Times* (1994).

3. *Testing with evaluation benchmarks datasets in official GIR evaluations and posterior experiments.* The approaches for GIR have been evaluated within the GeoCLEF GIR evaluations of 2005, 2006 and 2007 and posterior experiments with the full collection (2005, 2006, 2007, and 2008).

4. *Toponym Recognition is performed with NERC and Toponym Disambiguation uses partial and conservative context-independent heuristics based on Geographical Knowledge.*

5. *Evaluation of effectiveness measures.* The focus of the thesis is to improve effectiveness measures of GeoIA. In this thesis GIR is evaluated only with efectiveness measures:

concretely Mean Average Precision (MAP), R-Precision and Recall (these measures are defined in Chapter 3). Efficiency measures for GIR such as indexing and searching speed or index size are not reported and are out of the scope of this thesis. Statistical significance testing has been employed to compare GIR experiments but practical significance is not evaluated in this thesis.

1.11.2 Geographical Question Answering Scope

The scope of the GeoQA approaches implemented in this thesis is the following:

1. *Restricted-Domain and Open-Domain evaluations.* The main approach, which is an Scope-Based GeoQA system, has been evaluated in the Spanish geography domain. It means that the queries are focused on geographical entities, geographical names such as: cities, rivers, states or quantities such as: altitudes, population or extesion of places. The other approaches for GeoQA and Geographical Query Parsing have been evaluated in official GeoQA and Geographical Query parsing evaluation benchmarks: GikiCLEF (2009) and GeoQuery (2007).

2. *Text-based approaches. The approaches receive questions in natural language and perform QA over collections of textual documents.*

3. *Treatment of English and Spanish languages for GeoQA.*

4. *Geographic-scope based factoid questions at basic level of complexity.* The scope of the thesis deals with the treatment and processing of natural language queries in the form of factoid questions (”Where is Washington located?”) at a casual level of complexity (Carbonell et al., 2000). The kind of questions that can be answered are determined by a geographic scope (region, country, state) in which the system has to be adapted.

1.11.3 Textual Georeferencing Scope

The scope of the TG approaches implemented in this thesis is the following:

1. *One georeference per text.* The hypothesis of “one sense per discourse” applied in Word Sense Disambiguation (Gale et al., 1992) is applied in TG as “one georeference per text” in our formal and informal experiments (i.e. despite having many toponyms in the text, the predicted georeferencing of the text is a unique coordinates pair).

2. *Testing in official Textual Georeferencing evaluation benchmarks and posterior experiments with the datasets when evaluating informal documents.* The evaluation of the approaches has been performed in the context of the MediaEval 2010, 2011, and 2014 Placing Task multimodal georeferencing evaluation.

3. *A multilingual test collection is used when the approaches are evaluated with informal documents. The MediaEval Placing Task evaluation used informal documents from Flickr photos and videos (Flickr Meta-Data).*

4. *Toponym Recognition is performed with a Geographical Gazetteer and Toponym Disambiguation is performed with Geographical Knowledge and Population Heuristics.*

5. *English language is used when the approaches are evaluated with formal documents: a Wikipedia collection is used for testing with formal documents.*

6. *Geographical coordinates (latitude and longitude) use the WGS84 revision of the World Geodetic System.*

1.12 Summary of Contributions

This PhD thesis contributes to the state-of-the-art of Geographical Information Access (GeoIA) with the presentation and evaluation of several novel approaches that use effectively Geographical Knowledge and Natural Language Processing to deal with several tasks related to GeoIA. Several approaches have been implemented and evaluated in the tasks of Geographical Information Retrieval, Geographical Question Answering and Textual Georeferencing. Most of these algorithms have been presented in international benchmarking evaluations and some of them achieved state-of-the-art results (including some of the best results in GIR and TG tasks). The other experiments presented in this thesis achieved average or low performance compared with the other participants in the benchmarks but these experiments have shown scientific relevance by: a) showing that Geographical Knowlegdge and Natural Language Processing combined with Data-Driven methods can improve effectiveness measures of GeoIA tasks, or b) establishing baselines for future improvements in the task.

This section contains a brief summary of the contributions of this thesis (see in Chapter 7 a detailed description of these contributions). The main contributions of this thesis are:

1. **The presentation and description of several novel approaches for Geographical Information Access tasks.**

2. **The evaluation of these novel approaches for Geographical Information Access tasks.**

3. **The effective use of Geographical Knowledge and Natural Language Processing for the Geographical Information Retrieval tasks evaluated**.

4. ***Passage Retrieval Approaches for GIR***. Implementation and evaluation of two approaches that combine sucessfully Geographical Knowledge and Passage Retrieval presented at GeoCLEF 2005 and GeoCLEF 2006.

5. ***TALPGeoIR***. An approach that combines Geographical Knowledge Re-Ranking, Natural Language Processing and Relevance Feedback for Geographical Information Retrieval that achieved state-of-the-art results in official GeoCLEF benchmarks (Ferrés and Rodríguez, 2008a; Mandl et al., 2008) and posterior experiments (Ferrés and Rodríguez, 2015a).

6. ***GeoTALP-QA. A scope-based Geographical Question Answering Approach***. This thesis contributed to both GeoQA and Restricted-Domain QA state-of-the-art with the design and implementation of a Scope-based GeoQA system for Spanish and English and its evaluation with a set of questions of the Spanish geography (Ferrés and Rodríguez, 2006a).

7. ***State-of-the-art Textual Georeferencing approaches***. This thesis presented four novel approaches to generic Textual Georeferencing for informal and formal documents that achieved state-of-the-art results in evaluation benchmarks (Ferrés and Rodríguez, 2014) and posterior experiments (Ferrés and Rodríguez, 2011a; Ferrés and Rodríguez, 2015b).

8. ***A Geographical Query Parsing algorithm***. A Geographical Query Parsing algorithm that detects and extracts information from geographical queries that has been evaluated with search engine log queries in an evaluation benchmark (Ferrés and Rodríguez, 2008b).

9. ***GikiTALP: a simple Data-Driven baseline for Geographical Question Answering over Wikipedia.***

1.13 Publications

This section details the publications (grouped by chapter) that disseminate the research studies and results obtained from the work presented in this thesis.

Chapter 4. Geographical Information Retrieval Approaches. The following publications are related with approaches for GIR and its evaluation in the context of several GeoCLEF official evaluations and posterior experiments:

Book chapters

Daniel Ferrés, Alicia Ageno and Horacio Rodríguez.
The GeoTALP-IR System at GeoCLEF 2005: Experiments Using a QA-Based IR System, Linguistic Analysis, and a Geographical Thesaurus.
In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Lecture Notes in Computer Science. 2006. Vol 4022. Pages 947-955.

This paper presents GeoTALP-IR, the first GIR approach developed by the author is described and evaluated at GeoCLEF 2005 evaluation benchmark.

Daniel Ferrés and Horacio Rodríguez.
TALP at GeoCLEF 2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus.
Evaluation of Multilingual and Multi-modal Information Retrieval. Lecture Notes in Computer Science. Vol. 4730. Pages. 962-969. 2007.

This paper describes the second GIR approach proposed, TALPGeoIR 2006, and its evaluation in the context of the GeoCLEF 2006 evaluation benchmark.

Daniel Ferrés and Horacio Rodríguez.
TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier.
Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science. Vol. 5152. Pages 830-833. Springer. 2008.

This paper reports an analysis of the results of the TALPGeoIR 2007 approach evaluated at GeoCLEF 2007 evaluation benchmark, where it achieved the top-ranked position in Monolingual English GIR.

Daniel Ferrés and Horacio Rodríguez.
Evaluating Geographical Knowledge Re-Ranking, Linguistic Processing and Query Expansion Techniques for Geographical Information Retrieval .
Proceedings of the 22th International Symposium on String Processing and Information Retrieval (SPIRE 2015). September, 2015. London, UK. Lecture Notes in Computer Science. Vol. 9309, pages 311-323. Springer. 2015.

This paper shows the evaluation of the different components of the TALPGeoIR applied over three state-of-the-art IR algorithms: TF-IDF, BM25 and InL2. The components evaluated were: Geographical Knowledge-Reranking, Linguistic Processing, and Query Expansion with Relevance Feedback. The evaluation was done with the full GeoCLEF collections from 2005 to 2008 (100 topics) and showed improvement of the MAP effectiveness measure over of the best official results at GeoCLEF evaluations of 2005, 2006, and 2007.

Conference Proceedings

Daniel Ferrés and Horacio Rodríguez.
TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering
Working Notes for CLEF 2007 Workshop co-located with the 11th European Conference on Digital Libraries (ECDL 2007), Budapest, Hungary, September 19-21, 2007.

The third GIR approach proposed, TALPGeoIR 2007, is described and evaluated at GeoCLEF 2007 evaluation benchmark (where it achieved the top-ranked runs in Monolingual English GIR) in this paper.

Chapter 5. Geographical Question Answering Approaches: The following publications are related with GeoQA and Geographical Query Parsing approaches evaluated with closed collections or within evaluation benchmarks:

Book Chapters

Daniel Ferrés and Horacio Rodríguez.
TALP at GeoQuery 2007: Linguistic and Geographical Analysis for Query Parsing.
Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science, Vol. 5152. Pages 834-837. 2008.

This paper describes the system presented at GeoQuery2007 and analyzes the results.

Daniel Ferrés and Horacio Rodríguez.
TALP at GikiCLEF 2009.
Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments. Lecture Notes in Computer Science. Vol. 6241. Pages 322-325. 2010.

This paper describes experiments in Geographical Information Retrieval with the Wikipedia collection in the context of the participation in the GikiCLEF 2009 Multilingual task in English and Spanish.

Conference Proceedings

Daniel Ferrés and Horacio Rodríguez,
Experiments Adapting an Open-Domain Question Answering System to the Geographical Domain Using Scope-Based Resources.
Proceedings of the Multilingual Question Answering Workshop of the EACL 2006. 2006. Trento, Italy.

This paper describes an approach to adapt an existing multilingual Open-Domain Question Answering (ODQA) system for factoid questions to a Restricted Domain, the Geographical Domain.

Jordi Luque and Daniel Ferrés and Javier Hernando and José B. Mariño and Horacio Rodríguez.
GeoVAQA: A Voice Activated Geographical Question Answering System.
Actas de las IV Jornadas en Tecnología del Habla (4JTH). November, 2006, Zaragoza, Spain

This paper describes GeoVAQA, a voice-activated Geographical QA system. The author of this thesis contributed to this paper with a textual Geographical QA system that receives questions previously recognized by an automatic speech recognition (ASR) system.

Chapter 6. Textual Georeferencing Approaches: These publications are related with TG in the context of official Media Eval Placing Task (MEPT) evaluations and posterior experiments with the MEPT datasets:

Book Chapters

Daniel Ferrés and Horacio Rodríguez
Knowledge-Based and Data-Driven Approaches for Georeferencing of Informal Documents. Proceedings of the 8th International Conference on Text, Speech and Dialogue TSD 2015. September, 2015. Plzen, Czech Republic.
Lecture Notes in Computer Science. Vol 9302. Springer. Pages 452-460.

This paper describes four Georeferencing approaches, experiments, and results at the MediaEval 2014 Placing Task (ME2014PT) evaluation, and posterior experiments. Some of the approaches achieved state-of-the-art results at ME2014PT evaluation and posterior experiments, including the best results for distance accuracies of 1000km and 5,000km in the task where only the official training dataset can be used to predict the coordinates.

Conference Proceedings

Daniel Ferrés and Horacio Rodríguez
TALP at MediaEval 2010 Placing Task: Geographical Focus Detection of Flickr Textual Annotations.
Working Notes of the Mediaeval 2010 Evaluation. October 2010. Pisa, Italy.

This paper describes the textual georeferencing experiments in the context of the Multimedia Placing Task at the MediaEval 2010 evaluation benchmark. In these experiments only Geographical Knowledge (gazetteers) and limited NLP (stopwords and dictionaries) were used to predict.

Daniel Ferrés and Horacio Rodríguez
Georeferencing Textual Annotations and Tagsets with Geographical Knowledge and Language Models
Actas de la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural). September 2011. Huelva, Spain.

This paper proposed 4 new generic textual georeferencing approaches based on Geographical Knowledge Bases, Linguistic Knowledge, and Information Retrieval. These approaches have been evaluated with the MediaEval 2010 dataset and outperformed the best results in accuracy reported by the state-of-the art systems that participated at MediaEval 2010 official Placing task.

Daniel Ferrés and Horacio Rodríguez
TALP at MediaEval 2011 Placing Task: Georeferencing Flickr Videos with Geographical Knowledge and Information Retrieval.
Working Notes of the Mediaeval 2011 Evaluation. October 2011. Amsterdam, Holand.

This paper describes the textual georeferencing experiments in the context of the Multimedia Placing Task at the MediaEval 2011 evaluation benchmark.

Daniel Ferrés and Horacio Rodríguez
TALP-UPC at MediaEval 2014 Placing Task: Combining Geographical Knowledge Bases and Language Models for Large-Scale Textual Georeferencing.
Working Notes of the Mediaeval 2014 Evaluation. October 2014. Barcelona, Spain.

This paper describes the textual georeferencing experiments in the context of the Multimedia Placing Task at the MediaEval 2014 evaluation benchmark.

Annex G: Web Person Search experiments at WePS-3

Conference Proceedings

Daniel Ferrés and Horacio Rodríguez
TALP at WePS-3 2010.
CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy. Ed. by M. Braschler, D. Harman, and E. Pianta. Vol. 1176. CEUR Workshop Proceedings.

This paper describes the Web Person Search experiments at WePS-3.

1.14 Structure of the Document

The rest of this thesis is structured as follows:

Chapter 2 - Toponym Recognition and Disambiguation - State-of-the-art

Geographical Information Access applications require the recognition and sometimes the disambiguation of the toponyms. This Chapter describes the state of the art of Toponym Recognition and Toponym Disambiguation, including: 1) the two main approaches for TR: Geographical Gazetteers lookup and Named Entity Recognition and Classification based on NLP tools, and 2) state-of-the-art TD heuristics and systems.

Chapter 3 - Geographical Information Access Tasks - State-of-the-art

This Chapter describes the state-of-the-art of the three Geographical Information Access tasks that are treated in this thesis: GIR, GeoQA, and TG.

Chapter 4 - Geographical Information Retrieval Approaches

This Chapter describes the approaches that the author of this thesis presented at several GeoCLEF GIR evaluations (2005, 2006, and 2007) and performs a depth evaluation of the last approach presented with new experiments that have been performed in 2015 (Ferrés and Rodríguez, 2015a).

Chapter 5 - Geographical Question Answering Approaches

This Chapter describes the GeoQA approaches, the experiments to evaluate them and the results. First, describes a system that consists of an adaptation of an ODQA (Open Domain Question Answering) to the Geographical Domain. Then, a system for Geographical QA over the wikipedia is evaluated at CLEF's GikiCLEF 2009 evaluation. Finally, a system for analyzing Geographical queries is presented in the context of the CLEF's GeoQuery 2007 evaluation.

Chapter 6 - Textual Georeferencing Approaches

This Chapter describes generic approaches for georeferencing formal and informal documents. The informal documents have been evaluated in the context of several Media Eval Placing Task evaluations (2010, 2011, and 2014) and posterior experiments after these evaluations. The formal documents have been evaluated with an existing test set for georeferencing Wikipedia documents. Moreover a set of experiments with emergency scenarios have been performed using a subset of the Media Eval Placing Task 2014 dataset.

Chapter 7 - Conclusions

The last Chapter describes the contributions of the author of this dissertation to the research fields investigated, reports the limitations of the work, and proposes further work to develop in the researched areas.

Annex A - Test Collections

This annex describes the datasets employed in the experiments performed in this thesis and provides links to download them.

Annex B - GeoCLEF Topics List and Topics Classification

This annex shows the GeoCLEF topics used in the GeoCLEF evaluation benchmarks from 2005 to 2008 (a total of 100 topics). This annex also includes several classification types of these topics.

Annex C - GeoCLEF Per-Query Results

This annex shows the GeoCLEF Per Query results of the three IR algorithms used: TF-IDF, BM25 and HLM. The results include the experiments with the following fields: title alone, title and description, and all tags (title, description and narrative).

Annex D - Spanish GeoQA Questions

This annex shows the test set of the GeoQA experiments. These sets contain 62 questions about the Spanish geography.

Annex E - GikiCLEF Questions

This annex shows the GikiCLEF topics in English and Spanish used in the GikiCLEF 2009 evaluation benchmark (a total of 50 questions per language).

Annex F - Geographical Feature Types Mappings

This annex shows the geographical feature types data mappings developed in this thesis.

Annex G - Web Person Search experiments at WePS-3

This annex describes the Web Person Search experiments at the WePS-3 benchmark.

CHAPTER 2

Toponym Recognition and Disambiguation - State-of-the-art

Toponym Recognition (TR) is the task of automatically recognize geographical place names (toponyms) appearing in electronic texts. This task can be considered a sub-task of the more general problem of Named Entity Recognition and Classification (NERC), in which Named Entities (NE) are recognized in text and classified into a set of categories. Marrero et al. (2013) discusses that NE can be defined by four criteria according to the experts in the field: 1) grammatical category (proper nouns or common names acting as common nouns), 2) rigid designator 3) unique identifier and 4) purpose and domain of the application, being this later definition criteria the only one consistent with the literature, evaluation forums and tools.

The TR task recognises the toponyms but a further task is necessary to establish which is the appropiate referent according to the context. As an example, the toponym “Paris” recognized in text can refer to multiple places around the world, (over 140 places (M. Lieberman et al., 2010a)) and some referents can hava a feature type different from a city (e.g. “Paris” as a region or a river). Toponym Disambiguation is the task of automatically disambiguate the toponyms recognized in electronic texts by chosing their unique referents in the context (if possible). This task implies that (whenever it is possible) every geographical concept in the text must be classified in a geographical class (or feature type) and disambiguated into its geographical world referent. The process that involves Toponym Recognition and Toponym Disambiguation has been named “geotagging” in the literature (Amitay et al., 2005; M. Lieberman et al., 2010a).

Once the TD process has been completed the disambiguated entries can be associated to a unique spatial coordinates and regions. Geotagging thus allows to pass from the name-based and discrete informal way of expressing locations to the continuous, formal way of spatial representation of places by coordinates, polygons, boxes,... (Hill, 2006).

2.1 Toponym Recognition Approaches

Toponym Recognition implies also the task of Toponym Normalization (e.g. understanding that “UK”, “U.K.”, and “United Kingdom” are refering to the same geographical or geopolitical entity) and solve geo/non-geo ambiguities (e.g. deciding which word is correct in the context: “Metro” (noun) or “Metro” (city)). Although Toponym Normalization is easily solved with Gazetteers that include alternate names of toponyms, the geo/non-geo ambiguity is a more complex problem. For instance, Volz et al. (2007) show that 11,5% of WordNet 2.0 names intersect with geographic names. Amitay et al. (2005) presented approach to the geo/non-geo ambiguity that consists to create lists of place names with a very common non-geo sense by counting the names of the place names occurred in a large corpus and filtering those with a high frequency and those with small population. Their approach required manual pass to remove errors or add missing names.

The Toponym Recognition task is generally solved by using the following methods (explained above) alone or in combination: 1) Geographical Gazetteers, 2) Named Entity Recognition.

2.1.1 Geographical Gazetteers

Geographical Gazetteers can be defined as geospatial dictionaries of geographic names. Normally these places can be political and administrative areas, natural features, and manmade structures. They contain large lists of geographical entities, normally enriched with some information such as: geographical feature type (e.g. “city”, “country”), location (e.g. geographical coordinates such as longitude and latitude), elevation, population, language, inclusive relations (e.g. referent of the state and/or country where is located) (Hill, 2006). Some of the most relevant geographical gazetteers are described here:

- **GEOnet Names Server (GNS[1]).** A worldwide database of geographic feature names, excluding the United States and Antarctica, with 5.5 million entries. The coordinate system for data served by GNS is WGS84. Each gazetteer entry contains a geographical name (toponym) and its feature class and code, geographical coordinates (latitude, longitude), language of the geographical name and other features as country, first administrative division, etc.

- **Geographic Names Information System** (GNIS[2]). A gazetteer with 2.0 million entries about geographic features of the United States and its territories. GNIS gazetteer was developed by the U.S. Geological Survey in cooperation with the U.S. Board on Geographic Names (BGN). The entries contain the following fields: geographical name , feature Type, U.S. County, U.S. State, Geographical Coordinates, Elevation, etc.

- **Alexandria Digital Gazetteer[3] (ADL).** The ADL gazetteer is a geospatially defined geographic name datasets with about 4 million entries (Frew et al., 1998).

[1] **GNS**. `http://geonames.nga.mil/gns/html/namefiles.html`

[2] **GNIS**. `http://geonames.usgs.gov/domestic/download_data.htm`

[3] **ADL**. `http://legacy.alexandria.ucsb.edu/gazetteer/`

- **Alexandria Digital Gazetteer Feature Type Thesaurus (ADLFTT).** The ADL Feature Type Thesaurus [4] is a hierarchical (4 levels) set of about 210 geographical terms used to type named geographic places in English (Hill, 2000). The top level of the ADLFTT hierarchy has the following classes: administrative areas, hydrographic features, land parcels, manmade features, physiographic features, and regions.

- **GeoWorldMap**[5] gazetteer with approximately 40,594 entries (countries, regions and important cities). The countries data contains information associated to world countries such as:

  - Country Name (e.g. “India”, “United States”,...).
  - FIPS10-4 code[6]: a four letter code that identifies geopolitical entities. This code was established by the Federal Information Processing Standarts institution of U.S.A.
  - ISO2 code: is a 2 alpha-numeric characters code that represents a country (e.g. “US” represents the United States of America). More information on the ISO2 code can be found at the ISO 3166 standard[7].
  - ISO3 code: is a 3 alpha-numeric characters code that represents a country (e.g. “GTM” represents Guatemala). More information on the ISO3 code can be found at the ISO 3166 standard.
  - ISON: is the number column in the ISO 3166 document which lists each country with associated alpha and numeric codes (e.g. “036” represents Australia).
  - Internet; the “ccTLD” code designated by Internet Assigned Numbers Authority (IANA)[8]. This code is employed in the DNS to identify hosts in various countries around the world (e.g. the code “uk” is assigned to the United Kingdom). The “ccTLD” codes are based on ISO2 codes.
  - Capital; this field is a string representation of a given countries capital city, for example, the capital of United States is “Washington, DC”.
  - MapReference; this field is a string representation of a given countries, major reference point in the world, for example, the map reference for Canada is “North America”.
  - Nationality Demonyms in singular and plural: the singular and plural expression of a given countries nationality (e.g. “German” and “Germans” for Germany).
  - Currency: a string with the country currency (e.g. “Yen” for Japan and “US Dolar” for United States).
  - CurrencyCode: is a 3 character string representation of a given countries currency code based on the ISO 4217[9] standard. The first 2 characters are made up of the countries Internet code and the last is a currency designator (e.g. the currency code “USD” is the code for the US Dollar and “JPY” for Japan).

[4] ADLFTT. http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm

[5] **Geobytes Inc.**: http://www.geobytes.com/

[6] FIPS-104 publication. www.nimal.mil/gns/html/index.html

[7] http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html

[8] http://www.iana.org,

[9] www.iso.org

  - population: estimation of country population (2001).
  - Title: the country's title appearing in a sentence (e.g. "The United Kingdom" for the United Kingdom).

  The regions data contains a list of sub-country geographical entities such as states, provinces and territories, etc. Each region has a relationship with its country, an special 2 character code abbreviation (ISO 3166-2 codes) and the ADM1 code to identify a geopolitical region at sub-country level.

  The data about cities is a detailed list of cities of the world and its relationship of membership with countries and sub-country regions. Each city has also its longitude, latitude and time zone as an associated data.

- **UN-LOCODE**. The official gazetteer by the United Nations[10], with more than 36,000 locations in 234 countries.
- **Getty Thesaurus of Geographic Names (TGN)**. This gazetteer was compiled by the Getty Research Institute. The TGN includes names and associated information about places. Places in TGN include administrative political entities (e.g., cities, nations) and physical features (e.g., mountains, rivers). Current and historical places are included. The TGN is a structured vocabulary currently containing around 1,102,000 names and other information about places. Names for a place may include names in the vernacular language, English, other languages, historical names, names and in natural order and inverted order. Among these names, one is flagged as the preferred name. There are around 911,000 places in the TGN hierarchy with geographic coordinates, notes, sources for the data, and place types, role of the place (e.g., inhabited place and state capital) and temporal information coverage.
- **Heavens-Above GmbH Gazetteer**. Heavens Above is a private company which offers a gazetteer data to specify geographic location in order to orient sky charts, satellite fly-overs, etc.
- **World Gazetteer**[11]: a gazetteer with approximately 171,021 entries of towns, administrative divisions and agglomerations with their features and current population (see some example records in Table 2.1).

493866395 Ouroux-en-Morvan Ouroux locality 669 4718 395 France Bourgogne Nièvre
478809098 Cambridge locality 1217 4300 -8902 United States of America Wisconsin
478918662 Cambridge locality 1900 4303 -7338 United States of America New York
511974013 Cambridge Caergrawnt, Cambridge-Milton locality 128488 5221 13 United Kingdom England
511939112 Marienborn locality 531 5220 1112 Germany Sachsen-Anhalt Magdeburg Bördekreis

Table 2.1: World Gazetter records examples

- **Geonames**[12].

  The *Geonames* geographical database contains over 10 million geographical names and consists of 9 million unique features whereof 2.8 million populated places and

[10] **UN-LOCODE.** `http://www.unece.org/cefact/locode/service/main.htm`

[11] World Gazetteer copy (the original site is not online). `http://biit.cs.ut.ee/biodc/dataen.zip`

[12] **Geonames.** `http://www.geonames.org`

5.5 million alternate names. All features are categorized into one out of nine feature classes (see this classes in Table 2.2) and further subcategorized into one out of 645 feature codes (see the most important feature codes in Table 2.3). Geonames is integrating geographical data such as names, altitude, population and others from various sources (see the entries' fields description in Table 2.4). All lat/long coordinates are in WGS84 (World Geodetic System 1984). The sources used by this KB are: *NGA*: National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names (most names except US and CA), *GNIS* : U.S. Geological Survey Geographic Names Information System (names in US), `www.geobase.ca` (names in CA), *gtopo30* (elevation data), and *Wikipedia*.

Feature class	feature types
(A) Administrative Boundary Features	(country, state, region,...)
(H) Hydrographic Features	(stream, lake, ...)
(L) Area Features	(parks,area, ...)
(P) Populated Place Features	(city, village,...)
(R) Road / Railroad Features	(road, railroad,...)
(S) Spot Features	(spot, building, farm,...)
(T) Hypsographic Features	(mountain,hill,rock,...)
(U) Undersea Features	(undersea)
(V) Vegetation Features	(forest,heath,...)

Table 2.2: Geonames feature classes.

- **Pertaynims Gazetteers**. A set of nationalities-countries (e.g. "Japanese"-"Japan") lists can be obtained automatically from WordNet. These kind of relationships are called pertaynims (or demonyms). As shown in Greenwood (2004), pertaynims are useful for IR queries for QA, because answers to questions which include a location often occur in close proximity to the adjective form of the location, hence including the adjective form in the IR query increase the coverage of the retrieved documents.

Besides these toponym gazetteers some widely used ontologies offer a nice coverage of geographical entities. Between them Wikipedia[13], DBpedia[14], FreeBase[15], and YAGO (Suchanek et al., 2007).

2.1.2 Named Entity Recognition and Classification

Named Entity Recognition and Classification (NERC) is the task of recognizing and properly classifying named Noun Phrases in a set of predefined categories. NERC is a central issue in many basic NLP tasks such as co-reference resolution, document linking or topic detection, and also has currently become present in most of the Information Access and Text Mining applications. NERC can be seen as a two-step process: Named Entity Recognition (NER) and Named Entity Classification (NEC). NER consists on locating a sequence of one or more contiguous words that can be considered candidate to be a Named Entity and deciding if it is an actual one. NEC implies assigning a class from a closed dataset to the NE. Most NEC systems reduce this set to the basic 7 MUC classes: LOCATION, PERSON, etc. (see Table 2.5), while finer grained classification has been faced in extended NEC (Sekine et al.,

[13] `http://www.wikipedia.org`
[14] `http://wiki.dbpedia.org/`
[15] `http://www.freebase`

Feature class	feature types
A.ADM1	first-order administrative division (such as a state in U.S.A.)
A.ADM2	second-order administrative division
A.ADM3	third-order administrative division
A.ADM4	fourth-order administrative division
A.ADMD	administrative division
A.LTER	leased area
A.PCL	political entity
A.PCLD	dependent political entity
A.PCLF	freely associated state
A.PCLI	independent political entity
A.PCLIX	section of independent political entity
A.PCLS	semi-independent political entity
A.TERR	territory
A.ZN	zone
H.LK	lake
H.OCN	ocean
H.SEA	sea
H.STM	stream
L.AREA	area
L.CONT	continent
L.LCTY	locality
P.PPL	populated place (city, town, village,...)
P.PPLA	seat of a first-order administrative division
P.PPLA2	seat of a second-order administrative division
P.PPLA3	seat of a third-order administrative division
P.PPLA4	seat of a fourth-order administrative division
P.PPLC	capital of a political entity
S.CH	church
S.CMTY	cemetery
S.FRM	farm
S.SCH	school
T.HLL	hill
T.ISL	island
T.MT	mountain

Table 2.3: Some important geonames feature classes.

2002). Note that NERC can also be named NER (Named Entity Recognition) in the NLP literature (Ratinov and Roth (2009) and Marrero et al. (2013)) but in this thesis NER will refer only to the Named Entity detection part of the NERC process explained above.

Natural Language Processing (NLP) tools are widely used in NERC tasks applied to Toponym Recognition. NLP tools annotate detailed information about the words, relations between words and sentences. Current state of the art NERC tools and NERC systems for Toponym Recognition frequently only use basic level (lexical analysis) tools such as Tokenization, Sentence Splitting, and Part-of-Speech (POS) Tagging. Part-of-Speech tagging task consists in attaching the lexical category of each lexical unit of a sentence. The most common POS tag-set for English is Penn Tree-Bank[16] (PTB) tag-set. The EAGLES[17] group tag-set is used for other languages (e.g. Spanish). State-of-the art POS tagging techniques

[16] **Penn Tree-Bank (PTB).** `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`

[17] **EAGLES group**. `http://www.ilc.cnr.it/EAGLES96/home.html`

Databased Field	Description
geonameid (id)	geonames database identifier (id)
name	name of geographical point (utf8)
asciiname	name of geographical point in plain ascii characters
alternatenames	alternatenames
latitude	latitude in decimal degrees (wgs84)
longitude	longitude in decimal degrees (wgs84)
feature class	feature class (9 types)
feature code	feature code (635 types)
country code	ISO-3166 2-letter country code
cc2	alternate country codes
admin1 code	fipscode
admin2 code	code for the second administrative division (a county in the US)
admin3 code	code for third level administrative division
admin4 code	code for fourth level administrative division
population	population count
elevation	elevation in meters
dem	digital elevation model
timezone	timezone id
modification date	date of last modification in yyyy-MM-dd format

Table 2.4: Geonames Database Entry Fields Description.

achieve high effectiveness. The HMM statistical-based tagger Trigrams 'n' Tags (TnT) (Brants, 2000) performs 96.7% of accuracy in English when trained with the WSJ corpus. Collins (2002) used algorithms based on the perceptron algorithm and Viterbi decoding, performing an accuracy of 97.11% On the other hand, Toutanova et al. (2003) reported an accuracy of 97.24% over the Penn Tree-bank WSJ corpus using Bidirectional Dependency Networks. Finally, a Support Vector Machines approach, SVMTagger (Giménez and Márquez, 2004) outperformed TnT with a 97.2% of accuracy in the WSJ corpus. SVMTagger achieves also good results for Spanish: 96.89% of accuracy.

Different NERC systems have been evaluated in several NERC tasks in different international Information Extraction conferences and workshops. In 1996 the Multilingual Entity Task (Merchant and Okurowski, 1996; Sundheim, 1995a) in the Message Understanding Conference (MUC-6) (Sundheim, 1995b), was the first evaluation on NERC. In the MUC evaluations the following expressions to detect were considered: Named Entities (persons, locations and organizations), temporal expressions (time and date) and numeric expressions (percentage and money) (see Table 2.5). An example is given in Figure 2.1.

<ENAMEX TYPE="ORGANIZATION">Grupo Televisa</ENAMEX> and <ENAMEX TYPE="ORGANIZATION">Globo</ENAMEX> plan to offer national and local programming in Spanish and Portuguese. Initially, the venture's partners said they planned to invest <NUMEX TYPE="MONEY">$500 million</NUMEX>.
But a similar explosion <TIMEX TYPE="DATE">last year</TIMEX> delayed the plans of several American media companies to offer a package of satellite television services in <ENAMEX TYPE="LOCATION">Asia</ENAMEX>.

Figure 2.1: Example of NERC tagging from MUC-7 Conference.

Element	Entity Class	Expected Names
ENAMEX	ORGANIZATION	named corporate, governmental, or other organizational entity
	PERSON	named person or family
	LOCATION	name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)
TIMEX	DATE	complete or partial date expression
	TIME	complete or partial expression of time of day
NUMEX	MONEY	monetary expression
	PERCENT	percentage

Table 2.5: Tag elements and entity classes at the NE task of MUC conferences (Chinchor and Robinson, 1997).

Different Information Extraction contests organized Named Entity Extraction tasks: In 1996 MET (Multilingual Entity Task) for instance organized a task similar to the MUC one for Spanish, Chinese and Japanese (Merchant and Okurowski, 1996). MET task consisted allowed 10 entity types: Person, Organization, Location, Date, Time, Duration, Percent, Money, Measure, and Number. The conference on *Computational Natural Language Learning* (CoNLL)[18] organized a shared Task[19] [20] in 2002 and 2003 that consisted on NERC for four languages: English, German, Dutch and Spanish (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). This task consisted in an evaluation of different state-of-the-art algorithms for NERC. The classes used were: person, organization, location and others.

The NE task in MUC was inherited by the ACE project[21] in the U.S.A., where 2 new categories are added, GPE (Geographical and Political Entities, such as "France" or "New York") and facility, such as "Empire State Building" . In the ACE project were used 5 coarse classes (ENAMEX, TIMEX, NUMEX, MEASURE, CARDINAL) which could be expanded to 11 classes (Person, Organization, Location, GPE, Facility, Date, Time, Duration, Percent, Money, Measure, and Number).

Most state-of-the-art NEC systems use coarse-grained MUC-style datasets for performing the classification task reducing it to distinguish among LOCATION, PERSON, ORGANIZATION and so. Sekine et al. (2002) proposed an extended NE hierarchy of 150 types, while Manov et al. (2003) used 97 classes for the location sub-ontology. The approaches to NERC (Nadeau and Sekine, 2007) include manual rules, supervised or unsupervised Machine Learning, and hybrid approaches (see in Table 2.6 the evaluation of some NERC approaches over the CoNLL-2003 English dataset). Current state-of-the art open source and commercial NERC systems include: 1) LingPipe[22] , 2) Stanford CoreNLP[23] ,3)

[18] **CoNLL**. CoNLL is the yearly conference of SIGNLL, the Special Interest Group of the Association for Computational Linguistics on Machine Learning of Language; `http://www.aclweb.org/signll`

[19] **CoNLL Shared Task 2002.** `http://www.cnts.ua.ac.be/conll2002/ner/`

[20] **CoNLL Shared Task 2003.** `http://www.cnts.ua.ac.be/conll2003/ner/`

[21] **ACE project.** `http://www.itl.nist.gov/iad/894.01/tests/ace/`

[22] LingPipe. `http://alias-i.com/lingpipe/`

[23] `http://nlp.stanford.edu/software/corenlp.shtml`

OpenCalais[24], 4) Freeling[25] (for NER only) (Padró and Stanilovsky, 2012), 5) GATE[26], 6) OpenNLP[27], 7) Illinois Named Entity Tagger[28] (Ratinov and Roth, 2009), among others.

System	F1-measure
Carreras et al., 2003a	85.00%
Carreras et al., 2003b	84.30%
Chieu and Ng, 2003	88.31%
Curran and Clark, 2003	84.89%
Zhang and Lee, 2003	85.50%
Florian et al., 2003	88.76%
Klein et al., 2003	86.07%
Ratinov and Roth, 2009	90.80%
Passos et al., 2014	90.90%

Table 2.6: NERC Approaches to Named Entity Recognition and Classification for English evaluated with the CoNLL-2003 benchmark dataset.

The main features used for NERC are lexical features, part-of-speech tags, previously predicted NE tags, affix information (n-grams), orthographic information, gazetteers, chunk tags, orthographic patterns and global case information (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). Current state-of-the-art NERC systems such as Ratinov and Roth (2009) and Passos et al. (2014) use use Brown clusters and lexicon-infusee embeddings respectively.

Named Entity Evaluation tasks use the evaluation measures of Precision, Recall and F1. Precision is the percentage of NEs found that are predicted correctly:

$$\text{precision} = \frac{\#NEs\_predicted\_and\_correct}{\#NEs} \tag{2.1}$$

The recall measures the proportion of NE present in the corpus that are found by the system:

$$\text{recall} = \frac{\#NEs\_found}{\#NEs} \tag{2.2}$$

The F measure controls the relative importance of recall and precision. The general formula of the F measure is:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{2.3}$$

The β parameter can be used to tuned the relative importance of the recall and precision. NERC evaluations often use β set to 1. In this case the F-Measure is called F1.

$$F_{(\beta=1)} = \frac{2PR}{P + R} \tag{2.4}$$

[24] http://www.opencalais.com/

[25] nlp.lsi.upc.edu/freeling/

[26] https://gate.ac.uk/

[27] http://opennlp.apache.org/

[28] http://cogcomp.cs.illinois.edu/page/software_view/NETagger

2.2 Toponym Disambiguation Approaches

Toponym Disambiguation (TD) approaches usually have the following phases:

- **Toponym Feature Type Disambiguation** . Some approaches can apply fine subclassification using extended NE hierarchies to classify the Named Entities identitified as LOCATION as the Perseus system Smith and Crane (2001) or Ferrés et al. (2004b) for geographical NEs. Sekine et al. (2002) uses an extended NE hierarchy of 150 types and Manov et al. (2003) use 97 classes for the location sub-ontology. At this point sometimes a geographical feature type disambiguation procedure must be applied to decide at which subclass pertains the place name (e.g. in some contexts the Named Entity "Buffalo" could be a city or a river). A Geographical Name Place Class Disambiguator normally tries to disambiguate those NEs using features from the document in which the Named Entity appears and optionally features from external resources to decide in which subclass pertains. But some systems apply a class-based disambiguation in which some feature types are more important than others (e.g. the feature type "country" has priority over "city"),
- **Toponym Resolution**. The disambiguation phase in which a set of possible referents for each toponym in the text is computed and then a set of heuristics or algorithms is applied to reduce the number of referents per toponym or get a unique one (if possible).
- **Toponym Grounding**. Finally, the last process is the Grounding of geographical NEs, i.e. mapping a geographical NE to its appropriate physical (spatial) location (coordinates, area, etc.) using information derived from Geographical Gazetteers, as in Leidner et al. (2003).

Geographical ambiguity problems treated by TD systems include:

- **Referent ambiguity problem.** This problem occurs when the same name is used for several locations (of the same or different class). Some authors (H. Li et al., 2003) note the similarity of this problem to the Word Sense Disambiguation (WSD) problem.

  In some texts sometimes it is impossible to solve this ambiguity, and, in this case, we have to accept as correct all of the possible interpretations (or a superclass of them). Otherwise, a trigger phrase pattern can be used to resolve the ambiguity (e.g. "Madrid" is an ambiguous NE, but in the phrase, "State of Madrid", the ambiguity is solved by the feature type).

  The basic approaches to this problem are:

  1. **One referent per discourse**. Some of the approaches to the Toponym Disambiguatin task use the *one referent per discourse* heuristic following a similar approach to the Word Sense Disambiguation (WSD) heuristic *one sense per discourse* (Gale et al., 1992). In this method for WSD, it is assumed that a word appearing in a discourse refers to the same sense throughout the discourse. The approach for geographical referent disambiguation with this heuristic is to assume that a place name used in a discourse refers to the same location throughout the

discourse (Leidner et al., 2003). Obviously this heuristic is not error free and some texts (rarely) could include place names that are equal homonyms but refer to different places (e.g. “Georgia” (country) vs “Georgia” (US state)).

2. **Proximity of place names**. This approaches assumes that there is a high degree of spatial correlation in geographic references that are in textual proximity (Rauch et al., 2003).
3. **Spatial minimality heuristic**. This approach tries to disambiguate places assuming that the small region that is able to ground the whole set of places appearing in the discourse is the correct interpretation of these toponyms (Leidner et al., 2003). Buscaldi and Rosso (2008c), for instance, used GeoWordNet to implement a map-based disambiguation method based on the one proposed by Smith and Crane (2001).
4. **Contextual Pattern Matching**. Applying contextual patterns (e.g. “location1 at South of location2”, “city of X”) is the most widely used approach (H. Li et al., 2002; Rauch et al., 2003; Manov et al., 2003).
5. **Population heuristics**. Population data in geographical gazetteers is used in different ways: ignoring small places and/or promoting dense populated place (Leidner, 2007). Rauch et al. (2003) assumed that “A place with a high population is more likely mentioned than a place with a lower one”.
6. **Co-occurrence models**. H. Li et al. (2003) used discourse features based on co-occurring toponyms (e.g., a document with “Buffalo” , “Albany” and “Rochester” will likely have those toponyms disambiguated to New York State). S. Overell et al. (2006) applied co-occurrence models trained with Wikipedia for place name disambiguation. X. Wang et al. (2010) uses co-ocurrence of toponyms by computing semantic relations (metric, topological, and typological) and then using the Dempster-Shafer theory to select the correct candidate.
7. **Comma groups**. The study of M. Lieberman et al. (2010b) presents a set of heuristics to resolve toponym resolution of comma groups (three or more toponyms in a list) in news text. They found in their experiments that 49% of comma groups are resolved by the sibling heuristic in the geographic hierarchy, a 39% by population based prominence, and a 12% by distance-based proximity heuristic.
8. **Use of default and salience**. Some methods set a default location when a place name is ambiguous, the most common heuristic to decide the default place is the use of the candidate with the largest population.
9. **Topological and Knowledge based measures**. Buscaldi and Rosso (2008c) uses a Conceptual Density method based on WordNet. K. Roberts et al. (2010) shows the importance of event structures when disambiguating toponyms by using a a probabilistic model that only estimates probabilities of events. Bensalem and Kholladi (2010) proposed a new measure of geographical correlation called Geographical Density. Their toponym disambiguation heuristic is based on the arborescent proximity between toponyms (e.g. the proximity in the hierachical tree of the world places). Their results over the GeoSemCor corpus outperformed state-of-the art methods in the term of recall and coverage

10. **Machine Learning**. J. Santos et al. (2015) used the LambdaMART and Random Forest machine learning algorithms for some toponym disambiguation subtasks of English using both textual and geographical features. On the other hand, system such as Speriosu and Baldridge (2013) and DeLozier et al. (2015) used georeferenced language models learnt from georeferenced Wikipedia pages.

- **Reference ambiguity problem.** This problem occurs when the same location can have more than one name (in Spanish texts this frequently occurs as many place names occur in languages other than Spanish, as Basque, Catalan or Galician). Knowledge sources as GNS or TGN are used to deal with this problem. For instance, Ferrés and Rodríguez (2006a) applies a grouping process over GNS to create groups of place names that refer to the same locations. On the other hand, Leveling and Veiel (2006) implemented a metonymic location classifier trained with the manual annotated data from the GERMAN CONLL-2003 shared task. The classifier achieves a performance of 81.7% of F1-measure in differentiating between literal and metonymic senses of location names.

- **Referent Class Ambiguity**. The same name can be used for locations and also for other classes of Named Entities like persons or organizations. An example of this ambiguity can be the person name Paris (e.g. in "Paris Hilton") that can be erroneously tagged as a location when used without the Surname in some texts. Ferrés et al. (2005a) apply a NEC correction filter to correct the Person/Location ambiguity errors. This filter stores in a hash table all the tokens that compose the NEs classified as *person*. Then *location* or *organization* NEs are checked against the hash table. Z. Li et al. (2006) apply a set of rules for resolving the location-person ambiguity.

2.2.1 Toponym Recognition and Disambiguation Systems

This part presents some of the most relevant Toponym Resolution systems:

- Rauch et al. (2003) use data mining procedures and domain knowledge repositories (such as first names) to generate sets of contexts with positive or negative indicators. Positive context for geographic names could be trigger words before of after a name (e.g. "city", "mayor", "community college").

- S. Overell et al. (2006) applied co-occurrence models trained with Wikipedia for place name disambiguation with a Naive Bayes classifier.

- Garbin and Mani (2005) describes a corpus-based method for disambiguating toponyms with an unsupervised Machine Learning system that develops disambiguation rules. They used the ALTAS Gazetteer and the World Gazetteer and the LexScan tool. They used a Human Annotated Corpus of news (from TimeBank 1.2, and Gigaword NYT Sept. 2001 and June 2) (Section 5). This corpus contains 83,872 words with 1275 place names (435 distinct) annotated with 3 geographical classes: *national capital*, *civil politicaladministrative region*, and *populated place*. This method achieves a 78.5% of accuracy in the human-annotated corpus.

- Leidner (2006) presented the first systematic account of the utility of different heuristics for the toponym resolution task, based on experimental comparison on two novel large-scale gold-standard annotated corpora: *TR-CoNLL* (a gold-standard corpus of nearly 1,000 news articles from CoNLL 2003 with the correct referents annotated by humans) and *TR-MUC4* (an annotated corpus of 100 MUC-4 documents focused on Central America). Both corpora were annotated with these populated place classes: city, state, country, and continent. Leidner (2006) replicated two methods: Perseus (Smith and Crane, 2001) and LSW03 (Leidner et al., 2003) for a set of large-scale experiments. LSW03 outperformed Perseus in both corpora. LSW03 achieved 0.4736 and 0.4598 of Toponym Score (see the explanation of these evaluation metric at the end of this chapter) in TR-CoNLL and TR-MUC4 respectively. Perseus achieved 0.3431 and 0.4023 of Toponym Score in the same corpora.

- Yi Li et al. (2006) used a probabilistic approach for toponym resolution based on a five-level normalization of the gazetteer. Assigning more probabilities to the top levels (country or nations). Initial probabilities are also adjusted based on the following evidences: *local contextual information*: for example, geo-types in close proximity to each other (e.g. city, state), *population information*, *Trigger Words.* (e.g. "county", "river", etc.), *global contextual information*, occurrences in the document of country geo-terms that are gazetteer ancestors to the candidate, and *Mutual disambiguation*: Candidates that are closely related to each other in the gazetteer hierarchy boost each others' probability assignment for their respective terms. They used a hand annotated subset of the GeoCLEF corpus to determine the performance of the Named Entity Classification System, and the toponym disambiguation algorithm. The corpus consists of a set of 106 Glasgow Herald and 196 LA times news articles, which contained 2,311 tagged locations in total. LingPipe achieved a 50% of Precision and a 65% of recall . The TR algorithm achieved an accuracy of 90.3% on the 1502 place names identified by LingPipe. The disambiguation accuracy with respect to the total number of total locations achieved an accuracy of 60.8%.

- Buscaldi and Rosso (2008c) used two algorithms: a knowledge-based method and a map-based method evaluated over the GeoSemCor corpus and compared them. They used GeoWordNet to implement a map-based disambiguation method based on the one proposed by Smith and Crane (2001). The other method is Conceptual Density method based on WordNet. The Conceptual Density approach achieved better results than the Map-based evaluated at sentence, paragraph, and document level. The best results at document level in F-measure and their corresponding precision, recall and coverage were 0.832, 89.9%, 77.5%, and 86.2%.

- Bensalem and Kholladi (2010) proposed a Toponym Disambiguation system based on the arborescent proximity between toponyms (e.g. the proximity in the hierarchical tree of the world places). Their results over the GeoSemCor corpus outperformed state-of-the art methods presented by Buscaldi and Rosso (2008c) in the term of recall (87.4%) and coverage (99%) at document level.

- M. D. Lieberman and Samet (2012) presented an approach that used adaptative context features for toponym resolution in streaming news. They use a POS tagger and a NER package to recognize toponyms. For toponym disambiguation they use Geonames and Random Forests classifiers based on adaptative context features. The approach was evaluated with the following datasets: ACE, LGL, and CLUST.

- Speriosu and Baldridge (2013) used the non-toponym textual context at local or document level for toponym disambiguation. English NER models from the OpenNLP project are used for toponym recognition. For toponym disambiguation text-driven models were created using Geonames and geotagged Wikipedia articles. Their results on the TR-CoNLL and CWar datasets show that their text classifiers are more accurate than knowledge-based algorithms based on spatial proximity or metadata.
- Habib and Keulen (2013) presented a toponym extraction and disambiguation language-independent approach based on Hidden Markov Models (HMM), Geonames, and Support Vector Machines (SVM) over a dateaset of the EuroCottage portal in English, German, and Dutch. An HMM model was used to extract the toponyms that are then filtered using Geonames. Then a disambiguation process is applied using a clustering approach. Finally the SVM is used to classify between true positives and false positives.
- J. Santos et al. (2015) used machine learning for toponym disambiguation of English corpus of SpatialML, LGL and a subset of Wikipedia. They used the Stanford NER (3.2.0) for toponym extraction and a Wikipedia based Knowledge Base for toponym disambiguation combined with ML. Then a disambiguation process is applied with the following steps: 1) Query expansion to detect altenative names of the toponym in the text, 2) candidate generation: searches for entities with string similarity with the query (toponym) in a KB, 3) candidate ranking using the LambdaMART learning to rank algorithm with several textual and geographical features such as: candidate count, population count, geospatial (area,containment and distance), distance to the closest reference, area of the geometric hull, Jaccard similarity between geographical entities, missed geographical entities, 4) candidate validation using a Random Forest classifier with the same features of the candidate ranking.
- DeLozier et al. (2015) modeled the geographic distributions of words using georeferenced language models learnt from the GeoWiki (a subset of georeferenced Wikipedia pages). They also can use the Geonames Gazetteer and the Natural Earth gazetteer in combination with the georeferenced models. They used the TR-CoNLL, CWar and LGL corpus to perform an evaluation of the approach.
- Spitz et al. (2016) built a network of place similarities based on the text of the English Wikipedia that includes entity linking to WikiData for toponym disambiguation. They used the AIDA CoNLL-YAGO dataset to evaluate the approach.

2.2.2 Toponym Disambiguation Evaluation

Toponym Disambiguation is evaluated with test collections specifically annotated with correct referents and metrics adapted from Natural Language Processing task such Word Sense Disambiguation and NERC, usually precision, recall and F-1 although finer metrics are used for more specific tasks (Leidner, 2006). Some of these test collections are the following:

- **ACE corpus annotatted with SpatialML** SpatialML (Mani et al., n.d.) is an annotation scheme for marking up references to places in natural language. It covers both named and nominal references to places, grounding them where possible with geo-coordinates, including both relative and absolute locations, and characterizes relationships among places in terms of a region calculus.[29] A corpus of 428 ACE documents, originally from the University of Pennsylvania Linguistics Data Consortium (LDC), has been annotated in SpatialML. This corpus, drawn mainly from broadcast conversation, broadcast news, news magazine, newsgroups, and weblogs, contains 6338 PLACE tags, of which 4,783 are named PLACEs with geo-coordinates. This ACE SpatialML Corpus (ASC) has been re-released to the LDC, and is available to LDC members (LDC Catalog LDC2008T0313).

- **TR-CLEF** The TR-CLEF corpus was created by Andogah (2010) from the relevant documents of GeoCLEF 2006 evaluation contest. The current version of the TR-CLEF corpus consists of a subset of 321 documents relevant to GeoCLEF 2006 campaign topics. The place names in these documents have been resolved (5,783 toponym instances). These documents contain a total of 802 unique references to places with a 86% of references ambiguous.

- **GeoSemCor** GeoSemCor (Buscaldi and Rosso, 2008c) was obtained from SemCor, the most used corpus for the evaluation of WSD methods. SemCor is a collection of texts extracted from the Brown Corpus of American English, where each word has been labelled with a WordNet sense (synset). It contains 1,210 toponym instances in its final version.

- **LGL corpus** LGL (Local/Global Lexicon) (M. Lieberman et al., 2010a) is a corpus of 588 articles collected from 78 different data sources, containing 4,793 toponyms. This corpus is based on smaller newspapers with a localised audience.

- **TR-CoNLL** The TR-CoNLL corpus (Leidner, 2007) contains 946 news articles from the CoNLL 2003 shared task, in which 6,980 toponym instances have been annotated in TRML (Toponym Resolution Markup Language), an XML-based markup language created by Leidner. This data set can be used as a gold standard to evaluate automatic systems that can do toponym resolution.

- **TR-RNW** The TR-RNW corpus is derived from the Radio Netherlands Worldwide (RNW) summaries in English (Andogah, 2010). The TR-RNW consists of 556 news summaries from RNW. This corpus contains 2,339 toponym instances (432 are unique toponyms). The 76.5% of TR-RNW unique toponyms are ambiguous.

- **TR-MUC4** The TR-MUC4 (Leidner, 2007) is a corpus of 100 documents form the 4th Message Understanding Contest (MUC-4) (Sundheim, 1992). The collection is made up of intelligence reports covering Central America.

- **RCV1** The Reuters Corpus Volume 1 (RCV1) (Lewis et al., 2004) is an archive of 806,791 English language news stories coded for topic, region (geography) and industry. The type of regions referred to in a story can be: countries (e.g., UK), geographical groups (e.g., BENELUX), and, economic groupings (e.g., G7).

- **AIDA CoNLL-YAGO**: (Hoffart et al., 2011) is collection of 1393 news documents from the Reuters RCV-1 collection annotated with YAGO2 entities.

[29] `http://sourceforge.net/projects/spatialml`

- **CWar**: the Perseus Civil War and 19th Century American Collection (CWAR)[30] contains 341 books written primarily about and during the American Civil War (Crane, 2000). The corpus was annotated using a semi-automated process (Speriosu and Baldridge, 2013).
- **CLUST**: (M. D. Lieberman and Samet, 2011) is a corpus with articles from different news sources.

[30] http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:cwar

CHAPTER 3

Geographical Information Access Tasks - State-of-the-art

This section describes the state of the art of the three Geographical Information Access tasks faced in this thesis: 1) *Geographical Information Retrieval*, 2) *Geographical Question Answering*, and 3) *Textual Georeferencing.*

3.1 Geographical Information Retrieval - State-of-the-art

Geographical Information Retrieval (GIR) consists in searching documents with geographically restricted queries. GIR queries consist in requests that involve both thematic and geographic search (e.g. “rice exportation in Japan” or “shark attacks in California”). In Sanderson and Kohler (2004a) a geographic query is defined as: “A query which includes at least one of the following types of geographic terms: place names (e.g. Houston, Texas, US); other locators (e.g. postcode, ZIP code); adjectives of place (e.g. American, international, western); terms descriptive of location (e.g. state, country, city, site, street); geographic features (e.g. island, lake); and directions (e.g. north, south). “. GIR systems have very specific issues due to its restricted domain (geography) specificity. Some of these issues have been detailed in the GIR literature (C. B. Jones and R. S. Purves, 2008):

- geographical names detection (toponym recognition).
- spatial natural language qualifiers detection (e.g. “north”, “south of”, “near”, “close by”,...).
- toponym disambiguation (e.g. Paris, Texas (USA) vs Paris (France)).
- vague place names detection and interpretation. (e.g. “Scottish Trossachs”, “Midlands”,...).
- thematic and geospatial indexing and retrieval.

Since earlier 2000s GIR has become a popular task in the IR community mainly for these reasons: 1) the inclusion of new GIR evaluation benchmarks and GIR workshops in several international IR conferences, 2) the inclusion of geographic search in major search engines (C. Jones and R. Purves, 2005), 3) the publication of some PhD theses in GIR, and 4) the emergence of research projects in GIR.

The organization of several GIR evaluation benchmarks started initially with the inclusion a the GeoCLEF GIR track in CLEF 2005, 2006, 2007 and 2008 (F. Gey et al., 2005; F. Gey et al., 2006; Mandl et al., 2008; Mandl et al., 2008), and then with GeoTime statio-temporal GIR track in NTCIR in 2008 and 2009 (F. C. Gey et al., 2010; F. C. Gey et al., 2011). The GeoCLEF GIR evaluation forum took place during 4 years (1 as a pilot task) between 2005 and 2008 in the framework of the CLEF conferences[1]. The GeoCLEF contest evaluated GIR for monolingual and bilingual experiments in the following languages: English (in all the evaluations from 2005 to 2008), Portuguese (from 2006 to 2008), Spanish (2006), German (2005 to 2008), Japanese (only in topics, not in collections) in 2006. On the other side, the International Workshop on Geographic Information Retrieval (GIR) has been organized in several international IR conferences since 2004[2].

The following PhD theses related with GIR have been recently published: 1) Leidner (2007) proposed heuristics and algorithms, collections and metrics for Toponym Disambiguation, 2) Martins (2008) proposed Geographical Scope Detection of documents with page rank , 3) S. Overell (2009) mined and extracted and identified useful knowledge from the Wikipedia and used it to GIR and TD, 4) Andogah (2010) proposed methods for Geographical Scope Resolution and GIR, 5) Buscaldi (2010) applied TD algorithms for GIR, and 6) J. M. Perea-Ortega (2010) used query reformulation and geographical document re-ranking to improve GIR, 7) Villatoro-Tello (2010) proprosed a GIR re-ranking strategy using a set of example documents for relevance feedback, 8) Palacio (2010) presented and evaluated an approach that combines special indexing for spatial, temporal and thematic dimensions for GIR.

Some well-known past research projects and systems related with GIR were:

- *GIPSY*, and earlier GIR system that indexed text using spatial coordinates derived from knowledge-bases (Woodruff and Plaunt, 1994).

- *Web-a-where*, a system that performed Toponym Detection and Disambiguation and assigned to each page a geographic focus (Amitay et al., 2005).

- *SPIRIT*[3] (Spatially-Aware Information Retrieval on the Internet) was a research project (funded through the EC Fifth Framework Programme) that was engaged in the design and implementation of a search engine to find documents and datasets on the web relating to places or regions referred to in a query (C. Jones et al., 2002; C. B. Jones et al., 2004; R. S. Purves et al., 2007).

- *DIGMAP* was a project focused on historical digitized maps and provided access to historical cartography (Martins et al., 2007a).

- *STEWARD* was a spatio-textual search engine that allowed queries and visualization of textual references to geographic locations in unstructured text documents[4] (M.

[1] http://www.clef-initiative.eu

[2] http://www.geo.uzh.ch/~rsp/gir14/index.html

[3] **SPIRIT project**. http://www.geo-spirit.org

[4] STEWARD online demo. http://steward.umiacs.umd.edu/

Lieberman et al., 2007).

- *PIV.* The Virtual Itineraries in the Pyrenees (PIV) project was a web-based platform with an in-depth geo-semantic identification (Gaio et al., 2008).

- *NewsStand.* is an existing search engine that does Toponym Recognition and Disambiguation in news and plots them in a world map[5] (Teitler et al., 2008).

3.1.1 GIR Issues

GIR requires appropriate indexing and search structures and algorithms to determine both thematic and spatial relevance. Spatial relevance similarity measures require that the following aspects had to be taken into account: hierarchical containment, adjacency of places, connectivity, and proximity. Different multi-dimensional indexing approaches have been proposed to manage spatial data: such as grid indexes, quad-trees, R-trees, and k-d-trees (Martins et al., 2005). R-Tree, that allows efficient geographical search, is the most popular spatial indexing method (Guttman, 1984). On the other hand, c-squares is a grid indexing approach that uses grid representation and can be encoded in textual strings (Rees, 2003), so it could be easily implemented in a normal IR system.

On the other hand, Geographical Knowledge and Reasoning is required to deal with geographical resolution problems. These are the common issues related with the use of Geographical Knowledge for IR:

- *Using efficiently Geographic Knowledge in GIR queries.* F. Gey et al. (2005) reported on deteriorated performance when applying manual query expansion of geographic references. Guillén (2005) concludes that adding geographic information in the queries could not significantly improve retrieval performance. Metacarta (Kornai, 2005) improved its results using geographic bounding boxes, but with a bit low MAP. As reported by Toral et al. (2006) in GeoCLEF 2005 three of the top-4 systems for the English monolingual run were based only on IR (the remaining one used geographic NERC).

- *Person-Location ambiguity problems.* It is common for proper name of persons and places to be the same and this leads to potential false associations between articles mentioning persons with such name and particular places.

- *Multilinguality.* (i.e. handling toponyms in different languages). In many gazetteers, mostly English names are used.

- *Name Variants.* Leveling et al. (2005) defined these variants: 1) endonymic names: a local name for a geographical entity (e.g. “Wien”, “Köln”, and “Milano”), 2) exonym names: is a place name in a language used outside its region; (e.g. “Viena”, “Cologne”, and “Milán”), and 3) historical names: traditional names such as “New Amsterdam” for “New York”.

- *Multiword Names.* Two or more words form the place name (e.g. “Mount Cook”, “Island of Sylt”).

[5] `http://newsstand.umiacs.umd.edu/web/`

- *Semantic relations between toponyms and related concepts.* Concepts related to a toponym such as the language, inhabitants of a place, or other kind of phrases are not considered in geographic tagging (i.e. expressions such as “latin american countries”, “French speaking part”, “catholic communities”,...).
- *Temporal changes* in toponyms.
- *Metonymic usage.* Metonymy is defined as a figure of speech in which a speaker uses “one entity to refer to another that is related to it” (Leveling and Veiel, 2006). As an example in the following sentence: “At the meeting of France and Germany in Lisbon last year, Paris vetoed the decision”, Paris is a metonymy of France, and France and Germany are a metonymy of the French and German governors.
- *Query expansion.* Adding terms to the original query in order to increase the retrieval performance can lead to obtain additional relevant documents (i.e. increasing Recall), possibly at the expense of Precision.
- *Gazetteers problems.* Incompleteness of the major gazetteers. Fonseca et al. (2002) discussed about the problems of selecting and using gazetteers. As an example, the GeoNet Names Server geographical gazetteer presents the following problems: i) highly ambiguity on some names, ii) geographic entities that have a certain area/length (like rivers or large cities) but only a single latitude/longitude pair is given (Hauff et al., 2006), iii) bad data (out of range longitude/latitude pairs, parent information can overlap or is not fully accurate), iv) lack of data (Leveling et al., 2005) (e.g. lack of native language forms). v) relations or modifiers generate name variants not covered by a gazetteer (e.g. Southern Germany), vi) data representation may be inconsistent. (e.g. some streams or rivers are represented with only one point), vii) it does not provide sufficient information for a successful disambiguation from context (e.g. temporal information is missing), viii) incomplete ontological basis, and ix) uncovered mame inflection. Ahlers (2013) described some problems with the Geonames Gazetteer accuracy: including inaccuracies ranging from grid patterns, imprecise coordinates, overlaps, repetitions, and misclassifications.

3.1.2 GIR Approaches

Major approaches in GIR (F. Gey et al., 2005) include: adhoc techniques, QA modules, Gazetteer construction, Geoname Entity Extraction, Term expansion using WordNet, geographical thesauri, toponym resolution, NLP-Geofiltering predicates, latitude-longitude assignment, gazetteer based query expansion, conventional IR systems, geographic entity recognition, Knowledge Bases, query expansion strategies (e.g. blind feedback, addition of proper names, geographic reference expansion using hierarchical information on GKB), and geo-spatial query restriction strategies: minimum bounding box based, geo-scope based.

Despite of the diversity of approaches at GIR, two major phases are usually present in all the system architectures: Topic and Collection Processing and Document Retrieval. A generic system architecture of a GIR system is shown in Figure 3.1.

3.1.2.1 Topic Processing and Collection Processing

Topic and Collection Processing consists in analyze the topics and/or documents of the collection in order to enrich them with useful information derived from Natural Language

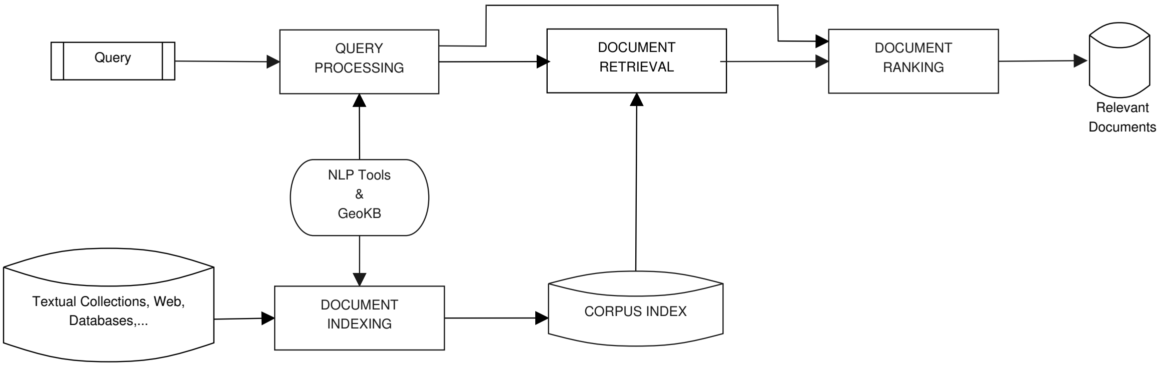


Figure 3.1: Generic architecture of a Geographical IR system.

Processing and/or Geographical Analysis.

3.1.2.1.1 Natural Language Processing Natural Language Processing (NLP) in GIR normally consists of applying linguistic analysis over the topics and/or the document collection for lexical purposes. Semantic parsing and lexical databases are rarely applied. Lexical analysis for GIR normally deals with NERC in order to detect place names. POS tagging is applied in most systems because sometimes is required for the NERC to have useful features. NERC approaches applied in GIR include both Machine Learning approaches and Rule-based ones: such as GATE (S. Overell et al., 2008a), Alias-I LingPipe (Andogah and Bouma, 2008; Buscaldi and Rosso, 2008a; Kölle et al., 2008; J. Perea-Ortega et al., 2008a), ABIONET Ferrés et al. (2005a), Stanford NER system, (Buscaldi and Rosso, 2008b). Buscaldi et al. (2005) used the WordNet ontology in the geographical domain, by applying a query expansion method, based on the synonymy and meronymy relationships, to geographical terms. Buscaldi et al. (2006) used also WN to perform an index expansion based on synonymy and holonymy relations. Leveling and Veiel (2006) employed multilayered extended semantic networks for the representation of knowledge, queries and documents for GIR with a syntactico-semantic parser (WOCADI).

3.1.2.1.2 Geographical Analysis Geographical Analysis of the topics and documents may consists on using Geographical Knowledge Bases (GKB) and Toponym Resolution algorithms. GKBs are used in order to detect geographical place names and its possible referents. Toponym Resolution is applied to decide which referent is used in a certain context.

Geographical Knowledge Bases. Geographical Knowledge Bases can be defined as geospatial dictionaries of geographic names with some relationships among place names. Usually these places can be political and administrative areas, natural features, and man-made structures. Relationships among place names are commonly downward (parent-child) relations (e.g. Asia - China) and upward (e.g. Germany - Europe). On the other hand some approaches define other relationships, Hu and Ge (2006) GKB includes relationships between entities such as part-of adjacency and similar (e.g. if two entities have a similarity such as being administrative divisions of the same country or if they are countries, ...). Lana-Serrano et al. (2006a) provided a flexible structure that allows define other types of relationships between resources: based on its languages ("latin america", "anglo-saxon countries") or religion ("catholic countries", "protestant towns",...).

The most commonly used GKBs in GeoCLEF evaluations are publicly available huge gazetteers such as: GeoNet Names Server, Geonames, GNIS, and WorldGazetteer. WorldGazetteer was widely used due to its population statistics (Cardoso et al., 2005; Leidner, 2005; Ferrés and Rodríguez, 2006b; F. Gey et al., 2005). Some approaches used the Wikipedia to collect information (Cardoso et al., 2005). S. Overell et al. (2006) and Yi Li et al. (2006) used the Getty Thesaurus of Geographic Names, a propietary Gazetteer.

GIR systems often tend to merge some these gazetteers into a unique one. Hauff et al. (2006) used a merge of GNS, GNIS, and World Gazetteer (WG), that provides information about the parent-child relationships. Andogah (2006) used geographic resources such as Wikipedia, World-Gazetteer, GeoNet names server, and WordNet. Hu and Ge (2006) joined several resources to build a GKB: a) FIPS 10-4 for countries and administrative divisions, b) World Factbook for border countries, coastlines, country capital cities, c) Wikipedia for oceans, seas, gulfs, rivers and regions, d) A set of large cities collected from TravelGis.com, e) The Standard Country and Area Codes Classifications (M49) for regions and continents, f) The ESRI Gazetteer server developed by the Environmental Systems Research Institute, Inc. for Minimum Boundary Rectangle (MBR) of countries, and g) WordNet for variant places names. Toral et al. (2006) used Geonames DB[6].

Toponym Disambiguation in GIR. Toponym Disambiguation is used in several GIR approaches. TR Algorithms usually decide the best referent candidate among a set of possible referents for a place name applying a set of heuristics (see Chapter 2 for a detailed explanation of the TR methods). Cardoso et al. (2005) at GeoCLEF 2005 and Martins et al. (2006) at GeoCLEF 2006 used the one single scope per document heuristic (Martins and Silva, 2005) with a PageRank variation graph based algorithm. Leidner (2005) used a maximum-population heuristic. S. Overell et al. (2006) applied co-occurrence models trained with Wikipedia for place name disambiguation with a Naive Bayes. Yi Li et al. (2006) used a probabilistic approach for toponym resolution based on the following evidences: *local contextual information*, *population information*, *Trigger Words*, *global contextual information*, and *Mutual disambiguation.*

Leveling and Veiel (2006) implemented a metonymic location classifier training with the manual annotated data from the GERMAN CONLL-2003 shared task and a subset of the GeoCLEF newspaper corpus. The features used were shallow (PoS tags, position of words in a sentence, word length and base forms of verbs). The classifier achieved a performance of 81.7% of F1-measure in differentiating between literal and metonymic senses of location names.

On the other hand few systems apply geo-disambiguation to resolve the *person- organization - location* ambiguity (i.e distinguish if the candidate was correctly tagged as a toponym or is really a person name or an organization name). Ferrés et al. (2005a) apply a NEC correction filter to correct these errors. Z. Li et al. (2006) apply a set of rules for resolving the location-person ambiguity.

[6] Geonames. `http://www.geonames.org`

3.1.2.2 Document Retrieval

The main goal of this phase is to retrieve a set of relevant documents to the topic. The main process of this phase is the Information Retrieval process which usually requires the use of an IR system. This phase can be complemented by a Query Expansion phase and a post phase of Document Filtering.

3.1.2.2.1 Query Expansion Query Expansion (QE) techniques in IR usually consist in adding related terms to the query manually or automatically in order to retrieve more relevant documents. In GIR is also normal to use conventional IR QE techniques in order to modify the thematic search. For instance, García-Vega et al. (2006a) performed a thesaurus-based expansion using words with a high rate of document co-occurrence. But for geographical IR, usually terms geographically related to the topic terms are added to the query. The GIR QE can be done guided by several heuristics based on spatial relations and location type.

Before QE, the desired keywords are extracted to compose the query. Some approaches apply special algorithms for this Query Processing or Query Parsing step. Toral et al. (2006) collected required words and geographical items. Required words are all the nouns of the topic, description and narrative without geographic ones, stopwords and guidance information,

Sometimes document expansion is applied previously. Document expansion and query expansion techniques are used to match the location in a query to all its gazetteer children and nearby locations. Yi Li et al. (2006) used a geographic-based query expansion, using a gazetteer to extend geospatial terms to "nearby" locations, and included sublocations. A geo-term in the query may be expanded upwards (for "close/near " relations, influencing all or some of its ancestors) or downwards (for "in" relations, extending the influence to all of its descendants in the gazetteer hierarchy).

Leidner (2005), Buscaldi et al. (2005), and Leveling and Veiel (2006) applied query expansion with meronyms (e.g. for California, "Orange County" and "Los Angeles" are included), and Toral et al. (2006) and García-Vega et al. (2006a) used automatic query expansion consisting in expanding the locations of the topics with geographical information from Geonames gazetteer. Leveling and Veiel (2006) also employed multilayered extended semantic networks for the representation of knowledge, queries and documents for GIR. Geographical concepts from the query network are expanded with semantically connected via topological, directional, and proximity relations.

3.1.2.2.2 Information Retrieval Information Retrieval approaches for GIR often use combined search (i.e. both thematic and geographical search). There are few systems that do not use Geographical Knowledge (GK) in IR (F. Gey et al., 2005; Guillén, 2005; Guillén, 2006; Toral et al., 2006). But some of these systems, based only in pure IR techniques achieved the top-ranked results in the GeoCLEF evaluations of 2005 and 2006, and competitive results in the following GeoCLEF evaluations in 2007 and 2008.

Boolean models are rarely used for GIR, if used, they are only used for geographical searches (Ferrés and Rodríguez, 2006b; Bischoff et al., 2006). Most of the IR engines at GeoCLEF are based on the Vector Space Model (Lucene, SMART, Zettair, etc.) or Probabilistic frameworks (Lemur (Indri), Terrier, Zapian, etc). Lucene with a TFIDF weighting scheme is used frequently by many approaches (Leidner, 2005; Buscaldi et al., 2006; Andogah, 2006). This system is preferentially used for thematic search rather than for geographical search. The Lemur toolkit (Indri) was also used for several systems (Guillén, 2006; García-Vega et al., 2006a; Hauff et al., 2006). Passage Retrieval was used by few approaches: Ferrés and Rodríguez (2006b) used JIRS for thematic and geographical search, and Ferrández et al. (2005) and Toral et al. (2006) used IR-n. On the other hand also RDBMS systems were used specially for geographical isolated search (i.e. queries with only geographical terms): Postgres (used by S. Overell et al. (2006)), MySQL (used by Hu and Ge (2006)), and Toral et al. (2006) used SQL queries over the Geonames DB.

Normal textual indexing is vastly used for all the systems. Some of them take profit of the "field search" capabilities of some IR search engines. Several systems indexed separately textual terms and geographical terms. Andogah (2006), for instance, indexed and searched separately geographical relevant terms (place names, geo-spatial relations, geographic concepts and geographic adjectives) and thematic terms. Ferrés et al. (2005a) and Yi Li et al. (2006) used hierarchically expanded geo-terms indexing (i.e. a concatenated string consisting of a candidate and its ancestors in the gazetteer). Z. Li et al. (2006) performed this idea with a different way: utilizes the inverted index to store all the explicit and implicit locations of documents.

Other systems employed indexing structures specially designed for Geographical IR: R-Tree structures were used by S. Overell et al. (2006), Z. Li et al. (2006) used grid indexing with a textual index IR engine dividing the surface of the earth into 1000x2000 grids, and Kornai (2005) used the Metacarta search engine with a bounding box derivation scheme.

Relevance Feedback (RF) (which consists in performing a new retrieval loop with a set of manually or automatically collected terms from the initial retrieved documents) has emerged as a efficient method for improving the results in GIR. Systems such as Guillén (2005) and F. Gey et al. (2005) achieved the best results at GeoCLEF 2005 using Relevance Feedback (RF) techniques with a Probabilistic IR approach, and Ferrés and Rodríguez (2008a) achieved the top-ranked results at GeoCLEF 2007 with an approach that includes RF among other features. Term weighting schemas applied for GIR systems are: TF-IDF, BM25, DFR, and Boolean. TF-IDF and Okapi's BM25 are the most widely used. Preprocessing techniques such as stemming, stopwords removal are extensively used in most of the systems. Porter's stemmer in combination with the SMART stop words list are used in some GIR systems (Guillén, 2005; Ruiz et al., 2006), etc. But other approaches use lemmatization in combination with stemming (Ferrés and Rodríguez, 2008a).

3.1.2.2.3 Document Filtering Document filtering strategies for GIR try to filter out geographically irrelevant documents by using GKBs. Hauff et al. (2006) retrieved by content and subsequently filtered by geographical relevance using a gazetteer and coordinates restrictions. Leidner (2005) used Minimal bounding Rectangles (MBR) to approximate the polygons described by the locations in the query.

3.1.2.3 Document Ranking

The Document Ranking (DR) phase consists in combining scores from thematic search and geographic search (i.e. geographically isolated terms search). Relevant approaches include linear interpolation (Leidner, 2005; Andogah, 2006) and geographic similarity ranking (Martins et al., 2006).

3.1.3 IR Evaluation Measures

IR Evaluation Measures indicate how fast (indexing and searching speed) and how well (efficiency and size index) the process is done. The measures that show the the speed of indexing and retrieval are called *efficiency measures.* There are various ways to measure how well the retrieved information matches the desired information. These ways are called *effectiveness measures* (Rijsbergen, 1979). The effectivenes measures involve the use the following information based on the Cranfield methodology (Cleverdon and Keen, 1996) commonly used in evaluation forums:

- *corpus of documents*: set of documents in which the user wants to find relevant documents.
- *topics*: set of user query needs. In TREC-style evaluations normally contains the following parameters: title, description and narrative.
- *relevance judgments*: a set of relevant documents associated to each topic. These relevant documents had been marked as relevant by human assessors. Pooling is usually used in TREC-style evaluations to have a set of documents (provided by the top-ranked documents of each participant system) to be reviewed by the assessors without having to judge the whole corpus.
- *system output*: a set of ranked documents associated to each topic.

Some of the most used IR effectiveness measures are reported here:

- *Precision.* The proportion of retrieved and relevant documents to all the documents retrieved:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{3.1}$$

  Precision can also be evaluated at a given cut-off rank, denoted *P@n*, instead of all retrieved documents.

- *Recall.* The proportion of relevant documents that are retrieved, out of all relevant documents available:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{3.2}$$

- *F1-measure.* The weighted harmonic mean of precision and recall.

$$F = \frac{PR}{2P + R} \tag{3.3}$$

- *Precision at N.* This measure computes the precision after N retrieved documents (N can be 1,5,20,100,...).

- *Average Precision.* Average Precision is the average of the precision after each relevant document is retrieved. Where r is the rank, N the number retrieved, rel() a binary function on the relevance of a given rank, and P() precision at a given cut-off rank:

$$AveP = \frac{\sum_{r=1}^{N}(P(r) \times \mathrm{rel}(r))}{\text{number of relevant documents}} \tag{3.4}$$

- *Mean Average Precision (MAP).* MAP is the most used effectiveness measure in evaluation exercises and IR research. Over a set of queries, find the mean of the average precisions, where Average Precision is the average of the precision after each relevant document is retrieved.

- *R-Precision (R-Prec).* R-Precision is the precision computed after R retrieved documents, being R the total number of relevant documents for the query. The average R-Precision for a set of queries is the mean of the R-Precision of all the queries.

3.1.4 Systems Comparison and Statistical Testing

In order to compare IR systems two or more effectiveness measures can be used. The difference in effectiveness can be influenced by several important factors such as: the relevance assesments, the evaluation measures themselves, the document collection size, and the topics.

T. Jones et al. (2014) reported good evidence that collection size and document source have and strong influence in comparing IR systems by showing that different collection subsets can produce different evaluation results. Urbano et al. (2013) studied the reliability of 43 TREC test collections evidencing that some of them are very little reliable. They showed that the ideal topic set size varies significantly across tasks and the traditional choice of 50 queries is not enough for stable rankings. Despite these factors influencing the difference in effectiveness among different IR systems, a realiable conclusion to rank systems should have into account the possibility of the difference in efectiveness is not real and could be produced by random chance. Statistical significance testing methods are employed in IR to get reliable conclusions to compare systems. Sanderson and Zobel (2005) compared the properties of the t-test, Wilcoxon and sign test concluding that the use of the t-test and the Wilcoxon tests allowed for more accurate prediction over the sign test of which run was better. According to Smucker et al. (2007) the most commonly used tests in IR (in 2007) were: Student's paired t-test, the Wilcoxon signed rank test and the sign test. Smucker et al. (2007) recommended to use the randomization test. Although, statistical significance tests are the most popular data analysis in IR they had been criticised in several ways that include over-use and mis-use (Gigerenzer, 2004). As reported by Ellis (2012): "A statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically nonsignificant and important."

3.1.5 GIR Relevant Approaches

This subsection reports some relevant approaches for GIR. Some special relevance has been given to the systems that achieved top-ranked results at official GeoCLEF evaluations or have used effectively its test collection in posterior experiments. These approaches use the different strategies to perform: 1) stand-alone probabilistic models (R. Larson et al., 2006), 2) combination of textual and geographical search (Martins et al., 2007b; Andogah, 2010; Zaila and Montesi, 2015), 3) filtering or reranking without geographical knowledge (Villatoro-Tello et al., 2010) 4) filtering or reranking the documents with geographical knowledge (Ferrés and Rodríguez, 2008a; Cumbreras et al., 2009; J. Perea-Ortega et al., 2011; Ferrés and Rodríguez, 2015a), 5) geographical query expansion (Buscaldi and Rosso, 2011; R. Wang and Neumann, 2008; Cumbreras et al., 2009), and 6) Machine Learning techniques for re-ranking (Martins and Calado, 2010).

Markowetz et al. (2005) presented a geographic search engine prototype for web pages of Germany using a grid indexing approach and a ranking that combines textual and geographic scores. R. Larson et al. (2006) used a logistic regression algorithm. Their system achieved the highest result with a MAP of 0.3936 at GeoCLEF 2005 in a run that used the spatial tags included in the topics. Martins et al. (2007b) presented a GIR sytem at GeoCLEF 2006 that used a geographical ontology. They used this ontology combined with a graph-ranking approach to detect scope of documents and topics and a relevance ranking that combined BM25 and a geographical similarity function for scopes. R. Wang and Neumann (2008) applied an approach that, besides including geographical knowledge, also included knowledge of natural and human events mined from Wikipedia. They use Query Expansion with ontologies both for events and geographic terms. Their system achieved the best MAP at GeoCLEF2008 with a 0.3037 with a run with manual work and a MAP of 0.2924 in an automatic run.

Cumbreras et al. (2009) compared the query-expansion and document filtering approaches for GIR. the toponym recognition phase used GATE for recognition and Geonames Gazetteer to verify. They used manual rules to find spatial relationships. Indexing was performed with Lemur creating a document index collection with stems and a geographical index with locations (continents, countries, cities, and other places). Both query expansion and document filtering approaches take into account both the query location type and the spatial relationship. They used the topics of GeoCLEF 2006 and GeoCLEF 2007 to evaluate the system. The document filtering approach outperformed the query-expansion but achieves similar results compared with the baseline BM25 with Pseudo-Relevance Feedback.

Palacio et al. (2010) presented the PIV GIR system, a system that combines 3 indexing-searching dimensions for GIR: geographic, temporal and thematic. They used functions that combine the results of the 3 dimensions, the Comb* functions and Borda count. The evaluation was performed over the MIDR_2010 test collection showing a 73.9% of improvement over the baseline.

Villatoro-Tello et al. (2010) presented and approach to re-rank the GIR results with Markov Random Field model that combines: 1) the original ranking, 2) the similarity between documents, and 3) a relevance feedback approach with full documents. The evaluation was done with the GeoCLEF dataset and queries of the 2005,2006,2007, and 2008 editions. The results showed an improvement of performance with respect to the baseline (TF-IDF with Lemur IR software) using 1, 5 or 10 documents for relevance feedback. The experiments using 10 documents achieved MAPs of 0.5910 (GeoCLEF2005), 0.6942 (GeoCLEF2006), 0.4960 (GeoCLEF2007) and 0.4959 (GeoCLEF2008), thus outperforming the best official results obtained at GeoCLEF evaluations.

In his PhD thesis, Andogah (2010) presented experiments combining non-geographic and geographic relevance scores. They used linear interpolation and weighted harmonic-means. The harmonic mean based combination achieved the best performance achieving a MAP of 0.2935 on the GeoCLE2007 dataset, outperforming the best official results at GeoCLEF 2007 (the best MAP was 0.2850).

Buscaldi and Rosso (2011) applied the GeoCLEF (2005-2008) topics to test diversity in GIR. They reformulated queries using the meronyms of the places contained in the original queries (using only the title field), with the help of a geographical ontology. J. Perea-Ortega et al. (2011) using the GeoCLEF data showed that in each evaluation a re-ranking based on the combination of geographical similarity and textual similarity outperforms the baseline (textual based IR). A geographical index was built with Geo-NER to recognize geographical entities. The textual index uses the stemmed and stopwords filtered text and the geographical entities in its original word form. They applied Lemur[7], Terrier[8] and Lucene[9] for the IR process. Lemur was applied with BM25 with Pseudo Relevance Feedback, Lucene with BM25 and Query Expansion, and Terrier with InL2 and Bo1 (Query Expansion).

Zaila and Montesi (2015) presented a GIR system with these tools: 1) GeoNW, a geo-ontology for toponym recognition based on GeoNames, WordNet and Wikipedia; 2) a toponym disambiguation algorithm based on candidates feature type frequencies for geo-class disambiguation between physical and administrative feature types and hierarchical distances based on GeoNW between toponyms in text for referential disambiguation, 3) dcoument geographic focus detector based on topological and distance influences and the geographic frequencies, 4) query geographical focus detection, and a 5) spatial similarity measure. The ranking is performed with a combination function that uses textual and spatial similarity. This function benefits documents retrieved in both lists and penalizes those that were retrieved only by their geographical information. The evaluation of the GIR system was performed with 25 topics of the GeoCLEF 2008 using the Terrier IR sofware with these three baselines: BM25, DLH13 and LGD models. The best results reported were obtained with the combination of textual and spatial similarity and the LGD approach with MAP values of 0.489. This result outperforms the best MAP results at GeoCLEF2008 (0.3037).

3.1.6 GIR Evaluation

Geographical Information Retrieval has been recently evaluated in some evaluation benchmarks such as GeoCLEF and GeoTime. This Section describes the GeoCLEF GIR evaluation task and reports the results of GeoCLEF evaluations from 2005 to 2008.

3.1.6.1 GeoCLEF

GeoCLEF was an IR task at the CLEF evaluation framework that evaluated GIR from 2005 to 2008. GeoCLEF started as a cross-language geographic retrieval task at the CLEF 2005

[7] `www.lemurproject.org/`

[8] `http://www.terrier.org`

[9] `http://lucene.apache.org`

campaign (F. Gey et al., 2005). The goal of the GeoCLEF task was to find as many relevant documents as possible from the document collections, using a topic set. Topics at GeoCLEF 2005 were textual descriptions with the following fields: title, description, narrative, location (e.g. geographical places like continents, regions, countries, cities, etc.) and a geographical operator (e.g. spatial relations like “in”, “near”, “north of”, etc.). From GeoCLEF 2006 the topics did not contain explicit expressions with geographic references and geographic operators (see an example of a topic of GeoCLEF 2006 in Figure 3.2). This implies that geographical references (geographic places, and geographic relations) were embedded in the title, description, and narrative. In addition, new geographic relationship were added, such as geographic distance (e.g. “within 100km of Frankfurt”) and complex geographic expressions (e.g. “Northern Germany”).

```
<top>
<num>GC033</num>
<EN-title> International sports competitions in the Ruhr area</EN-title>
<EN-desc> World Championships and international tournaments in
the Ruhr area</EN-desc>
<EN-narr> Relevant documents state the type or name of the competition,
the city and possibly results. Irrelevant are documents where only part of the competition takes place
in the Ruhr area of Germany, e.g. Tour de France,
Champions League or UEFA-Cup games.</EN-narr>
</top>
```

Figure 3.2: Example of a topic of the GeoCLEF 2006 edition.

In GeoCLEF 2007 (Mandl et al., 2008) the following three difficulties were introduced: 1) specifying complex (multiply defined) geographic relations: “East Coast of Scotland”; “Europe excluding the Alps”, “main roads north of Perth”, “Mediterranean coast”, “Portuguese islands”, and “the region between the UK and the Continent”, 2) politically defined regions smaller than countries such as: “French speaking part of Switzerland”, “the Bosphorus”, “Northern Italy”, “Grande Lisboa”, or larger than countries: “East European countries”, “Africa and north western Europe”, 3) finer geographic subjects, such as: “lakes”, “airports”, “F1 circuits”, and even one cathedral as place. Mandl et al. (2009) reports that in order to increase the difficulty of the topic set, the following issues were explicitly included in the topics of GeoCLEF 2008: 1) imprecise /vague geographic regions (“Sub-Saharan Africa”, “Western Europe”), 2) geographical relations beyond IN (“forest fires on Spanish islands”), 3) granularity below the country level (“fairs in Lower Saxony”), 4) terms which are not explicitly mentioned in documents (“Portuguese communities in other countries”).

The relevance judgements were binary, i.e. the document either meets the information need expressed in a topic (1) or not (0) (Leidner, 2005). The test collections for English are composed of 100 topics (25 topics per year from 2005 to 2008). The GeoCLEF English document collection consists of 169,477 documents composed by stories from the British newspaper *The Glasgow Herald* of 1995 (GH95) and the American newspaper *Los Angeles Times* of 1994 (LAT94).

In Martins and Calado (2010) the different kind of geographical topics at GeoCLEF GIR evaluations are reported:

- Feature types with non-geographic restrictions (e.g. “rivers with vineyards”).
- Feature type with geographical place restriction (e.g. “cities in Germany”).

- Thematic subject associated to a toponym (e.g. “independence of Quebec”).
- Topics with a non-geographic subject that is a complex function of place (e.g. “European football cup matches”).
- Vague topics (e.g. “Sub-Saharan Africa”).
- Geographical relations among toponyms (e.g. “Oil and gas extraction found between the UK and the Continent”).
- Geographical relations among events (e.g. “F1 circuits where Ayrton Senna competed in 1994”).
- Relations between events in specific toponyms (e.g. “Casualties in fights in Nagorno-Karabakh”).

3.1.6.2 Results of the Monolingual English Systems at GeoCLEF

The official results of the GeoCLEF evaluations for Monolingual English systems (from 2005 to 2008) are presented in Tables 3.1, 3.2, 3.3, and 3.4 respectively. In GeoCLEF 2005 and 2006 the best results were obtained by probabilistic IR systems (including Logistic Regression) that use BM25 and do not use Geographical Knowledge: F. Gey et al. (2005) and Guillén (2005) at GeoCLEF 2005 and F. Gey et al. (2006), Guillén (2006), and Toral et al. (2006) at GeoCLEF 2006. It must be take into account that some of the top-performing systems (F. Gey et al., 2005; Martins et al., 2006) achieved good results using manual Query Expansion.

Group	IR	GeoKB	QE	RF	MAP
(R. Larson et al., 2006)	BM25	no	auto	BF	0.3936
(R. Larson et al., 2006)	BM25	yes	auto	-	0.3879
(Leidner, 2005)	TFIDF	yes	yes	-	0.1850
(Cardoso et al., 2005)	TFIDF	yes	manual	-	0.2253
(Kornai, 2005)	TFIDF	yes	auto	-	0.1700
(Buscaldi et al., 2005)	TFIDF	yes	auto	-	0.1464
(Ferrés et al., 2005a)	TFIDF	yes	auto	-	0.2231
(Ferrández et al., 2005)	BM25	no	auto	-	0.3495
(Kornai, 2005)	TFIDF	yes	auto	-	0.1700
(Lana-Serrano, et al., 2006a)	trie	yes	auto	-	0.2653
(Guillén, 2005)	DFR	no	auto	PRF	0.3616
(Guillén, 2005)	DFR	yes	auto	PRF	0.3032

Table 3.1: Best MAP configurations of the approaches in the official GeoCLEF 2005 evaluation.

Group	IR	GeoKB	QE	RF	MAP
(R. Larson and F. Gey, 2006)	BM25	no	auto	BF	0.2656
(R. Larson and F. Gey, 2006)	BM25	no	manual	BF	0.2887
(Martins et al., 2006)	BM25	yes	manual	-	0.2080
	BM25	yes	manual	BF	0.2150
(Buscaldi et al., 2006)	TFIDF	yes	auto	-	0.2660
(Hu and Ge, 2006)	TFIDF	yes	auto	-	0.2758
(Lana-Serrano, et al., 2006b)	BM25	yes	auto	-	0.2000
(Andogah, 2006)	TFIDF	yes	auto	-	0.2195
(Yi Li et al., 2006)	BM25	yes	auto	-	0.2464
(García-Vega, et al., 2006a)	mix	yes	auto	-	0.2403
(García-Vega, et al., 2006b)	BM25	yes	auto	PRF	0.2403
(Ruiz et al., 2006)	TFIDF	yes	manual	RF	0.2446
(Bischoff et al., 2006)	boolean	yes	auto	BF	0.1875
(S. Overell et al., 2006)	binTF	yes	auto	-	0.1953
(Z. Li et al., 2006)	BM25	yes	manual	-	0.2395
(Lana-Serrano, et al., 2006a)	trie	yes	auto	-	0.2653
(Guillén, 2006)	DFR	no	auto	-	0.2857
(Toral et al., 2006)	DFR	no	auto	-	0.2985
(Hauff et al., 2006)	-	yes	auto	-	0.1875
(Ferrés and Rodríguez, 2006b)	-	yes	auto	-	0.1370

Table 3.2: Best MAP configurations of the GIR approaches in the context of the official GeoCLEF 2006 evaluation.

By contrast in GeoCLEF 2007 and GeoCLEF 2008, the top-ranked systems (Ferrés and Rodríguez, 2008a; R. Wang and Neumann, 2009) used Geographical Knowledge.

Group	IR	GeoKB	QE	RF	MAP
(R. R. Larson, 2007)	BM25	no	auto	BR	0.2642
(Andogah and Bouma, 2007)	TFIDF	yes		-	0.2515
(Cardoso et al., 2007)	BM25	yes	yes	yes	0.2180
(Z. Li et al., 2007a)	BM25	no	no	no	0.1519
(Buscaldi and Rosso, 2007)	TFIDF	no	-	-	0.2636
(Ferrés and Rodríguez, 2007b)	TFIDF	yes	auto	RF	0.2850
(Kölle et al., 2007)	TFIDF	no	-	-	0.1535
(S. E. Overell et al., 2007)	TFIDF	no	-	-	0.1850
Moscow state Univ.	-	-	-	-	0.1761
(Hughes, 2005)	zettair	yes	-	-	0.2514
(J. M. Perea-Ortega et al., 2007)	BM25	yes			0.2605
(Guillén, 2007)	InL2	no	auto	RF	0.21

Table 3.3: Best MAP configurations of the approaches at GeoCLEF 2007 evaluation.

Group	IR	GeoKB	QE	RF	MAP
(R. R. Larson, 2008)	BM25	no	auto	BR	0.2685
(R. Wang and Neumann, 2008)	TFIDF	yes	manual	-	0.3037
(Buscaldi and Rosso, 2008b)	TFIDF	yes	auto	-	0.254
(Villatoro-Tello et al., 2008)	TFIDF	yes	auto	-	0.318
(S. Overell et al., 2008b)	TFIDF	yes	auto	-	0.264
(Pu et al., 2008)	TFIDF	yes	auto	-	0.2624
(Perea-Ortega, et al., 2008b)	Fusion	-	-	-	0.286
(Perea-Ortega, et al., 2008a)	okapi	yes		PRF	0.2841
(Guillén, 2008b)	InL2	no	auto	-	0.16

Table 3.4: Best MAP configurations of the approaches at GeoCLEF 2008 evaluation.

Several researchers performed experiments with the GeoCLEF collections and some of them improved the official MAP results (see in Table 3.5 these systems).

Table 3.5: Best MAP results at official GeoCLEF evaluation and other posterior results that outperformed them (all combinations of Topic, Description and Narrative allowed)

Approach	GeoCLEF2005	GeoCLEF2006	GeoCLEF2007	GeoCLEF2008
Best official results	0.3936[10]	0.3034[11]	0.2850[12]	0.3037[13]
(Andogah, 2010)	-	-	0.2935	-
(Villatoro-Tello et al., 2010)	**0.5910**	**0.6942**	**0.4960**	**0.4959**
(J. M. Perea-Ortega, 2010)	0.4034	-	-	0.3270
(Ferrés and Rodríguez, 2015a)	0.3974	0.3390	0.2937	-
(Zaila and Montesi, 2015)	-	-	-	0.4890

[10] (R. Larson et al., 2006)

[11] (Martins et al., 2007b)

[12] (Ferrés and Rodríguez, 2008a)

[13] (R. Wang and Neumann, 2009)

3.2 Geographical Question Answering - State-of-the-art

Question Answering (QA) is the task of, given a question expressed in natural language, retrieving its correct answer(s) (single items, text snippets,...) from closed collections or the Web. This task could be considered a step beyond Information Retrieval (IR). IR systems retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza-Yates and Ribeiro-Neto, 1999). In IR, the user query normally consists on a set of relevant keywords and/or regular expressions. In QA the user request is a natural language question that demands for a single piece of information instead of an entire document, and the input is a question expressed in Natural Language instead of a query. Document collections, digital libraries, web search engines and other sources of electronic documents are often used in both IR and QA.

Geographical Question Answering (GeoQA) can be defined as Question Answering type in which questions include geographical expressions (e.g "What is the capital of United States?"). This Section describes specific GeoQA systems and the generic QA approaches and modules that can be used and adapted for the geographical domain.

Some well-known past research projects and systems related with Geographical QA are: *CITYTOUR*, *GeoQuery*, *START*. The *CITYTOUR* project (Andre et al., 1986) was designed to answer natural language questions about the spatial relationship between objects in a city. *GeoQuery*[14] is a learned Natural Language Interface to a US Geography Database (Zelle and Mooney, 1996). Geoquery contains a small database of information about United States geography and can be trained with machine learning for semantic parsing to map novel natural language queries. *START*[15], was the first Web-based question answering system (Katz et al., 2002) that was dealing efficiently with geography. START was able to resolve several types of geographical questions by consulting knowledge databases.

Waldinger et al. (2003) presented GeoLogica, a system that deduces answers to geographical questions based on querying agents (knowledge bases). GeoLogica translate questions into logic by a natural language parser called Gemini. GeoLogica incorporates the Alexandria Digital Library feature type classification. The agents used were: the Alexandria Digital Library Gazetteer, the CIA World Factbook, the ASCS search engine, map providers (TerraVision, NIMA's Geospatial Engine, Generic Mapping Tools, the NASA Goddard Distributed Active Archive, and the NASA Landsat Project), and procedures for performing numerical and geographical computations.

Minock (2005) presented a demo of a the STEP system for natural language access to relational databases. The demo allowed to perform queries to a geographical database. Behrangi et al. (2007) proposed a density-based algorithm that uses fuzzy logic without NLP tools. They selected 400 "where-is" questions from TREC and used google snippets to find the answers achieving a 62.40% of accuracy.

Mishra et al. (2010) used a Semantic approach for textual-based Geographical-Domain Geographical Question Answering with mapping abilities. Some parts of this approach are similar to the approach of Ferrés and Rodríguez (2006a) (explained in Chapter 5, Section 1). Text collections from various cities of India were collected from Wikipedia and other sites. They applied Stanford NLP PoS-Tagging, NERC, and Parsing tools and WordNet 3.0 to perform the analysis of queries and texts. Both Question classification and Expected

[14] **GeoQuery**. http://www.cs.utexas.edu/users/ml/geo-demo.html

[15] START. http://start.csail.mit.edu/

Answer Type are performed. Passage Retrieval is based on pre-processed texts. The Answer Extraction phase employs measures of confidence estimation over semantic similarities (through Wornet) among the question and the documents. The evaluation of the systems was performed with a closed collection of documents about 5 cities of region of India and a set of 152 questions. The results were about 80% of accuracy.

Younis et al. (2012) performed hybrid structured question answering over a structured database, the DBpedia.

Chen (2014) uses a knowledge-based GeoQA system that relies on GIS instead of texts to treat some types of geographical questions using the following techniques: 1) ontologies of spatial relationships, 2) GIS (Geographical Information Systems) data 3) NLP to treat the queries, 4) ontological reasoning, and 5) spatial SQL queries. The type of geographical questions treated in Chen (2014) are: 1) location (e.g. "Where is X?"), 2) distace ("How far is X from Y?",3) "Which city is the nearest to X?"), and 4) proximity buffer (e.g "What cities are within 5 miles of Columbus?").

3.2.1 Classification of QA Systems

Question Answering systems can be classified from different points of view. This section presents different QA categorization types. This categorization types can be applied also to Geographical Question Answering systems.

- *Classification by Knowledge Used.* Moldovan et al. (1999) provided a taxonomy of QA based on the necessary knowledge to resolve the questions. They considered important the three following criteria: Knowledge Bases (KB), Reasoning, and Natural Language Processing (NLP) indexing techniques. Knowledge bases and reasoning provide the medium for building question contexts and matching them against text documents. Indexing identifies the text passages where answers may lie, and natural language processing provides a framework for answer extraction (Moldovan et al., 1999).

- *Classification by Question Types.* There are several types of questions: definitional, list, context, interactive and factoid questions. Factoid questions are the most common researched ones and have been largely evaluated in several international QA contests such as TREC, CLEF, and NTCIR. Factoid questions are questions that seek short fact-based answers like entities, organizations, persons, dates,...(e.g. *What is the capital of France?*, *Who is the President of the United States?*, *Which is the color of the sky?*). Usually the answer is a noun (e.g. blue), a noun phrase (slightly blue) or a Named Entity (e.g. 1979, Paris, George Bush). But some times an adjective or an adverb could be the answer (i.e. in most of the "How" questions).

- *Classification by Domain.* Open-Domain Question Answering (ODQA) systems deal with general questions about many themes. Normally these systems use huge corpus and/or the World Wide Web to extract the answer. On the other hand, Restricted-Domain QA (RDQA) systems deal with questions about a specific domain (e.g. geography, medicine, etc.) (Molla and J. Vicedo, 2005). They often use domain-specific knowledge bases and corpus. Usually, for RDQA, the answers are searched in relatively small domain specific collections, so methods based on exploiting the redundancy of answers in several documents are not useful. Furthermore, a highly accurate Passage Retrieval module is required because frequently the answer occurs in a very small set of passages. RDQAs are frequently task-based. So, the repertory of question patterns is limited allowing a good accuracy in Question Processing with limited effort. User requirements regarding the quality of the answer tend to be higher in RDQA. As Chung et al. (2004) pointed out, "no answer" is preferred to a wrong answer. In RDQA not only NEs but also domain specific terminology plays a central role. This fact usually implies that domain specific lexicons and gazetteers have to be used. In some cases, as in Geographical Domain, many documents included in the collections are far to be standard NL texts but contain tables, lists, ill-formed sentences, etc. sometimes following a more or less defined structure. Thus, extraction systems based on the linguistic structure of the sentences have to be relaxed in some way to deal with this kind of texts.

- **Classification by Information Access** Two basic types of Question Answering systems can be distinguished depending of the structure of the knowledge that they use to answer.

  - **Database-oriented**: systems that access to structured information contained in a database in order to answer the questions. The main challenge of these systems is to transform a natural language question into a database query (Monz, 2003). The fact that this systems are focused in good results, but expand to other domains is a hard task, expertise is required.
  - **Text-based**: Most systems use unstructured information such as plain texts: newspapers, manuals, encyclopedias, etc. to find the answer. Textual question answering systems match the question with text units, e.g., phrases or sentences, in the document collection, and within those units, identify the element the question is asking for. The task of identifying elements of the appropriate type is closely related to the research area of Information Extraction and Named Entity Recognition and Classification. Moreover, for text-based QA system data redundancy plays and important role for answer extraction (i.e. more data implies higher chance that appear occurrences in text where this information is expressed in a way similar to the question). On the other hand, huge amounts of data increases the computational costs of finding an answer.

3.2.2 Architecture of QA Systems

The common architecture of most of the existing QA and GeoQA systems is generally divided into 3 phases: Question Classification, Passage Retrieval (sometimes divided into Document Retrieval and true Passage Retrieval) and Answer Extraction (sometimes divided into Candidates Extraction and Answer Selection) (see in Figure 3.3 a generic QA architecture). In most systems, these phases are executed sequentially, but some systems such as PowerAnswer (S. Harabagiu et al., 2005) perform several iterations in order to get the correct answer. In every system NLP and IR techniques are applied, some times with manually built rules or databases or learned approaches using ML techniques.

3.2.2.1 Question Classification

The Question Classification task consists in: given a question q, assign one or more class labels c_i from a class set C to the question. Question Classification for QA could be seen

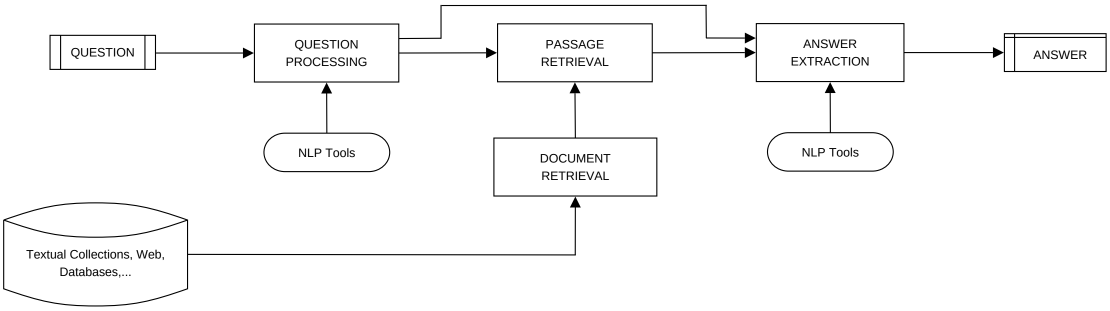


Figure 3.3: Generic architecture of a Question Answering system.

as a multi-class single-label or multi-label classification problem. Depending on the QA typology, question's ambiguity among classes could be allowed and a multi-label tagging could be accepted. For example, the question *Who designed the Eiffel Tower?* in some typologies could be seen as a *Who-person* question and/or a *definee* question.

Question Classification (QC) is a crucial issue in QA because a the question class leads the Answer Extraction system to extract the correct expected answer. Consequently, question categories strongly depend on the Named Entity set of the extraction component employed to tag the documents of the collection. Depending on the system, several entity sets were employed (typically the MUC set).

Early works in theoretical QA proposed question categorization schemes. Lehnert (1978) grouped together questions under 13 conceptual categories. Arthur Graesser's Taxonomy of Inquiries (Graesser et al., 1992) has foundations both in theory and in empirical research. It uses Lehnert's 13 categories to which have been added 4 new categories. Graesser showed that its taxonomy is able to accommodate all inquiries that occur in a discourse.

Open-domain QA systems started with few categories, normally related with the expected noun classes (Named Entities) to be returned as the answer and strongly based on the interrogative pronouns used in the question. The approaches to the QC task are based on manually built rules or Machine Learning techniques that use sets of lexical, semantic or syntactic features to perform the task. Manual rules based on patterns to detect questions of the same answer type (Breck et al., 1999; Prager et al., 2000) used on particular words and on part-of-speech tags. (e.g for example if the pattern *<how (large|small|big)>* is matched, the type MEASURE is returned.) Pasca and S. M. Harabagiu (2001a) presented the first QA system to use patterns with syntactic parsing features and semantic information from WordNet. The WebClopedia (E. H. Hovy et al., 2000) project annotated a QA typology based in the user's intention. They analyzed a set of 17,384 questions and answers to create the typology. The QA Typology contains 94 nodes, of which 47 are leaf nodes and includes classes such as Why-Famous (for *Who was Christopher Columbus?*), Abbreviation-Expansion (for *What does NASA stand for?*).

Although manually hand-crafted rules allow a rapid development of a simple QC. A low coverage and a lack of adaptability are the main problems of this approach. On the other hand Machine Learning approaches to open-domain QC have reached successful results in the last years. This methods require a large amount of data to build good classifiers automatically. Radev et al. (2002) applied decision rule induction using Ripper with 17 question types (person, place, date, number, definition, organization, description, abbreviation, knowfor, rate, length, money, reason, duration, purpose, nominal, and other) and using for learning 1200 questions from TREC-8, TREC-9 and TREC-10. The Results of using Ripper to identify question types with primitive lexical features were 30% of error in the testing using the TREC-10 and the other collections to train. On the other hand, X. Li and Roth (2002) used SNoW (Winnow algorithm) to learn two simple classifiers (a coarse classifier and a fine one). They used two-layer taxonomy which represents a semantic classification of typical TREC questions. The hierarchy contains 6 coarse classes (ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION and NUMERIC VALUE) and 50 finer classes. This was a successful approach, achieving a 98.80% of precision for coarse classes with question features and 95% for the fine classes over the 500 TREC-10 questions. Zhang and Lee (2003) experimented with five machine learning algorithms: Nearest Neighbors (NN), Nave Bayes (NB), Decision Tree (DT), Sparse Network of Winnows (SNoW), and Support Vector Machines (SVM) using two kinds of features: bag-of-words and bag-of-ngrams. They used the same two-layer taxonomy and training-testing datasets of X. Li and Roth (2002). Their experiments results showed that with only surface text features SVM outperforms the other four methods for this task. They also discussed about the importance of the syntactic structures of questions because SVMs with a kernel tree can improve the results of a single-layer SNoW using the same syntactic features. Suzuki et al. (2003) used a hierarchical SVM to experiment with feature sets that include words, named entities and semantic information. They measured a question type hierarchy at different depths and achieved accuracy rate ranging from 95% at depth 1 to 75% at depth 4. Solorio et al. (2004) proposed an algorithm for a Language-independent QC based on Support Vector Machines (SVMs) using 7 classes (person, place, date, measure, organization, object, other). They employed lexical features and Internet features avoiding semantic and syntactic information. They used the DISEQuA corpus (Magnini et al., 2003) , which consists in 450 questions formulated in four languages: Dutch, English, Italian and Spanish, with 10-fold cross-validation for English, Spanish and Italian, obtaining a results of 81.77% for English, 88.70% for Italian and 81.45% for Spanish. X. Li and Roth (2004) repeated their experiments in QC using their framework test set with the use of semantic information sources for this task. Their experiments results show that semantic information can improve the performance of the QA task. Classification accuracies over 1,000 TREC-2002 questions reached 92.5% for 6 coarse classes and 89.3% percent for 50 fine grained classes. On the other hand Shen et al. (2006) obtained a classification accuracy of 80.8 % in fine grained classes with the previous experiments. They used a Language Modelling approach with a Kneser-Ney smoothing for bigram features.

Most systems perform in parallel QC and extraction of information from the question (as question keywords, expected answer type, etc.). Keywords Selection is one of the most important steps in the Question Processing phase. Lexical terms (keywords) from the question normally used as a query to an IR/PR system lead to the relevant documents. These keywords could possibly expanded with lexical/semantic variations.

3.2.2.2 Passage Retrieval

Given a textual corpus and a user query (a question, a set of keywords, ...), Passage Retrieval (PR) could be defined as the task of retrieving a set of passages from the textual corpus relevant to the user query. Obviously, a passage is considered a portion of a whole document. A passage could have fixed size (words, bytes or sentences) or a dynamic size (paragraph, sentence,...). In QA the aim of Passage Retrieval is to get small fragments of text (with enough context) which probably contain the answer of the question.

A typical PR system has normally two phases, an indexing phase and a searching phase. The first one, called Indexing, consists in processing all the collection and extract its essential information. Then, in a following step of the same process, the information is stored in a structure that allows an easy recovery of the primordial data by querying for some features.

The core of each PR system has an Information Retrieval algorithm. IR techniques can be sub-classified in tree classes depending on its mathematical model:

- **Set Models**. These models represent documents by sets. The *Standard Boolean model* is the most popular.

- **Algebraic Models**. These algorithms represent documents and queries usually as vectors, matrices or tuples. The *Vector Space model* is the algebraic model most widely used in the IR community. In the vector space model, all the documents are mapped into a N-dimensional space in which each term represents a dimension. Each document and query is represented as a vector in this vectorial space. Document relevance with respect to a query is computed using distance measures between the document vector and the query vector. Term weighting is usually performed by TFIDF (Salton and Buckley, 1988) or Okapi's BM25 (Robertson and Walker, 1994) schemas.

- **Probabilistic Models**. These models represent similarities as probabilities. In the probabilistic models the estimated relevance of a document to a query is a function of the estimated probabilities that each of the various terms in the document occur in at least one relevant document but in no irrelevant documents. Currently, *Language models* (LM) and *Divergence From Randomness* (DFR) models (Amati, 2003) are ones of the most established probabilistic models.

Information Retrieval engines are the core of most text-based QA and GeoQA systems. This paragraph lists and describes some of the most relevant existing IR engines.

- **Lucene**. Lucene[16] IR system uses the standard tf.idf weighting scheme with the cosine similarity measure, and it allows ranked and boolean queries.

- **Terrier**[17]. Performing very well at TREC Terrier includes: parameter-free probabilistic retrieval approaches such as Divergence from Randomness (DFR) models (Ounis et al., 2006), the TF-IDF (with Robertson's TF) weighting scheme, other recent language modelling approaches, and the well-established Okapi's BM25 probabilistic ranking formula.

- **Indri (Lemur project)**. Indri[18] (an IR component of the Lemur toolkit) is an Information Retrieval system that supports retrieval algorithms based on Language Modelling (Ogilvie and Callan, 2001).

[16] **Lucene**. `http://lucene.apache.org/java/docs/`

[17] **Terrier**. `http://ir.dcs.gla.ac.uk/terrier/`

[18] **Indri**. `http://www.lemurproject.org/`

- **JIRS**. The JAVA Information Retrieval System (JIRS) software (Soriano et al., 2005) is used to retrieve relevant passages related to a question. JIRS[19] was specially designed for Question Answering (QA). This system gets passages with a high similarity between the largests n-grams of the question and the ones in the passage. It has 3 modes: simple n-gram model, term weight n-gram model, and distance n-gram model.

- **Sphinx**. Sphinx[20] is a full-text search engine that provides fast, size-efficient and relevant full-text search functions to other applications. Sphinx has two types of weighting functions: Phrase rank and Statistical rank. Phrase rank is based on a length of longest common subsequence (LCS) of search words between document body and query phrase. Statistical rank is based on classic BM25 function which only takes word frequencies into account.

Indexing

Rijsbergen (1979) defined an index language as the language used to describe documents and requests. The elements of the index language are index terms, which may be derived from the text of the document to be described, or attached to it. Usually, documents are indexed using its words as an *indexed terms*. In the indexing phase some dimensional reduction techniques (Term Normalization) are applied. The most popular indexing technique is the use of Inverted Indexes, that consists in having a inverted list for each index term. Some pre-process over the terms before indexing include:

- **stopwords removal**: avoids the indexing of irrelevant information by filtering out words with high frequency of occurrences is text that they lose their utility as search keywords and/or words without semantic importance such as articles, prepositions, pronouns, etc.

- **stemming**: a stemmer is an algorithm that given a word form determines its stem form. The stem is not necessarily identical to the root of the word. As an example, for English, an stemmer will possibly identify the string “build” as the stem of the following word forms: “building”, “builders”. The Porter algorithm is very widely used as a standard stemmer for English (Porter, 1997). This method removes the commoner morphological and inflexional endings from words in English.

- **lemmatization**: a lemmatizer is an algorithm that given a word form determines its lemma by using the part of speech of the word in a sentence. It requires a lexicon that store the necessary knowledge of the language (i.e. a lemma and its associated lexeme, the pair <word form, part-of-speech>). lemmatization differs from Stemming in the fact that requires the knowledge of the POS tag of the word in the sentence and needs a knowledge base of lexemes. Stemming does not take into account the function of the word in the sentence, does not require a great knowledge of the language, and normally works by stripping morphological and inflexional endings of the words. As an example, the word “went” has “go” as a lemma, but its stem is the word form itself.

[19] **JIRS.** `http://sourceforge.net/projects/jirs/`

[20] **Sphinx**. `http://www.sphinxsearch.com`

- **Named Entity indexing**: indexing Named Entities as a multi-word class can improve the recall and avoid noise in the retrieval. However, a high precision NERC is required in order to lose recall. (Prager et al., 2000) started this approach by indexing Named Entities and their class (predictive annotation). This method identifies potential answers in the text and then indexes their corresponding Named Entity class or Expected Answer Type.

- **semantic indexing**: using WordNet synsets to index collections can improve the recall of IR systems respect to word based indexing. Gonzalo et al. (1998) used the SMART IR and SemCor (a disambiguated collection) to index by synsets with dubious results. In fact the increase in recall (29%) has a decrease in precision counterpart due to polysemy. What is true is that with accurate WSD module (currently not existing) the results could be good. Mihalcea and Moldovan (2000) experiments indexing by synsets reported also an improvement in IR effectiveness using the Cranfield collection. Liu et al. (2004) used effectively WordNet to disambiguate word senses of query terms.

Searching

Searching documents in IR systems implies the use of a textual query in a boolean or ranked manner to obtain a set of ordered or unordered relevant documents. Boolean searches involve the use of logical operators such as: AND, OR, and NOT over the query terms to find a set of documents that satisfy the logical expression. Ranked retrieval, on the other side, does a ranking over a set of documents based on keywords similarities.

IR systems sometimes offer capabilities like phrasal search (searching for a phrase or a specific sequence of words (e.g. "Tom Cruise")), fuzzy matches (e.g. "*at" will match "Pat" or "rat"), regular expression (regexp) matches or boosting terms (i.e. weighting search terms). A frequent approach in Searching is Query Expansion (QE). The QE approach is often used to increase the recall of the system by adding similar terms to the ones in the original query. WordNet has been used for this purpose by expanding terms with its synonyms, hyponyms, and hypernyms[21]. On the other hand, Gazetteers, encyclopedic knowledge, and abbreviations, can be used in certain domains to realize QEs.

The number of documents to retrieve depends on the task. In QA, normally it depends on the document processing capability of the system. The processing capability depends on the computational resources available to process and the computational costs of the algorithms designed to process the documents. Sometimes deep NLP approaches might require expensive computational resources and processing time and use only few documents (and/or passages), and some simple approaches with lesser requirements can cope with more data.[22]

In the Information Retrieval field, for research purpose the first top 1,000 documents are taken into account to evaluate the systems (e.g. TREC, and CLEF adhoc IR tasks). In the real world, normally the user wants the search engines for no more than 50 documents. For QA, usually few documents/passages are used to extract the answer. In PR the searching

[21] Without a good WSD this kind of expansion has to be done very carefully for avoiding the introduction of noisy terms.

[22] In online-QA the response time is a critical constraint while in TREC or CLEF contests time process can be huge.

process retrieves passages sometimes with overlapping and sometimes with fixed size. Jorg Tiedemann (2004) does comparison of different IR systems for QA, in which Zettair and Lucene obtained the best results.

An often used approach to improve searching is Relevance Feedback for IR/PR. Relevance Feedback (RF) consists in using the most relevant terms collected from the top ranked documents of an initial query to compose manually or automatically a second query with more information.

3.2.2.3 Answer Extraction

The Answer Extraction phase has the aim of recover the answer(s) of a certain question. This phase normally takes place after Question Processing and Passage Retrieval and processing.

After passage processing the AE algorithms can use simple and fast answer pattern matching or sophisticated reasoning modules. The Answer Extraction phase is often composed by three subphases: Sentence Retrieval, Answer Ranking, and Answer Selection.

Current approaches to Answer Extraction can be divided into the following points depending on the use of different NLP processing techniques and the type of data used to search the answer:

1. **Linguistic Pattern Matching.**

   Answer pattern matching is one of the most common approaches to the QA task. Answer patterns consists of series of regular expressions based on lexical, syntactic and/or semantic features that allows easily to match the answer sentence context to extract properly the answer. As an example, the following lexical pattern *<X;is/are;[a/an/the];A>* matches “Michigan’s state flower is the apple blossom”. On the other hand, the semantic pattern *<PERSON> was born in <BIRTHDAY>* matches “Mozart was born in 1756”.

   Several groups used manually built rules with great success. Soubbotin (2001); for instance, obtained the best results at the TREC 2001 QA evaluation task (MRR: 0.676) with a system that uses massively indicative lexical answer patterns for a broad range of question types.

   Ravichandran and E. H. Hovy (2002) presented an approach for automatically learning answer patterns (regular expressions) from the web, for certain types of questions. Their method uses bootstrapping learning to build a large tagged corpus staring with only a few examples of QA pairs.

2. **Semantic Matching.**

   Semantic matching is performed using ontologies (e.g WordNet, SUMO, or CYC) sometimes helped by syntactic parsing structures. J. L. Vicedo (2002) used the Semantic Content of the Concept, a semantic representation of questions and sentences based on weights obtained by using *idf* weights and WordNet relationships: synonymy, hypernymy and hyponiny. Ferrés et al. (2004a) represents semantically sentences and questions with binary and unary predicates and applies an iterative relaxation approach by means of structural and hierarchical relaxation of predicates. Lo and Lam (2006) presented a system with a sophisticated grammatical framework that parses the question and candidate answers and the semantic relations are obtained. Then, these relations are compared base on the level of consistency as well as the linkages from the Wikipedia.

3. **Context-based Linguistic Features.** This method uses linguistic features from the candidate's context to perform a ranking of the candidates. FALCON (S. M. Harabagiu et al., 2000), for instance, was an early advanced QA system that applied these approach integrating semantic information using WordNet, Expected Answer Type, Query Expansion, syntactic parsing with Collins' parser and abductive reasoning.

4. **Lexical Matching with Expected Answer Type**

   Expected Answer Type (EAT) matching (Pasca and S. M. Harabagiu, 2001a) is a common strategy for the Answer selection process in most of the current QA systems. Detecting the EAT of a question could be useful in the Passage Retrieval and the Answer Extraction phases. A mapping of answer types to Named Entity types is required. During the PR phases it can be used filtering out the passages without concepts of the same category as the expected answer type. Finally, in the Answer Extraction phase the EAT can be used to select the candidates with the same type. Pasca and S. M. Harabagiu (2001b) used an answer taxonomy that includes 8707 concepts from 129 WordNet subhierarchies. Predictive Annotation and Virtual Annotation are also successful techniques for Answer Extraction introduced by Prager et al. (2000).

5. **Data-Driven Statistical Modelling.**

   Statistical modelling for answer extraction relies in Statistical Machine Learning using annotated corpus of question-answer pairs to learn probability models. Whittaker et al. (2006) presented a non-linguistic multilingual data-driven statistical QA system trained with the TREC QA evaluation datasets and the Knowledge Master KM data[23]. Ittycheriah et al. (2001) created statistical algorithms for both expected answer type prediction and named entity tagging. The answer selection model used maximum entropy with the following feature sets: sentence features, entity features, definition features, and linguistic features.

6. **Cache-Based Services.**

   Although is a simple strategy, some QA systems such as QUARTZ (Jijkoun et al., 2004), Aranea (J. Lin and Katz, 2003) among others have a Database of question-answer pairs that it is consulted before using the QA algorithms given a question.

7. **Inference & Reasoning.** This methods require the use of ontologies and Bases of Knowledge for inferences. LCC's language logic prover, COGEX (Moldovan et al., 2003), is an example of abductive reasoning for QA.

8. **Web-based External Knowledge Mining**. Using the Web as a data source to extract the answer and then apply this information into the extraction process has emerged as new research line in QA. Major search engines and confident data sources

[23] *Knowledge Master data.* Academic Hallmarks, `http://www.greatauk.com`. A non-free library of 142,000 questions about different subjects

as Wikipedia are often used. Systems such as: Aranea (J. Lin and Katz, 2003), and PowerAnswer 3 (Moldovan et al., 2006), among others used this technique.

9. **Complex Hypothesis Generation and Evidence Scoring.** This is a unique approach created by IBM that uses massive resources to perform the generation of several hypothesis from the questions, get candidates from different evidence sources, and get the answer with high confidence. DeepQA (Watson) (Ferrucci, 2012) performs state-of-the-art ML and NLP techniques such as parsing, entity disambiguation, relations detection and textual entailment.

10. **Linked Data Answer Extraction.** These systems consult Knowledge Bases with natural language queries. Pradel et al. (2011) presented the SWIP system, a system that allows consult Knowledge Bases with natural languages queries. QAKiS (Cabrio et al., 2012) is a system for ODQA over linked data that uses relational textual patterns.

3.2.3 Evaluation Benchmarks of QA systems

QA has become a popular task in the NL Processing (NLP) research community in the framework of different international ODQA evaluation contests such as: Text Retrieval Conference (TREC) for English (Voorhees, 2003), Cross-Lingual Evaluation Forum (CLEF) for European languages (Magnini et al., 2003), and NII-NACSIS Test Collection for IR Systems (NTCIR) for Asian languages (Sasaki et al., 2005). More recent QA evaluations focused on several domains and approaches such as Biomedical (BioASQ track from CLEF 2013 to CLEF 2015, and ACL 2016-2017), Geographical (GikiP and GikiCLEF at CLEF 2008 and 2009), Linked Data (QALD track from CLEF 2011 to CLEF 2015, ESWC 2016., and CLEF 2017),.... QA evaluation contests usually provide test collections (data sets usable for experiments) and unified evaluation procedures for experiment results (Voorhees and Tice, 1999). Each participating group conducts research and experiments using the common data provided by the organization with various approaches. The TREC[24] conference is the most popular international evaluation framework in the field of Information Retrieval for English. It has different tracks (areas of IR) that propose different tasks related to IR. NIST provides participating groups with test sets and evaluates the results of the participants. From 1999 to 2007, a special open domain QA track was carried out every year. The TREC conference has fostered and has inspired a substantial set of publications and current QA systems. The CLEF[25] is an international evaluation framework for IR in European Languages. CLEF provides the infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts. Within the framework of the CLEF, a Multilingual QA track for was succesfully carried out from 2003 to 2008. Both QA tracks (at CLEF and TREC) included some geographical questions in their questions. GikiCLEF (at 2009) was an evaluation task under the scope of CLEF with the aim to evaluate systems which find Wikipedia entries/documents that answer a particular information need. For GikiCLEF, systems needed to answer or address geographically challenging topics on the Wikipedia collections returning Wikipedia document titles as list of answers in all languages it can find answers.

[24] **TREC**. `http://trec.nist.gov`

[25] **CLEF**. `http://www.clef-initiative.eu`

3.2.4 Evaluation Metrics of Factoid QA Systems

The main capabilities in QA were discussed by the QA road-map committee (Burger et al., 2000). The following capabilities are the most relevant.

- **Timeliness.** The answer to a question must be provided in real-time, and the question could refer to most recent events and facts.
- **Accuracy.** The precision of QA systems is extremely important as incorrect answers are worse than no answers. To be accurate, a QA system must incorporate world knowledge and mechanisms that mimic common sense inference.
- **Usability.** This capability implies the rapid prototyping of domain-specific knowledge and its incorporation in the open-domain ontologies, the use of heterogeneous data sources, deal with heterogeneous data formats and allow the user to describe the context of the question.
- **Completeness.** Complete answers to a user's question is desirable. Some times answer fusion is required.
- **Relevance.** The answer to a user's question must be relevant within a specific context. The evaluation of QA system must be user-centered: humans are the ultimate judges of the usefulness and relevance of QA systems and of the ease with which they can be used.

The main issues on evaluation of the different components of QA systems are reported here:

Question Processing. The Question Processing phase consists in the analysis of the question using NLP tools (morphosyntactic analyzers, syntactic parsers, Named Entity Recognizers, semantic parsing,...). Although is not common, the evaluation of this part is an important step to avoid cumulative errors in the following phases. So, for example, Named Entity Recognition and Classification could be influenced by POS-tagging errors and semantic pre-processing could depend on the errors in the NERC and the syntactic parsing steps.

Question Classification. In the Question Classification phase normally, is evaluated its global accuracy (number of correct questions classified divided by the total number of questions) and the accuracy of the classifiers for a specific class c.

$$Accuracy = \frac{\#correct\_predictions}{\#predictions} \tag{3.5}$$

$$Accuracy(c) = \frac{\#correct\_predictions\_of\_class\_c}{\#predictions\_of\_class\_c} \tag{3.6}$$

Passage Retrieval. Let Q be the question set, D the document (or passage) collection, $A_{D,q}$ the subset of D which contains correct answers for $q \in Q$, and $R^S_{D,q,n}$ be the n top-ranked documents (or passages) in D retrieved by a retrieval system S given question q. The following metrics of a retrieval system S for a question set Q and document collection D at rank n are defined:

- **Coverage** (or Accuracy): the Coverage, sometimes called Accuracy, gives the proportion of the question set for which a correct answer can be found within the top n documents retrieved for each question.

$$coverage^S(Q, D, n) = \frac{|\{q \in Q|R^S_{D,q,n} \cap A_{D,q} \neq \emptyset\}}{|Q|}$$

- **Redundancy**: the answer redundancy gives the average number, per question, of passages within the top n ranks retrieved which contain a correct answer.

$$redundancy^S(Q, D, n) = \frac{\sum_{q \in Q} |R^S_{D,q,n} \cap A_{D,q}|}{|Q|}$$

- **Maximum Redundancy**: the maximum answer redundancy any system could achieve.

$$maximum\_redundancy(Q, D, n) = \frac{\sum_{q \in Q} |A_{D,q}|}{|Q|}$$

- **Precision**: the precision of a system for a given question set and document collection at rank n is the average proportion of the n returned documents or passages that contain a correct answer.

- **Recall**: the Recall is the average proportion of answer bearing documents that are present in the top n returned documents or passages.

The most useful evaluation metrics to evaluate PR for QA are *coverage* and *redundancy*. On the other hand, *precision* and *recall* are not helpful for PR in a QA context (I. Roberts and Gaizauskas, 2004). Precision cannot capture the goodness of the overall queries, which is crucial for QA, the evaluation is done over a set of questions and these measures can be confusing. Recall is not as unhelpful as precision, because it can show how the retrieved document set approaches to the maximum redundancy obtainable. Redundancy, on the other hand, tells one only how many answering bearing passages per question are being returned on average. However, redundancy gives a measure of how many chances per question on average an answer extraction component has to extract an answer. In addition, Ferrés et al. (2005b) designed two different measures to evaluate the Passage Retrieval for Factoid questions: the first one (called *answer*) is the accuracy taking into account the questions that have a correct answer in its set of passages. The second one (called *answer+docID*) is the accuracy taking into account the questions that have a minimum of one passage with a correct answer and a correct document identifier in its set of passages.

Answer Extraction. The evaluation of the Answer Extraction module can be done in different modes depending on the number of sub-tasks that has this module. When the Answer Extraction is a single module the evaluation takes into account the retrieved passages with a correct answer to perform an evaluation of the Answer Extraction accuracy.

$$AnwserExtractionAccuracy = \frac{\#questionsWithCorrectAnswerExtracted\&1passageEntailsAnswer}{\#questions\_withAtLeast1passageEntailsAnswer} \tag{3.7}$$

Sometimes Answer Extraction uses two steps: Candidate Extraction (CE) module, and Answer Selection module. Then every step can be evaluated separately.

$$CandidatesExtractionAccuracy = \frac{\#questionsWithCorrectCandidateExtracted\&1passageEntailsAnswer}{\#questionsWithAtLeast1passageEntailsAnswer} \tag{3.8}$$

$$AnswerSelectionnAccuracy = \frac{\#questionsWithCorrectAnswerExtractedFromASupportedPassage}{\#questionsWithCorrectCandidateFromASupportedPassage} \tag{3.9}$$

Question Answering. QA judgements of factoid questions in current QA evaluations often consider a response as a single pair of answer-string and document identifier. If a pair <answer-string, document-identifier> pair is given as a response, the answer-string must contain nothing other than the answer, and the document identifier must be the global identifier of a document in the collection that supports answer-string as an answer. Sometimes if the system detects that there is no answer in the collection the response pair reflects that the question answer is nil. These answers will be judged correct if there is no answer known to exist in the document collection; otherwise it will be judged as incorrect. An answer string must contain a complete, exact answer and nothing else. As with correctness, exactness will be in the opinion of the assessor. Responses will be judged by human assessors who will assign one of four possible judgments to a response:

- *incorrect*: the answer-string does not contain a correct answer or the answer is not responsive.
- *unsupported*: the answer-string contains a correct answer but the document returned does not support that answer (i.e does not textually entails the answer).
- *non-exact*: the answer-string contains a correct answer and the document supports that answer, but the string contains more than just the answer (or is missing parts of the answer).
- *correct*: the answer-string consists of exactly a correct answer and that answer is supported by the document returned.

Mean Reciprocal Rank (MRR). MRR represents the mean score over all questions and is one of the most used evaluation measure in QA. MRR takes into consideration both recall and precision of the systems performance, and can range between 0 (no correct responses) and 1 (all the queries have a correct answer at position one). Two versions of MRR can be applied in a QA evaluation: a) 'strict', where unsupported responses are counted as wrong, and b) 'lenient' where unsupported responses are counted as correct.

$$MRR = \frac{\sum_{i=1}^{|Q|} \frac{1}{far(i)}}{|Q|} \tag{3.10}$$

Accuracy. The accuracy measure is commonly used in all the QA evaluations (TREC, CLEF, NTCIR). The accuracy measures the precision giving the answer at the top-N rank of answers. Accuracy is the fraction of questions judged to have at least one correct answer in the first n answers to the questions. Let C be the correct answers.

$$accuracy^s(Q, D, n) = \frac{|\{q \in Q | A_{D,q,n}\}|}{|Q|} \tag{3.11}$$

F-measure. The F measure is controls the relative importance of recall and precision (Voorhees, 2003). The general formula of the F measure is:

$$F = \frac{\beta^2 PR}{(\beta^2 + 1)P + R} \tag{3.12}$$

The β parameter can be used to tune the relative importance of the recall and precision.

3.3 Textual Georeferencing - State-of-the-art

Textual Georeferencing consists in assigning a set of geographical coordinates to formal (news, reports,..) or informal (blogs, social networks, chats, tagsets,...) texts and documents. The approaches to deal with the Textual georeferencing task generally use: 1) data-driven models to predict, and/or 2) Geographical Knowledge bases for toponym recognition, toponym disambiguation, and toponym grounding. But this prediction task it is not possible, or very difficult to perform, in some texts because of lack of enough information. Currently some platforms allow users to georeference (geotag) their content automatically (GPS-enabled cameras) or manually, but most of existing textual and media content is not georeferenced, and thus this task could be applied in many data sources.

The automatic understanding or prediction of the georeference of informal texts from social networks (and other document sources) can be applied to different areas such as tourism (Zheng et al., 2012), discovery of Points of Interest (Skovsgaard et al., 2014), and emergency scenarios (De Longueville et al., 2009) among many others. For instance Zheng et al. (2012) extracts topological characteristics of travel routes by using a pool of geotagged photos of the internet, (Skovsgaard et al., 2014) proposes automatic discovery of Points of Interest from geo-tagged microblog posts, and De Longueville et al. (2009) analyzed the temporal, spatial and social dynamics of Twitter activity during a major forest fire event in the South of France in July 2009.

3.3.1 Approaches for Textual Georeferencing

Many approaches to Textual Georeferencing of formal and informal texts such as user generated data have been presented in last years. Most of them addressed to solve the following problems:

- predict users' location from textual content (Mahmud et al., 2014; Chang et al., 2012; Kordopatis-Zilos et al., 2016),
- predict the geographical focus of the text (Van Laere et al., 2013; Serdyukov et al., 2009; Intagorn and Lerman, 2014),
- predict both the users' location and the geographical focus of the text. (Han et al., 2014; Schulz et al., 2013).

Some of these approaches have been applied over different textual inputs such as Twitter messages, Flickr metadata, Wikipedia pages or other collections. Many approaches to the Georeferencing task for predicting the most appropiate coordinates for tagged images have been presented in last years (Hays and Efros, 2008; Crandall et al., 2009; Serdyukov et al., 2009; Van Laere et al., 2010b) using mainly textual features. Also some of the georeferencing research authors use their own and different corpus collections, and another ones participate and use corpora from specific evaluation benchmarks such as the MediaEval Placing Task. Crandall et al. (2009) presented a system that uses textual (tags), visual, and temporal features for placing Flickr images on map. They used automatic classifiers based on Naïve Bayes and Support Vector Machines trained over a corpus of 35 million images. Their results show that visual and texual features together outperform text features alone by a significant margin. Serdyukov et al. (2009) used a language model based on the tags provided by the users to predict the location of Flickr images. The language model follows a Multinomial distribution. In addition several smoothing strategies were taken into account to test: 1) spatial neighbourhood for tags, 2) cell relevance probabilities, 3) toponym-based smoothing, 4) spatial ambiguity-aware smoothing. They used a set of 140,000 geo-tagged Flicker photos in which there is at most one photo per user with the same tagset. They used 120,000 to train models, 10,000 for tuning parameters and 10,000 for testing purposes. All the smoothing strategies outperformed the Language Modelling baseline. Van Laere et al. (2010b) presented a system that uses clustering (k-medoids) and classification (Naïve Bayes) algorithms to predict geographic coordinates of Flickr photos by using users tags. Hays and Efros (2008) used only visual features and dataset of over 6 million GPS-tagged images from the Flickr online photo collection to estimate the geographical coordinates of the image. Their results evaluating over a set of 237 images. They use a measure of visual features similarity between test image and dataset image to find the 1-NN (Nearests Neighbours) and the 120-NN. Mean-shift clustering then is applied to the estimated probability distribution derived from the 120-NN photos with GPS-coordinates.They report results in prediction locations up to 30 times better than chance.

Rattenbury and Naaman (2009) extracts place and event semantics for tags using Flickr data. C. Wang et al. (2007) for instance tries to find relationships between tags and countries. Baba et al. (2010) research introduces tag georeferencing. They proposed a method for extracting places related to Flickr tags using the co-occurrence of a tag and a geolocation.

Some approaches use twitter data to predict georeferences. Schulz et al. (2013), for instance, uses a multi-indicator approach that uses mapping to polygons to determine the location where a tweet was created and the location of the user's residence. Intagorn and Lerman (2014) presented a framework to predict the location of short text using spatial granularity prediction with confidence. Mahmud et al. (2014) presented a hierarchical ensemble algorithm to predict the home location of twitter users at different granularities using location classification approaches, content-based heuristics, and behavior-based time zone classifiers. Han et al. (2014) investigated the impact of several factors in geolocation prediction accuracy in Twitter: explicit words selection, non-geotagged tweets, language influences, metadata and temporal posting.

Regarding the approaches over formal documents, B. Wing and Baldridge (2011) used language modelling with geodesic grid cells of 1°x 1°with Kullback-Leibler Divergence to find the most probable cell for a given document. They evaluated the approach with Wikipedia and Twitter corpora, obtaining a median prediction error of just 11.8 kilometers for Wikipedia documents. Roller et al. (2012) used and adaptative-grid strategy with a k-d tree with Language Models for geolocation. The adaptative grid cells are created with documents labelled with latitude/longitude coordinates. They also evaluated the approach on Wikipedia and Twitter corpora. B. Wing and Baldridge (2014) employed logistic regresion models on a hierarchic k-d tree grid evaluated over both Wikipedia and Twitter corpora separately. They obtained a median error of 15.3 Km with the English Wikipedia.

On the other hand, Van Laere et al. (2014) applied probabilistic language models trained on Flickr and Twitter to geolocate Wikipedia articles. They showed that language models substantially outperform methods based on Gazetteers and that social media data outperform Wikipedia data alone to geolocate. They used Yahoo! Placemaker and Geonames as the gazetteer based methods. Yahoo! Placemaker is capable of georeferencing documents and webpages and Geonames was used in combination of Natural Language Toolkit and Stanford Named Entity Recognizer to extract toponyms. Then Geonames was used to retrieve the coordinates of the extracted toponyms. To assign coordinates to the Wikipedia article they used the medoid of the places that were found, choosing for each place the nearest coordinates in case of ambiguity.

3.3.2 Textual Geographical Focus Detection

Textual Geographical Focus Detection consists of detecting which place name (disambiguated) represents the scope of the text (Andogah et al., 2008). This task is very similar to the Textual Georeferencing task but it assigns geographical scopes to generic texts or web pages at different levels instead of predicting the geographical coordinates related with texts. Andogah et al. (2008) for instance assigns geographical scopes up to six levels: continent, continent-directional, country, country- directional, province and province-directional. The experiments of Amitay et al. (2005) with the Web-a-Where system for geotagging Web content performed 80% of accuracy in place names tagging and a 91% of accuracy in web page focus detection. Amitay et al. (2005) used a selected gazetteer of 75,000 place names (includes cities, states and countries) extracted from different sources. Martins and Silva (2005) computes the geographic scope of a document using references extracted from the text, information from an ontology, and the PageRank algorithm. The GSR sophisticated approach reported by Andogah et al. (2008) exploits the following Geographical Knowledge to perform the task: placename frequency of occurrence, geographical adjectives, place type (e.g., city), place importance (e.g., based-on population size and place type), and vertical (transitive parent/child) and horizontal (adjacency) relationships among places. Both Amitay et al. (2005) and Andogah et al. (2008) use population heuristics and hierarchical relationships between toponyms. TextGrounder system (Speriosu et al., 2010) performs geolocation connecting natural language texts, expressions, and individual words to geographical coordinates by topic-region probabilistic models.

3.3.3 The MediaEval Placing Task Geo-Estimation Challenge

The MediaEval Placing Task is a multi-modal georeferencing challenge to evaluate algorithms that can predict the location of randomly selected photos and videos from Flickr (Choi et al., 2014; M. Larson et al., 2015). It has been organized in seven editions from 2010 to 2016. The challenge has the following evaluation criteria (Choi et al., 2014) :

- Multiple modalities: georeferencing can exploit the information of the different modalities of the multimedia (audio, video, textual metadata, users info).
- Scalability: the dataset contains images and videos of the entire world. Thus, the algorithms must build locations models of the entire world.
- Noise handling: the benchmark dataset is not filtered by content and has been sampled randomly, and limited only in images/video per used to avoid user bias.
- Location bias and sparsity problem: the distribution of the training data is uneven. Some locations are highly represented, and other locations have little or no data.

3.3.4 Georeferencing Systems at MediaEval 2010-2013 Evaluations

The MediaEval Placing tasks from 2010 to 2013 required that participants automatically assign geographical coordinates (latitude and longitude) to Flickr videos/images using one or more of: Flickr metadata, visual content, audio content, and social information (M. Larson et al., 2015). Evaluation of results is done by calculating the Haversine distance from the actual point (assigned by a Flickr user) to the predicted point (assigned by a participant). Runs are evaluated finding how many videos were placed at least within some threshold distances. Some of the top performing and relevant textual based algorithms presented at MediaEval from 2010 to 2013 are reported in this subsection.

Van Laere et al. (2010a) obtained the best results at the MediaEval Placing Task 2010 obtaining a 67,23% of accuracy predicting georeferences up to 100 Km (5,091 videos) with a system that applies Language Modelling and Clustering. They used a corpus of 8,685,711 annotated metadata from Flickr photos to train the models. Kelm et al. (2010) presented an approach at MediaEval 2010 that combines three different methods to estimate geographical regions: a natural language processing approach with geographical knowledge filtering, probabilistic latent semantic analysis (pLSA) document indexing, and classification method based on colour and edge visual features to train a support vector machine (SVM). In order to apply NLP processors they detected and translated the textual annotations to English using the Google translate service. Their NLP approach is based on the use of a NLP processor, Wikipedia[26] and Geonames[27] with the use of population and higher-level categories salience. They achieve their best results in accuracy for prediction at a maximum of 100 Km with the combination of NLP and Geographic Knowledge method and visual features with a result of 60.46% of accuracy. J. Perea-Ortega et al. (2010) presented a system at MediaEval 2010 that uses an approach based on applying a geographical Named Entity Recognizer (Geo-NER) on the textual annotations. Geo-NER is a geographical entity recognizer that makes use of Wikipedia and GeoNames. Ferrés and Rodríguez (2010b) presented a Geographical Knowledge based approach to predict geographical points from textual user annotations (including tags) at MediaEval 2010, achieving an accuracy of 52% georeferencing up to a distance of 100 Km (this approach is explained in Chapter 6).

Ferrés and Rodríguez (2011b) presented an approach at MediaEval 2011 that used geographical knowledge and Data-Driven IR algorithms derived from previous work (Ferrés and Rodríguez, 2011a) (this approach is presented in detail in Chapter 6). They achieved the best results with the combination of both techniques with up to 59.30% of videos correctly georeferenced within a margin of error of 100km. Hauff and Houben (2011) used a geographical spread selection process to filter out terms that are very geographically spread achieving a competitive margin of error at 100km of 82.6% of 5,347 videos correctly georeferenced at MediaEval Placting Task 2011. Van Laere et al. (2011b) extended their approach presented in 2010 using language models with a more adaptative granularity and taking into account the home location of the user. They achieve a 62.47% of accuracy up to 100km in the run in which they only used textual information. In MediaEval 2012 Popescu and Ballas (2012) explored the use of users models with successful results, Van Laere et al. (2012) extended again their approach including the geospread feature selection introduced by Hauff and Houben (2011). In the same evaluation Trevisiol et al. (2012) used a grid based approach using square grids of 0.1°to compute a co-ocurrence matrix of tags associated with each area in which appear and BM25 feature weighting.

Popescu (2013) presented a system that uses machine tags to predict in combination with user modeling, geographicity and location models based on external training data (90 million items) that achieves the bests results at Media Eval Placing Task 2013. In the same

[26] **Wikipedia.** `http://www.wikipedia.org`

[27] **Geonames.** `http://www.geonames.org`

evaluation Cao (2013) presented a system that uses language models and similarity search (Van Laere et al., 2013) as baseline in combination with photo set refinement (photos within the same collection) and the user's location.

3.3.5 Georeferencing Systems at MediaEval 2014

The Media Eval 2014 Placing Task (MEPT2014) (Choi et al., 2014) introduced a web-scale geo-tagged dataset that contains 5.5 million photos and 35,000 videos. This large scale dataset addresses the research challenges in multimedia georeferencing.

The MEPT2014 dataset was extracted from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset[28][29] which contains the metadata for 99.2 million photos and 0.8 million videos that have been uploaded to Flickr with a Creative Commons license by the uploader. A subset of the geotagged videos of the YFCC100M (about a half of the dataset) were used to create the dataset. 5 million images and 25,000 videos were selected for the training set and 500,000 images and 10,000 videos for the test set. The selection was done with the following constraints: 1) each user only contributed at most 250 images and 50 videos, and that the recordings for a given user were all made more than 10 minutes apart from each other; 2) none of the users who contributed videos or images to the training set also contributed to the test set, and vice versa. This subsection presents all the approaches presented at MEPT2014 with the exception of the thesis author's approach (*TALP-UPC* (Ferrés and Rodríguez, 2014)) that is described in detail in Chapter 6. The results of these approaches are presented in Tables 3.6 and 3.7.

USEMP (Popescu et al., 2014) presented a text metadata based approach that used probabilistic place modeling and also machine tag and/or user modeling. The approach used was similar to their participation in MediaEval 2013 (Popescu, 2013). At MediaEval 2014 they used rectangular cells of size 0.01°of latitude and longitude degree (approximaltely of 1 km2 size) and compute the probability of a tag (taken from user tags and titles) in a cell by dividing its user count in that cell by its total user counts in all cells. The most probable cell for a given set of tags from the metadata of the image or video to georeference was found suming the contributions of individual tags. The user modelling technique computes the most probable cell of a user using 500 geotagged images per user. The machine tag modelling method models only machine tags that are strongly associated to locations. The fusion schema of the different modeling methods (machine tags, location and user) was done in the following way: machine tags are used in priority, if are not available then location models were used; finally if there are no tags available or the prediction score is below a threshold (empirically determined on the validation set) the user model is used by placing the photo to the most problable cell of the user who uploaded it. The user models were used for the 30% of test images which had the lowest placing scores. Their best results were achieved in when the three models (locations tags, machine tags, user tags) are used in combination and trained with all geotagged metadata from the YFCC100M dataset (removing all test items). *CEALIST* (Popescu et al., 2014) presented an approach similar to *USEMP* but trained with less amount data.

SonSens-CERTH (Kordopatis-Zilos et al., 2014) used a text metadata based approach in some of the runs of MEPT2014. Their approach is based on a geographical-tag model built from the tags and locations of the trainig set. This baseline approach uses a grid of

[28] `http://bit.ly/yfcc100md`

[29] YFCC100M Documentation. `http://arxiv.org/pdf/1503.01817v1`

rectangular cells with a side length of 0.01°for both latitude and longitude and the language model approach of Popescu (2013) with some extensions including: 1) similarity search, 2) internal grid, and 3) spatial tag entropy. Similarity search is based on Van Laere et al. (2011a). The internal grid extension was done for reliable prediction of finer granularities. Thus, a cell site of 0.001°for both latitude and longitude has been used to build a language model. Similarity search is be applied inside the finer (0.001°) instead of the coarse (0.01 °) cell depending if the estimate based on the finer granularity falls inside the coarse. The spatial tag entropy values are used to adjust the original languag model tag probabilities with a Gaussioan weight function.

In Run 1, using the language model, similarity search, internal grid and spatial entropy, they achieved accuracies of 23.02, 39.92 and 46.87 for margin of errors of 1, 10 and 100 kms. respectively. Run 4, using the language model and the center of cells as estimated location, hey achieved accuracies of 21.87, 38.96 and 46.13 for margin of errors of 1, 10 and 100 kms. respectively. And finally, Run 5, using the language model and similarity search they achieved accuracies of 22.24, 38.96, and 46.13 for margin of errors of 1, 10 and 100 kms. respectively.

The approach of *RECOD* (L. Li et al., 2014) at MediaEval 2014 combines runs with only textual features for prediction and another runs with a combination of textual and visual features using re-ranking and clustering approaches to geocode multimedia items based on the similarity of ranked lists. The textual runs used the following features of the metadata: title, description, and tags of photos/videos. They used the BM25 and TF-IDF retrieval models implemented by a Lucene index with stemming and stopwords filtering. In Run 1, they achieved best accuracies of a textual based approach with 21.04, 37.59 and 46.16 for margin of errors of 1, 10 and 100 kms. respectively.

The *UQ-DKE* approach (Cao et al., 2014) uses a language model-based document retrieval model (Metzler and Croft, 2004) in combination with a spatial-aware tag weighting schema to find the most similar item of the training set given a query (test item). The Flickr tags (excluding title and description) are used as a document. The query is constructed with the tags of the test item and a weighting process is applied to give different weight to each tag in the query. They used the Ripley's K statistic (Ripley, 2005) to calculate tag weighting and Bayesian Smoothing with Dirichlet priors is applied. They use also collection geo-correlation to predict test items without tags by using the most frequent location of well estimated test items within the same collection. In run 1 they applied the spatial aware tag weighting schema and a default location (a New York city coordinate) was used for test items without tags. In this run they achieved the accuracies of 19.57, 41.71 and 52.46 for margin of errors of 1, 10 and 100 kms. respectively. In run 3 ,spatial aware tag weighting and geo-correlation were applied. In this run they achieved the accuracies of 20.23 ,43.68 and 56.03 for margin of errors of 1, 10 and 100 kms. respectively. They showed that both methods improve geotagging accuracy by comparing with a baseline run. In Run 1, they achieved best accuracies of a textual based approach with 21.04, 37.59 and 46.16 for margin of errors of 1, 10 and 100 kms. respectively.

The *ICSI/TU Delft* system (Choi and X. Li, 2014) presented at MEPT2014 used two text-based approaches: spatial variance and graphical model framework. These algorithms used the user tags, title and also machine tags from the textual metadata. The spatial variance algorithm (Friedland et al., 2011) tries to find the estimation by finding the lowest spatial variance of the keywords. The graphical model framework (Choi et al., 2012) uses a Gaussian Mixture Model for the distribution of the location given a particular tag. In run

1 they applied the spatial variance algorithm with the following accuracies 16.65, 34.70 and 45.58 for margin of errors of 1, 10 and 100 kms. respectively. In run 3 they applied the graphical model framework algorithm with the following accuracies 16.28, 46.20 and 52.81 for margin of errors of 1, 10 and 100 kms. respectively.

Table 3.6: Media Eval 2014 Placing Task Run 1 results- Use provided dataset only. Percentage of correctly georeferenced photos/videos within several margins of kilometers.

	accuracy percentage						
System	10m	100m	1km	10km	100km	1000km	5000km
CEALIST	0.01	0.61	22.62	40.00	47.36	61.17	74.94
RECOD	0.55	**6.06**	21.04	37.59	46.14	61.69	76.76
SonSens-CERTH	0.50	5.85	23.02	39.92	46.87	60.11	74.80
TALP-UPC	0.29	4.12	16.54	34.34	51.06	**64.67**	**78.63**
UQ-DKE	**1.07**	4.98	19.57	**41.71**	**52.46**	63.61	77.28
USEMP	0.78	1.61	**23.48**	40.77	48.11	61.79	75.30
ICSI/TUDelft	0.24	3.15	16.65	34.70	45.58	60.67	75.03

Table 3.7: Media Eval 2014 Placing Task Overall best results - anything allowed except crawling the exact items of the test set. Percentage of correctly georeferenced photos/videos within certain margin of kilometers.

	accuracy percentage						
System	10m	100m	1km	10km	100km	1000km	5000km
CEALIST	0.01	1.22	40.25	55.98	62.26	72.14	81.95
RECOD	0.59	**6.26**	21.15	37.50	46.03	61.41	75.07
SonSens-CERTH	0.50	5.85	23.02	39.92	46.87	60.11	74.80
TALP-UPC	0.23	3.00	15.90	38.52	52.47	65.87	79.29
UQ-DKE	1.08	5.05	20.23	43.68	56.03	69.08	81.14
USEMP	**2.56**	4.33	**44.14**	**61.34**	**69.10**	**78.69**	**86.52**
ICSI/ TUDelft	0.32	3.41	12.13	19.95	22.82	33.79	53.06

3.3.6 Georeferencing Systems at MediaEval 2015-2016

The MediaEval 2015 and 2016 Placing Task (Choi et al., 2015; Choi et al., 2016) evaluation benchmarks presented new tasks slightly different from the original georeferencing task offered from 2010 to 2014. These new tasks were the following ones: i) estimation-based subtask (in 2015 and 2016), ii) mobility-based subtask (in 2015), and iii) verification-based subtask (in 2016). In the estimation-based subtask (which was originally called locale-based sub-task in 2015) participants were given a geographic hierarchy that ranged across neighbourhoods, cities, regions, countries, and continents. Given a photo or video to georeference, the subtask consisted into select the its hierarchy node. Participants could predict directly the coordinates or a node in the hierarchy depending on the confidence on the prediction. The mobility based subtask (introduced in 2015) consisted into predict the geographic reference in coordinates of some photos corresponding to a set of photos of a user taken in a certain city. On the other hand, the verification-based subtask (introduced in 2016) consisted into verify if a media item was captured in a given location. The evaluation of geographic distances is performed with Karney's formula (Karney, 2013). In both evaluations the organizers used sampled subsets of the YFCC100M for training and testing, see Table 3.8. The sampling was created with the following criteria: no user appeared both in the training set and in the test set, and each user was limited to contributing at most 250 photos and 50 videos, where no photos/videos were included that were taken by a user less than 10 minutes apart.

Table 3.8: Media Eval 2015-2016 Placing Task Training set.

	Training		Testing	
Task	#Photos	#Videos	#Photos	#Videos
MEPT2015 locale task	4,672,382	22,767	931,573	18,316
MEPT2016 estimation task	4,991,679	24,955	1,497,464	29,934

The run1 in both MEPT2015 and MEPT2016 required that participants use only textual metadata to predict the georeference. In MEPT2015 Duong-Trung et al. (2015), Kelm et al. (2015), Kordopatis-Zilos et al. (2015), and L. T. Li et al. (2015) submitted the run1. The best results were obtained by the approach of Kordopatis-Zilos et al. (2015) in the ranges 1m to 100Km, and by the approach of Kelm et al. (2015) in the ranges 1000km to 10000km. The approach of Kordopatis-Zilos et al., 2015 is based on the Popescu (2013) approach and uses a Language Model with a fine grid of cell side length of 0.001°and adaptable to use information from coarser grid cells, feature selection and similarity search. On the other hand, Kelm et al. (2015) presented and approach at MEPT2015 run1 that used a hierachical ranking model and text similarity with BM25. In MEPT2016 Kordopatis-Zilos et al. (2016), Muñoz et al. (2016), and Singh and Rafiei (2016) submitted the run1. Kordopatis-Zilos et al. (2016) obtained the best results (ranges from 10m to 100km). Their approach is based on the Popescu (2013) approach and uses a probabilistic Language Model using both coarse and finer rectangular cells of 0.01°x 0.01°and 0.001°x 0.001°with term-cell probabilities and feature selection. They also used the similarity search with the Jaccard similarity measure to get the georeference.

CHAPTER 4

Geographical Information Retrieval Approaches

This chapter describes three approaches to the Geographical Information Retrieval task in English and its evaluation in the context of the 2005, 2006 and 2007 GeoCLEF evaluations and posterior experiments with the full GeoCLEF test collections (from 2005 to 2008) in English (see Table 4.1 for some details and differences of these approaches). These approaches use Geographical Knowledge and some NLP techniques in order to improve their baseline results. The approaches are the following:

1. *GeoTALP-IR (2005).* The first approach, GeoTALP-IR (Ferrés et al., 2005a), uses a modified version of a Passage Retrieval module designed for Question Answering. This system uses a Keyword Selection algorithm based on a Linguistic and Geographical Analysis of the topics and a Geographical Thesaurus (GT) that has been build using a set of Geographical Gazetteers and a Geographical Ontology.

2. *TALPGeoIR (2006).* The second approach, TALPGeoIR 2006 (Ferrés and Rodríguez, 2007a) was a modified version of the GeoTALP-IR system presented at GeoCLEF 2005 (Ferrés et al., 2005a) with some changes in the retrieval modes and the Geographical Knowledge Base (KB). The TALPGeoIR 2006 used JIRS, a Passage Retrieval algorithm, to perform the IR phase. This approach also introduced the use of *Alexandria Digital Library (ADL) Feature Type Thesaurus* for GIR.

3. *TALPGeoIR (2007-2015).* The third approach, TALPGeoIR 2007 (Ferrés and Rodríguez, 2007b; Ferrés and Rodríguez, 2008a; Ferrés and Rodríguez, 2015a), is an evolution of the TALPGeoIR 2006 that uses the Terrier IR system and a custom build geographical index. Partial and conservative toponym disambiguation is applied in combination with automatic Query Expansion with Relevance Feedback.

Approach	Textual IR	Geo IR	GeoKB	NERC	Feature Types	Disambig.	QE-RF.
GeoTALP-IR 2005	Lucene	Lucene	Yes	Yes	No	No	No
TALP-GeoIR 2006	JIRS	Lucene	Yes	Yes	Yes	No	No
TALP-GeoIR 2007	Terrier	custom	Yes	Yes	Yes	Partial	Yes

Table 4.1: Details of the proposed and implemented approaches to Geographical IR.

4.1 GeoTALP-IR 2005 Approach

This section describes the GeoTALP-IR 2005 approach and its evaluation in the context of the GeoCLEF 2005 Geographical IR challenge. The system architecture of the GeoTALP-IR 2005 has two phases that are performed sequentially (as shown in Figure 4.1): Topic Analysis (TA) and Document Retrieval (DR). A collection pre-processing process was carried out in advance.

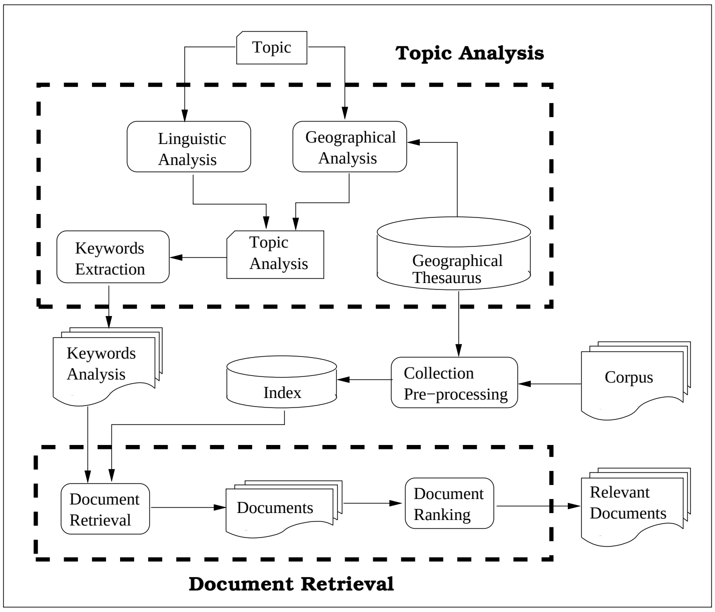


Figure 4.1: Architecture of GeoTALP-IR system.

4.1.1 Collection Pre-processing

The *Lucene*[1] Information Retrieval (IR) engine has been used to perform the DR task. The *Glasgow Herald* (1995) (GH95) and *Los Angeles Times* (1994) (LAT94) collections (i.e. 169,477 documents) were processed with linguistic tools (described in the next sub-section) to annotate the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process the collection is analyzed with a Geographical Thesaurus (described in the next

[1] http://jakarta.apache.org/lucene

sub-section). This information was used to build an index (see an example in Table 4.2) that contains the following fields for each document:

- **Form Field**: this field stores the original text (word forms) with the Named Entities recognized.
- **Lemma Field**: this part is built using the lemmas of the words, the POS tags, and the results of the Named Entity Recognition and Classification (NERC) module and the Geographical Thesaurus.
- **Geo Field**: it contains all NEs classified as *location* or *organization* that appear in the Geographical Thesaurus. This part has the geographical information about these NE: including geographical coordinates and geographical relations with the corresponding places of its path to the top of the geographical ontology (i.e. a city like “Barcelona” contains its state, country, sub-continent and continent). If a NE is an ambiguous location, all the possible ambiguous places are stored in this field.

Field	Indexed Content
Form	Watson flew off with his wife for a weekend in Barcelona, returned to London on Monday,
Lemma	Watson#NNP#PERSON fly#VBD off#RP with#IN his#PRP$ wife#NN for#IN a#DT weekend#NN in#IN Barcelona#NNP#LOCATION#city ,#, return#VBD to#TO London#NNP#LOCATION#capital on#IN monday#NNP ,#,
Geo	Europe#Europe#Spain#Cataluña#Barcelona#41.383_2.183 Europe#Europe#United_Kingdom#England#London#51.517_-0.105

Table 4.2: Example of an indexed document.

4.1.2 Topic Analysis

The goal of this phase is to extract all the relevant keywords from the topics enriching them as a result of the analysis. These keywords are then used by the Document Retrieval phase. The Topic Analysis phase has three main components: a Linguistic Analysis, a Geographical Analysis and a Keyword Selection algorithm.

4.1.2.1 Linguistic Analysis

This process extracts lexico-semantic and syntactic information using the following set of NLP tools:

- **Morphological components**, a statistical POS tagger (*TnT*) (Brants, 2000) and the WordNet 2.0 (Fellbaum, 1998) lemmatizer are used to obtain POS tags and lemmas. The *TnT* POS tagger used a pre-defined model trained on the Wall Street Journal corpus.
- **Spear**, a modified version of Collins parser that performs full parsing and robust detection of verbal predicate arguments (Collins, 1999). It is limited to three predicate arguments: agent, direct object (or theme), and indirect object (benefactive or instrument).

- **A Maximum Entropy based NERC**, a Named Entity Recognizer and Classifier that identifies and classifies NEs in basic categories (person, place, organization and other). This NERC has been trained with the CONLL-2003 shared task English data set (Tjong Kim Sang and De Meulder, 2003).
- **Gazetteers**, with the following information: location-nationality relations (e.g. Spain-Spanish) and actor-action relations (e.g. write-writer).

4.1.2.2 Geographical Analysis

The Geographical Analysis is applied to the Named Entities provided by the location tag (<EN-location>), and the Named Entities from the Title and Description tags that have been classified as *location* or *organization* by the NERC module. This analysis has two main components:

- **Geographical Thesaurus:** this component has been built joining three gazetteers that contain entries with places and their geographical class, coordinates, and other information: *GeoNet Names Server* (GNS), *Geographic Names Information System* (GNIS) (using only a subset of 39,906 of the most important geographical names), and *GeoWorldMap.*

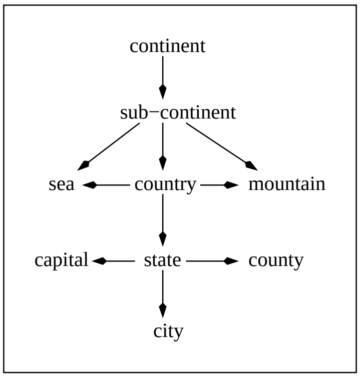


Figure 4.2: Geographical ontology.

Each one of these gazetteers have a different set of classes. These sets of classes have been mapped to our set of classes (see Figure 4.2), which includes the most common classes and the most important ones (e.g. country is not common, but is important). The resulting thesaurus contains approximately 3.7 million places with its geographical class. This approach is similar to that used in Manov et al. (2003), but they used a limited number of locations (only the 50,000 most important ones).

- **NEC correction filter**: a filter to correct some common errors in the *location-person* and *organization-person* ambiguity classes has been implemented. This filter stores all the NEs classified as *person* in the document; for each one of these NEs it extracts and stores in a hash table all the tokens that compose the NE. Then, for each NE of the document classified as *location* or *organization* it checks whether the NE exists in the document hash. If the NE exists then its class is changed to *person.*

4.1.2.3 Topic Keywords Selection

We designed an algorithm to extract the most relevant keywords of each topic (see an example in Table 4.3). These keywords are then passed to the Document Retrieval phase. The algorithm is applied after the Linguistic and Geographical analysis and has the following steps:

1. Initial Filtering. First, all the punctuation symbols and stopwords are removed from the analysis of the title, description and geographical tags.
2. Title Words Extraction. All the words from the title tag are obtained.
3. Description Chunks Filtering. All the Noun Phrase base chunks from the description tag that contain a word with a lemma that appears in one or more words from the title are extracted.
4. Description Words Extraction. The words belonging to the chunks extracted in the previous step and do not have a lemma appearing in the words of the title are extracted.
5. Append Title, Description and Location Words Analysis. The words extracted from the title and description and the geographical tag are appended.

Topic	EN-title	Environmental concerns in and around the Scottish Trossachs
	EN-desc	Find articles about environmental issues and concerns in the Trossachs region of Scotland.
	EN-location	the Scottish Trossachs
Keyword Selection	Title Stopword Filtering	Environmental concerns Scottish Trossachs
	Title Extracted words	Environmental, concerns, Scottish, and Trossachs
	Description Chunks	[environmental issues] [Trossachs region]
	Description Words Extraction	issues and region
	Selected Keywords	Environmental#environmental#JJ concerns#concern#NNS issues#issue#NNS region#region#NN scottish#Scottish#NNP#misc#location(”Scotland”) Trossachs#trossachs#NNP

Table 4.3: Keyword Selection example.

4.1.3 Document Retrieval

The Document Retrieval phase uses a modified version of the Passage Retrieval module of the TALP Question Answering (QA) system presented at CLEF 2004 (Ferrés et al., 2004a) and TREC 2004 (Ferrés et al., 2005c) with the *Lucene* (described in Chapter 3.1) IR

Topic ID	Topic Title
GC001	Shark Attacks off Australia and California
GC002	Vegetable Exporters of Europe
GC003	AI in Latin America
GC004	Actions against the fur industry in Europe and the U.S.A.
GC005	Japanese Rice Imports
GC006	Oil Accidents and Birds in Europe
GC007	Trade Unions in Europe
GC008	Milk Consumption in Europe
GC009	Child Labor in Asia
GC010	Flooding in Holland and Germany
GC011	Roman cities in the UK and Germany
GC012	Cathedrals in Europe
GC013	Visits of the American president to Germany
GC014	Environmentally hazardous Incidents in the North Sea
GC015	Consequences of the genocide in Rwanda
GC016	Oil prospecting and ecological problems in Siberia and the Caspian Sea
GC017	American Troops in Sarajevo, Bosnia-Herzegovina
GC018	Walking holidays in Scotland
GC019	Golf tournaments in Europe
GC020	Wind power in the Scottish Islands
GC021	Sea rescue in North Sea
GC022	Restored buildings in Southern Scotland
GC023	Murders and violence in South-West Scotland
GC024	Factors influencing tourist industry in Scottish Highlands
GC025	Environmental concerns in and around the Scottish Trossachs

Table 4.4: Topic titles of GeoCLEF 2005 in English.

engine API. The main function of the Document Retrieval component is to retrieve relevant documents that are likely to contain the information needed by the user. Document retrieval is performed using the *Lucene* Information Retrieval system. *Lucene* uses the standard TF-IDF weighting scheme with the cosine similarity measure, and it allows ranked and boolean queries. The document retrieval algorithm uses a data-driven query relaxation technique: if too few documents are retrieved, the query is relaxed by discarding the keywords with the lowest priority. The reverse happens when too many documents are extracted. Each keyword is assigned a priority using a series of heuristics fairly similar to Moldovan et al. (1999) (See Table 4.5). For example, a proper noun is assigned a higher priority than a common noun, the adverb is assigned the lowest priority, and stop words are removed.

The main options of the Document Retrieval phase are:

- Query types:
  - Boolean: all the keywords must appear in the documents retrieved. *Lucene* allows boolean queries and returns a score for each retrieved document.
  - Ranked: *Lucene* does ranked queries with tf-idf and cosine similarity.
  - Boolean+Ranked: this mode joins documents retrieved from boolean and ranked queries, giving priority to the documents from the boolean query.

Priority	Heuristic for current word (w)
9 (max)	stopword(w)==false AND has_quotes(w)
8	stopword(w)==false AND QFW(w)==false AND (isNNP(w) OR Number(w) OR Date(w))
7	isAdjNP(w) AND stopword(w)==false AND isQFW(w)==false AND isPunct(w)==false
6	isNonAdjNP(w) AND stopword(w)==false AND isQFW(w)==false AND isPunct(w)==false
5	isAdjective(w) AND stopword(w)==false AND isQFW(w)==false AND isPunct(w)==false
4	isNoun(w) AND stopword(w)==false AND isQFW(w)==false
3	(isVerb(w) || isAdverb(w)) AND stopword(w)==false AND isQFW(w)==false
2	isQFW(w)
1 (min)	stopword(w)==false AND isPunct(w)==false AND isInterrogativePronoun(w)==false

Table 4.5: Heuristics that assign keywords priority in Passage Retrieval

- Geographical Search Mode:
  - Lemma Field: this search mode implies that all the keywords that are Named Entities detected as *location* are searched in the "Lemma" field part of the index.
  - Geo Field: this search means that the NEs tagged as *location* and detected as keywords will be searched at the "Geo" index field.
- Geographical Search Policy:
  - Strict: this search policy can be enabled when the "Geo" Field search is running, and is used to find a *location* with exactly all this ontological path and coordinates for the following classes: country and region. In example, the form used to search "Australia" in the index is:
    Oceania#Oceania#Australia#-25.0_135.0
  - Relaxed: this search policy can also be enabled when the "Geo" field search is running. This mode searches without coordinates. The form used to search "Australia" in the index for this kind of search policy is:
    Oceania#Oceania#Australia
    In this case, the search is flexible and all the cities and regions of Australia will be returned. An example of a location found with the previous query is:
    Oceania#Oceania#Australia#Western_Australia#Perth#-31.966_115.8167

4.1.4 Document Ranking

This component joins the documents provided by the Document Retrieval phase. If the Query type is *boolean* or *ranked* it returns the first 1,000 top documents with their *Lucene* score. In the case of a query mode *boolean+ranked*, it first gives priority to the documents retrieved from the boolean query and holds their score. The documents provided by the ranked query are added to the list of relevant documents, but their score is then re-scaled using the score of the last boolean document retrieved (the document with lower score of the boolean retrieval). Finally, the first 1,000 top documents are selected.

4.1.5 Experiments and Results at GeoCLEF 2005

The GeoCLEF was a cross-language geographic retrieval task at the CLEF 2005 campaign. The goal of the task was to find as many relevant documents as possible from the document collections, using a topic set of 25 topics (see in Chapter 3.1 an example of a GeoCLEF 2005 topic). Topics are textual descriptions with the following fields: title, description, narrative, location (e.g. geographical places like continents, regions, countries, cities, etc.) and a geographical operator (e.g. spatial relations like in, near, north of, etc.). Table 4.4 presents the titles of the 25 topics of the GeoCLEF 2005.

A set of four experiments that consist in applying different query strategies and tags to an automatic GIR system (see Table 4.6) were designed to participate in the GeoCLEF 2005 GIR benchmark. Two baseline experiments were performed: the runs *geotalpIR1* and *geotalpIR2*. These runs differed uniquely in the Query type used: a *boolean+ranked* retrieval in *geotalpIR1* run and only *ranked* retrieval in *geotalpIR2* run. These runs consider the Title and Description (TD) tags, and they use the "lemma" index field. The third run (*geotalpIR3*) differs from the previous ones in the use of the Location tag (considering Title, Description and Location (TDL)) and uses the "Geo" field instead of the "lemma" field. The "Geo" field is used with a Strict Query search policy. This run also performs a *boolean+ranked* retrieval. The fourth run (*geotalpIR4*) is very similar to the third run (*geotalpIR3*), but uses a Relaxed Query search policy. On the other hand, the "spatialrelation" tag included in the topics (e.g. "south", "in", "near",...) is not used because the system treats all queries as "in" queries.

Table 4.6: Description of the Experiments at GeoCLEF 2005.

Run	Run type	Tags	Query Type	Geo. Index	Geo. Search
geotalpIR1	automatic	TD	Boolean+Ranked	Lemma	-
geotalpIR2	automatic	TD	Ranked	Lemma	-
geotalpIR3	automatic	TDL	Boolean+Ranked	Geo	Strict
geotalpIR4	automatic	TDL	Boolean+Ranked	Geo	Relaxed

The results of the GeoTalpIR system at the GeoCLEF 2005 Monolingual English task are summarized in Table 4.7 and Figure 4.3. This table shows the following IR measures for each run: *Average Precision*, *R-Precision*, *Recall*, and the increment over the median of the average precision (0.2063) obtained by all the systems that participated in the GeoCLEF 2005 Monolingual English task.

Table 4.7: GeoCLEF 2005 results.

Run	Tags	MAP	R-Prec.	Recall (%)	Recall	Δ MAP Diff.(%) over GeoCLEF avg. MAP
geotalpIR1	TD	0.1923	0.2249	49.51%	509/1028	-6.78%
geotalpIR2	TD	0.1933	0.2129	49.22%	506/1028	-6.30%
geotalpIR3	TDL	0.2140	0.2377	62.35%	641/1028	+3.73%
geotalpIR4	TDL	**0.2231**	**0.2508**	**66.83%**	**687/1028**	**+8.14%**

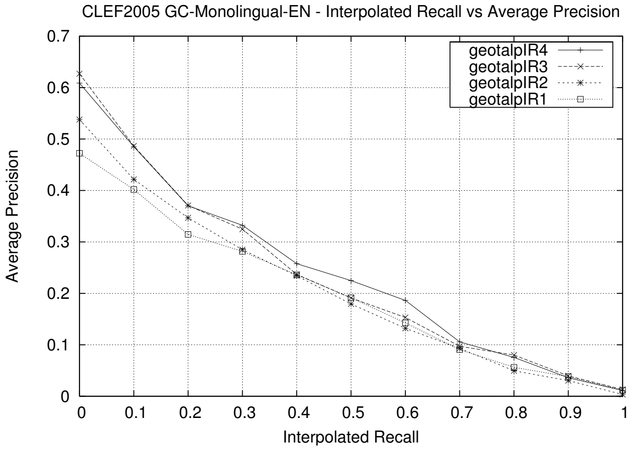


Figure 4.3: Interpolated Recall vs Average Precision.

The results show a substantial difference between the effectivenes measures of the two first runs and the two last ones, specially in the recall measure: 49.51% and 49.22% respectively in the first and second run (*geotalpIR1* and *geotalpIR2*) and 62.35% and 66.38% respectively in the third and fourth run (*geotalpIR3* and *geotalpIR4*). The recall is also improved by the use of Geographical Knowledge and a relaxed policy over the "Geo" Field as it is seen in run four (*geotalpIR4*). Finally, in the last run (*geotalpIR4*) we obtained results about +8.14% better than the median of the average obtained by all runs (0.2063). The use of the "Location" field by the last two runs means that the Geographical place names appearing in the Title and the Description detected by the Toponym Recognition are not taken into account to find the geographical relevance of the topics. Although the use of the geographical index field index has shown an improvement of the effectiveness measures, in comparison with the top ranked systems the results of GeoTALP-IR system in MAP (0.2231) are low. The three top ranked approaches achieved a MAP of 0.3936 (F. Gey et al., 2005), 0.3613 (Guillén, 2005) and 0.3495 (Ferrández et al., 2005) respectively.

4.2 TALPGeoIR 2006 Approach

This section describes the TALPGeoIR 2006 approach and the experiments and results of the system in the context of our participation in the CLEF 2006 GeoCLEF Monolingual English task. The TALPGeoIR system is a modified version of the GeoTALP-IR system presented at GeoCLEF 2005 (Ferrés et al., 2005a) with some changes in the retrieval modes and the Geographical Knowledge Base (KB) and the inclusion of a feature type thesaurus. The experiments with the GeoTALP-IR in the context of the GeoCLEF 2005 showed that the Geographical Knowledge improved the effectiveness measures but our results were low with respect to the top ranked approaches.

For this reason the following changes were done: 1) the textual IR retrieval system has been changed from Lucene to JIRS in order to test JIRS, a Passage Retrieval QA-oriented system, in a GIR challenge, 2) the Geographical Knowledge Base has been improved with the inclusion of the *World-Gazetteer* and the extension and mapping of the allowed feature types to the ones used in the *Alexandria Digital Libray Feature Type Thesaurus*.

The TALPGeoIR 2006 has four phases performed sequentially: 1) a Keywords Selection algorithm based on a linguistic and geographical analysis of the topics, 2) a geographical document retrieval with *Lucene*, 3) a textual document retrieval with the *JIRS* Passage Retrieval (PR) software, and 4) a Document Ranking phase.

4.2.1 Collection Processing

The Glasgow Herald 1995 (GH95) and Los Angeles Times 1994 (LAT94) collections (i.e. 169,477 documents) were processed with linguistic tools (described in the next sub-section) to annotate the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process the collection is analyzed with a Geographical Thesaurus (described in the next sub-section). This information was used to build two indexes: one with the geographical information of the documents and another one with the textual and geographical information of the documents. Two Information Retrieval (IR) systems were used to create these indexes: *Lucene* for the geographical index and *JIRS* for the textual and geographical index (see a sample of both indexes in Table 4.8). These indexes are described below:

- **Geographical Index**: this index contains the geographical information of the documents and its Named Entities. The Geographical index contains the following fields for each document:
  - **docid**: this field stores the document identifier.
  - **ftt**: this field indexes the feature type of each geographical name and the Named Entity classes of all the NEs appearing in the document.
  - **geo**: this field indexes the geographical names and the Named Entities of the document. It also stores the geographical information (hierarchical ancestors' path, and coordinates) about the place names. Even if the place is ambiguous all the possible referents are indexed.
- **Textual and Geographical Index**: this index stores the lemmatized content of the document and the geographical information (feature type, hierarchical ancestors' path, and coordinates) about the geographic place names appearing in the text. If

the geographical place is ambiguous then this information is not added to the indexed content.

System	Indexed Content	
Lucene	docid	GH950102000000
	ftt	regions@land_regions@continents administrative_areas@political_areas@countries_1st_order_divisions administrative_areas@populated_places@cities administrative_areas@political_areas@countries ...
	geo	Europe Asia@Western_Asia@Saudi_Arabia@Hejaz@24.5_38.5 America@Northern_America@United_States@South_Carolina @Lodge@32.9817_-80.952 America@Northern_America@United_States@38.91_-96.19 ...
JIRS	...the role of the wheel in lamatrekking , and where be the good place to air your string vest. pity the crew who accompany him on his travel as sayle of Arabia *countries_1st_order_divisions* **Asia Western_Asia Kuwait Arabia 25.0_45.0** along the Hejaz *countries_1st_order_divisions***Asia Western_Asia Saudi_Arabia Hejaz 24.5_38.5** railway line from Aleppo *countries_1st_order_divisions* **Asia Middle_East Syria Aleppo 36.0_37.0** in Northern_Syria *countries* **Asia Middle_East Syria 35.0_38.0** to Aqaba *cities* **Asia Western_Asia Jordan Maán Aqaba 29.517_35** in Jordan *countries* **Asia Western_Asia Jordan 31.0_36.0**. as he journey through the searing heat in an age East German ‘ biscuit tin ‘ , his good humour be sorely test ...	

Table 4.8: Samples of an indexed document with *Lucene* and *JIRS*.

4.2.2 Topic Analysis

The goal of this phase is to extract all the relevant keywords (with its analysis) from the topics. These keywords are then used by the document retrieval phases. The Topic Analysis phase has three main components: a Linguistic Analysis, a Geographical Analysis, and a Keyword Selection algorithm.

4.2.2.1 Linguistic Analysis.

This process extracts lexico-semantic and syntactic information using the same NLP tools employed at the GeoTALP-IR 2005 system described in Section 4.1.2: i) *TnT* ii) *WordNet 2.0 lemmatizer*, iii) *Spear*.

4.2.2.2 Geographical Analysis.

The Geographical Analysis is applied to the NEs from the title, description, and narrative tags that have been classified as *location* or *organization* by the NERC tool. This analysis has two components:

- **Geographical Knowledge Base:** this component has been built joining four geographical gazetteers: *GEOnet Names Server (GNS)*, *Geographic Names Information System (GNIS)* (using only a subset of 39,906 entries with the most important places), *GeoWorldMap Gazetteer*, and the *World Gazetteer* (adding only the 29,924 cities with more than 5,000 inhabitants).
- **Geographical Feature Type Thesaurus**: the feature type thesaurus of our Geographical KB is the *ADL Feature Type Thesaurus (ADLFTT)*. The *ADL Feature Type Thesaurus* is a hierarchical set of geographical terms used to type named geographic places in English (Hill, 2000). Both *GNIS* and *GNS* gazetteers have been mapped to the *ADLFTT*, with a resulting set of 575 geographical types. Our *GNIS* mapping is similar to the one exposed in Hill (2000).

4.2.2.3 Topic Keywords Selection.

This algorithm extracts the most relevant keywords of each topic. The algorithm was designed for GeoCLEF 2005 (Ferrés et al., 2005a). Once the keywords are extracted, three different Keyword Sets (KS) are created (see an example in Table 4.9):

- **All**: all the keywords extracted from the topic tags.
- **Geo**: geographical places or feature types appearing in the topic tags.
- **NotGeo**: all the keywords extracted from the topic tags that are not geographical place names or geographical types.

Topic	EN-title	Wine regions around rivers in Europe
	EN-desc	Documents about wine regions along the banks of European rivers.
	EN-narr	Relevant documents describe a wine region along a major river in European countries. To be relevant the document must name the region and the river.
Keywords Set (KS)	Not Geo	wine European
	Geo	Europe#location#regions@land_regions@continents#Europe regions hydrographic_features@streams@rivers
	All	wine regions rivers European Europe

Table 4.9: Keyword sets sample of Topic 026.

4.2.3 Geographical Document Retrieval with *Lucene*

Lucene is used to retrieve geographically relevant documents given a specific Geographical IR query. *Lucene* uses the standard TF-IDF weighting scheme with the cosine similarity measure and allows ranked and boolean queries. We used boolean queries with a *Relaxed geographical search policy* (as explained in the previous section (Ferrés et al., 2005a)). This search policy allows to retrieve all the documents that have a token that matches totally or partially (a sub-path) the geographical keyword. As an example, the keyword America@Northern_America@United_States will retrieve all the U.S. places (e.g. America@Northern_America@United_States@Ohio).

4.2.4 Document Retrieval using the *JIRS* Passage Retriever

The *JIRS* Passage Retrieval System (Soriano et al., 2005) is used to retrieve relevant documents related to a GIR query. *JIRS* is a Passage Retriever specially designed for QA. This system gets passages with a high similarity between the largests n-grams of the question and the ones in the passage. *JIRS* has the following steps executed sequentially: 1) a Passage Retrieval phase that gets a set of relevant passages using a set of keywords from the question, and 2) a search for all the n-grams of the question in the retrieved passage and weighting of them using the number and the weight of the n-grams appearing in the passages Soriano et al., 2005. In the specific context of our GIR system, *JIRS* is used considering a topic keyword set as a question. Then, a set of relevant passages are retrieved using the n-gram distance model of *JIRS* with a length of 11 sentences per passage. The first 100,000 top-scored passages per topic are obtained. Finally, a process selects the relevant documents from the set of retrieved passages. Two document scoring strategies were used:

- **Best**: the document score is the score of the top-scored passage in the set of the retrieved passages that belong to this document.
- **Accumulative**: the document score is the sum of the scores of all the retrieved passages that belong to this document.

4.2.5 Document Ranking

This component ranks the documents retrieved by *Lucene* and *JIRS*. First, the top-scored documents retrieved by *JIRS* that appear in the document set retrieved by *Lucene* are selected. Then, if the set of selected documents is less than 1,000, the top-scored documents of *JIRS* that don't appear in the document set of *Lucene* are selected with a lower priority than the previous ones. Finally, the first 1,000 top-scored documents are selected. On the other hand, when the system uses only *JIRS* for retrieval, only the first 1,000 top-scored documents by *JIRS* are selected.

4.2.6 Experiments and Results at GeoCLEF 2006

GeoCLEF 2006 was a cross-language geographic retrieval task at the CLEF 2006 campaign. Like the first GIR task in GeoCLEF 2005 (F. Gey et al., 2005), the goal of the GeoCLEF task was to find as many relevant documents as possible from the document collections, using a topic set of 25 topics. Topics are textual descriptions with the following fields: title, description, narrative. In GeoCLEF 2006 the fields "location", "concept" and "spatialrelation" (that were available in GeoCLEF 2005) are not included in the topics. See in Figure 4.4 an example of one topic of the GeoCLEF 2006 test collection and the complete list of topic titles in Table 4.10.

```
<num> GC034 </num>
<EN-title> Malaria in the tropics </EN-title>
<EN-desc> Malaria outbreaks in tropical regions and preventive
vaccination </EN-desc>
<EN-narr> Relevant documents state cases of malaria in tropical regions
and possible preventive measures like chances to vaccinate against the
disease. Outbreaks must be of epidemic scope. Tropics are defined
as the region between the Tropic of Capricorn, latitude 23.5 degrees
South and the Tropic of Cancer, latitude 23.5 degrees North. Not relevant
are documents about a single person's infection.</EN-narr>
```

Figure 4.4: Example of two topics of the GeoCLEF 2006 edition.

Topic ID	Topic Title
GC026	Wine regions around rivers in Europe
GC027	Cities within 100km of Frankfurt
GC028	Snowstorms in North America
GC029	Diamond trade in Angola and South Africa
GC030	Car bombings near Madrid
GC031	Combats and embargo in the northern part of Iraq
GC032	Independence movement in Quebec
GC033	International sports competitions in the Ruhr area
GC034	Malaria in the tropics
GC035	Credits to the former Eastern Bloc
GC036	Automotive industry around the Sea of Japan
GC037	Archeology in the Middle East
GC038	Solar or lunar eclipse in Southeast Asia
GC039	Russian troops in the southern Caucasus
GC040	Cities near active volcanoes
GC041	Shipwrecks in the Atlantic Ocean
GC042	Regional elections in Northern Germany
GC043	Scientific research in New England Universities
GC044	Arms sales in former Yugoslavia
GC045	Tourism in Northeast Brazil
GC046	Forest fires in Northern Portugal
GC047	Champions League games near the Mediterranean
GC048	Fishing in Newfoundland and Greenland
GC049	ETA in France
GC050	Cities along the Danube and the Rhine

Table 4.10: Topic titles of GeoCLEF 2006 in English.

For GeoCLEF 2006, a set of five experiments was designed. These experiments consist in applying different IR systems, query keyword sets, and tags (see Table 4.11). Basically, these experiments can be divided in two groups depending on the retrieval engines used:

- **JIRS**. There are two baseline experiments: the runs *TALPGeoIRTD1* and *TALPGeoIRTDN1*. These runs differ uniquely in the use of the narrative tag in the second one. Both runs use one retrieval system, *JIRS*, and they use all the keywords to perform the query. The experiment *TALPGeoIRTDN3* is similar to the previous ones but uses a Cumulative scoring strategy to select the documents with *JIRS*.
- **JIRS & Lucene**. The runs *TALPGeoIRTD2* and *TALPGeoIRTDN2* use *JIRS* for

textual document retrieval and *Lucene* for geographical document retrieval. Both runs use the *Geo* keywords set for *Lucene* and the *NotGeo* keywords set for *JIRS*.

Table 4.11: Description of the experiments at GeoCLEF 2006.

Automatic Runs	Tags	IR System	JIRS KS	Lucene KS	*JIRS* Score
TALPGeoIRTD1	TD	JIRS	All	-	Best
TALPGeoIRTD2	TD	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTDN1	TDN	JIRS	All	-	Best
TALPGeoIRTDN2	TDN	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTDN3	TDN	JIRS	All	-	Cumulative

The results of the TALP-GeoIR system at the CLEF 2006 GeoCLEF Monolingual English task are summarized in Table 4.12 and Figure 4.5. This table has the following IR measures for each run: *MAP*, *R-Precision*, and *Recall*.

Table 4.12: TALP-GeoIR results at GeoCLEF 2006 Monolingual English task.

Automatic Runs	Tags	IR System	MAP	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	JIRS	**0.1342**	**0.1370**	60.84%	230/378
TALPGeoIRTD2	TD	JIRS+Lucene	0.0766	0.0884	32.53%	123/378
TALPGeoIRTDN1	TDN	JIRS	0.1179	0.1316	**68.78%**	**260/378**
TALPGeoIRTDN2	TDN	JIRS+Lucene	0.0638	0.0813	47.88%	181/378
TALPGeoIRTDN3	TDN	JIRS	0.0997	0.0985	64.28%	243/378

The results show a substantial difference between the two sets of experiments. The runs that use only *JIRS* have a better *MAP*, *R-Precision*, and *Recall* than the ones that use *JIRS* and *Lucene*. The run with the best *MAP* is *TALPGeoIRTD1* with 0.1342. The best *Recall* measure is obtained by the run *TALPGeoIRTDN1* with a 68.78% of the relevant documents retrieved. This run has the same configuration that the*TALPGeoIRTD1* run but uses the narrative tag. Finally, we obtained poor results in comparison with the average MAP (0.1975) obtained by all the systems that participated in the GeoCLEF 2006 Monolingual English task.

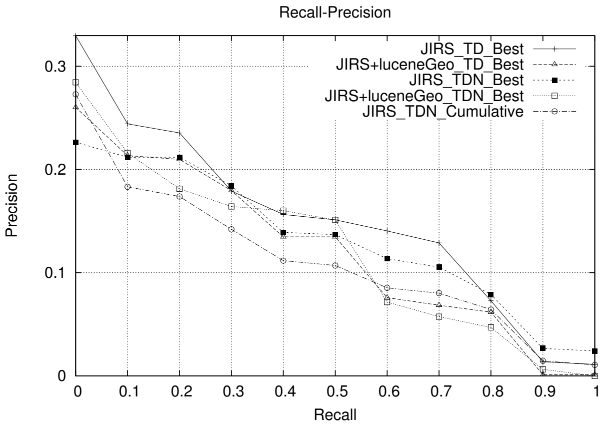


Figure 4.5: Interpolated Recall vs Average Precision.

The approach with only *JIRS* was better than the one with *JIRS* and *Lucene* combined. This happens because in the JIRS+Lucene runs the Non-Geo keywords have been used to retrieve the documents from JIRS in a Textual based IR instead of using the full set of keywords as in the runs that use only JIRS. Comparatively with the Mean Average Precision (MAP) of all the runs participating at GeoCLEF 2006 Monolingual English task our best results in MAP are very low (0.1342). This fact can be due to several reasons: 1) the *JIRS* PR was originally designed for the Question Answering task and maybe it was not used appropriately for GIR or is not suitable for GIR, 2) our system is not dealing with geographical ambiguities, 3) the lack of textual query expansion methods, 4) the need of Relevance Feedback methods, and 5) errors in the Topic Analysis phase.

4.2.7 Experiments after GeoCLEF 2006

After GeoCLEF 2006 a set of experiments were performed to evaluate the impact of using different sets of keywords (All. and NotGeo) when using the JIRS IR system for Textual IR (see Table 4.13) alone or in combination with Lucene. Some of these experiments had been already performed in the GeoCLEF 2006 evaluation but were repeated and included again here for comparison purposes.

Table 4.13: Description of the experiments at GeoCLEF 2006.

Automatic Runs	Tags	IR System	JIRS KS	Lucene KS	*JIRS* Score
TALPGeoIRTD1	TD	JIRS	All	-	Best
TALPGeoIRTD2	TD	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTD3	TD	JIRS+Lucene	All	Geo	Best
TALPGeoIRTDN1	TDN	JIRS	All	-	Best
TALPGeoIRTDN2	TDN	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTDN3	TDN	JIRS	All	-	Cumulative
TALPGeoIRTDN4	TDN	JIRS+Lucene	All	Geo	Best
TALPGeoIRTDN5	TDN	JIRS+Lucene	All	Geo	Cumulative

Table 4.14: TALP-GeoIR results after GeoCLEF 2006 Monolingual English task.

Automatic Runs	Tags	IR System	MAP	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	JIRS	0.1342	0.1370	60.84%	230/378
TALPGeoIRTD2	TD	JIRS+Lucene	0.0805	0.0978	44.17%	167/378
TALPGeoIRTD3	TD	JIRS+Lucene	**0.1531**	**0.1707**	64.02%	242/378
TALPGeoIRTDN1	TDN	JIRS	0.1179	0.1316	68.78%	260/378
TALPGeoIRTDN2	TDN	JIRS+Lucene	0.1106	0.1228	61.37%	232/378
TALPGeoIRTDN3	TDN	JIRS	0.0997	0.0985	64.28%	243/378
TALPGeoIRTDN4	TDN	JIRS+Lucene	0.1222	0.1343	**69.31%**	**262/378**
TALPGeoIRTDN5	TDN	JIRS+Lucene	0.1050	0.1020	66.67%	252/378

The results of these experiments are summarized in Table 4.14. The results show that the experiments that combined JIRS and Lucene using all the keywords in JIRS (i.e. All set) and the geographical keywords for Lucene (i.e. Geo keywords set) outperformed the experiments that use JIRS alone. It also has to be noted that during the experiments it was detected that the run that used JIRS and Lucene with the NotGeo keywords set (experiments *TALPGeoIRTD2* and *TALPGeoIRTDN2*) had an small error in the official GeoCLEF 2006 experiments and its results have changed in these new experiments. The best MAP achieved was 0.1531 and compared with the official results (best MAP was 0.3034) was still a very low MAP.

4.3 TALPGeoIR 2007 Approach

The TALPGeoIR 2007 GIR system is a modified version of TALPGeoIR 2006 system presented at GeoCLEF 2006 (Ferrés and Rodríguez, 2006b) with some changes in the retrieval modes and the Geographical Knowledge Base. The system has four phases performed sequentially: i) a Linguistic and Geographical Analysis of the topics, ii) a thematic Document Retrieval with Terrier, iii) a Geographical Retrieval task with Geographical Knowledge Bases (GKBs), and iv) a Document Filtering phase. In addition, a toolbox based on 'shape files'[2] for countries (Pouliquen et al., 2004) has been added to the system.

The system is composed of two main phases: 1) Textual and Geographical Indexing, 2) Geographical Information Retrieval. The IR software used in both indexing and retrieval phases is Terrier (version 4.0) (Ounis et al., 2006). We used the TF-IDF. BM25, and InL2 IR algorithms implemented in the Terrier IR engine. Stopwords filtering is applied by our system using the stopwords list provided in the Terrier IR engine. The baseline system uses all the terms from the topics. This means that no separation between thematic and geographical terms and themes or events is performed by the textual search.

4.3.1 Textual and Geographical Indexing

The GH95 and LAT94 collections (used in the previous GeoCLEF evaluations) were processed with linguistic tools (described in the next sub-section) to annotate the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process the collection is analyzed with a Geographical Knowledge Base and conservative Toponym Disambiguation heuristics (both components are described in the next sub-section). This information was used to built two types of indexes:

- Geographical Index. This is a custom-build index that contains the geographical information of the documents. For each toponym in the document (detected with the NERC detector) the feature type, GeoKB ontology information and coordinates are stored in the index. Even if the place is ambiguous all the possible geographical referents are indexed.

- Textual Indexes. These are Terrier based indexes that store the original or the linguistically processed information of the document. Note that in all these indexes geographical entities (toponyms) have been indexed without linguistic processing with exception of the stemmed indexes. The following indexes have been created: 1) *original index with word forms*, 2) *lemmatized index*, 3) *stemmed index* (using the Porter Stemmer, and 4) *lemmatized and stemmed index* (the Porter Stemmer applied over the lemmatized content).

4.3.2 Geographical Information Retrieval

The retrieval system has four phases performed sequentially: 1) a Linguistic and Geographical Processing of the topics, 2) a textual Document Retrieval with Terrier, 3) a Geographical Document Retrieval with Geographical Knowledge Bases (GKBs), and 4) a Geographical Re-Ranking phase.

[2] http://www.esri.com

4.3.2.1 Linguistic and Geographical Knowledge Processing of the topics and the collections

The goal of this phase is to extract all the relevant keywords (with its analysis) from the topics. These keywords are then used by the Textual and Geographical Document Retrieval phases. The Topic Analysis phase has two main sub-phases: a Linguistic Analysis and a Geographical Analysis. The Linguistic Analysis sub-phase extracts lexico-semantic and syntactic information using the following set of Natural Language Processing (NLP) tools: 1) *TnT* an statistical POS tagger (Brants, 2000), 2) *WordNet lemmatizer* (version 2.0), 3) A Maximum Entropy based NERC trained with the CONLL-2003 shared task English data set, 4) a list of demonyms relationships for each country (e.g. Japanese - Japan). The Geographical Analysis is applied to the Named Entities from the Title and Description and Narrative tags of the topics that have been classified as LOCATION or ORGANIZATION by the NERC module. This analysis uses a Geographical Knowledge Base that has two main components: a Geographical Thesaurus and a Feature type thesaurus.

The Geographical Thesaurus used is the same used in the TALPGeoIR 2006 approach (see Section 4.2.2). This thesaurus has been built joining four gazetteers (GNS, GNIS, GeoWorldMap, and World Gazetteer) that contain entries with places and their geographical class, coordinates, part-of relationships and other information. Each one of these gazetteers has a different set of classes that have been mapped to the ADL Feature Type Thesaurus (ADLFTT) with a resulting set of 575 geographical types. The ADL Feature Type Thesaurus is a hierarchical collection of geographical terms used to type named geographic places in English (Hill, 2000). The GNIS mapping is similar to the one exposed in Hill (2000).

The following Toponym Disambiguation heuristics are applied using the information from the GeoKB:

- *H1. Hierarchical ranked ontology of feature types.* The ranked hierarchy of the feature types ontology is applied when a toponym can refer to several kinds of feature types (e.g. Africa (the continent) vs Africa, Mexico). The following list of ordered priorities for the different feature types is used: 1) continent, 2) subcontinent (e.g. South America), 3) country capital, 4) country, 5) first order administrative divisions (e.g. states), 6) sea, 7) summit, 8) river, 9) county, 10) important city , 11) other place (can include less important cities and other types).

- *H2. Important places are disambiguated excluding other places with the same name.* The data of the GeoWorldMap and the Word Gazetteer has priority to disambiguate places because contains less but important places compared with GNIS and GNS.

- *H3. Treatment of toponym vs person name type of Geo/Non-Geo ambiguity when the toponym has the lowest priority (12).* A list of common first and last names is used to filter out Named Entities erroneously recognized as toponyms.

- *H4. Small places are not taken into account (only for USA).* Due to the high amount of places in the GNIS (USA only) gazetteer, only a small part of its data is used (the US concise gazetteer).

- *H5. Lowest priority toponyms are not disambiguated.* Toponyms with the lowest priority in the hierarchy are not disambiguated and all the possible geographical referents are taken into account in the collection processing and indexing, and the topic analisys phases.

These processes are applied to the topics but have been applied also to the entire document collection before indexing. The GeoKB and the Toponym Disambiguation processes take into account the part-of relationships of the toponyms detected and are used in the retrieval and indexing process (e.g. the toponym "United States" is indexed as *America@North_America@United_States*). Geographical coordinates (point-based) for each toponym are also included in the index with exception to the continent and subcontinent feature types. The feature types of each toponym disambiguated is also detected and stored (e.g the toponym "United States" will have the following feature type associated *administrative_areas@political_areas@countries*).

In addition to the gazetteers and the ADL feature type thesaurus, a Shape Files toolbox has been implemented. This idea has been inspired by the work of Pouliquen et al. (2004), that propose the use of a publicly available database of 'shape files' for countries. The shape file used is the ESRI First Level World Administrative Boundaries 1998. The ESRI 1998 shapefile contains the political boundaries for each country, and states/provinces within each country as of 1998.[3] Each 'shape file' contains a set of non overlapping regions (represented as polygons), each one consisting of a set of points (X-Y coordinates) representing the 'border' of the area. For most countries the 'shape file' contains only one area but some of them contain more than one, for instance, Italy contains 22 areas (the continental area and several islands). In order to cope with 'shape files' a toolbox was implemented to obtain the following information:

- Obtaining the border points of a country.
- Detecting if a point belongs to a country or area.
- Obtaining a polygon which encodes a certain area of a country using a 9-grid zone division (North, North-West, North-East, West, Central, East, South, South-West, Sout-East).
- Getting the border points around a point P at a distance D.
- Getting near points around a point P.

To deal with the toolboox operations, each country is assigned several areas corresponding to unconnected areas of its territories. One of these areas is assigned to be the main zone and this zone is used as default one when no information points to another zone. In our setting we compute the area of each zone and consider the main zone as the most extense one. For instance, for Spain the main zone corresponds to the Spanish territory within the Iberian peninsula while other small zones correspond to each of the Cantary and Balearic Islands, Spanish cities in Morocco and so... For dealing with expression as "in the north of Italy" we further divide each area into 9 subzones (9-grid) using the covering rectangle of the zone into 3 horitzonal and 3 vertical subzones.

[3] ESRI Admin98 shapefile.`http://geodata.grid.unep.ch/download/admin98_li_shp.zip`

4.3.2.2 Textual Document Retrieval

The textual IR phase is performed retrieving the top 10,000 documents related to the topic using the Terrier IR software. The default stopwords in English of the IR engine Terrier are used. This phase can perform Stemming (Porter's algorithm) and automatic Query Expansion (QE) using two state-of-the art Query Expansion models based on Divergence From Randomness: Bose-Einstein 1 (Bo1) and Kullback-Leibler (KL) (Amati, 2003). This pseudo-relevance feedback option extracts the T most informative terms from the X top-returned documents in first-pass retrieval as the expanded query terms [4].

The Bose-Einstein 1 term weighting model uses Bose-Einstein statistics Amati, 2003 to weight the terms in the top-returned documents. The Equation 1 shows the weight of a term t in the top X documents: tf_X is the term frequency in the top X documents, p_tc is the probability of the term t in the collection c, and tf_N is the frequency of the term t in the collection c (being N the number of documents in the collection).

$$p_tc = \frac{tf_N}{N} \qquad (eq.\ 1)$$

$$weight(t) = tf_X * log_2(\frac{1+p_tc}{p_tc}) + log_2(1+p_tc) \qquad (eq.\ 2)$$

The Kullback-Liebler term weighting model uses the Kullback-Liebler divergence (Amati, 2003) to get the most relevant terms. The Equation 4 shows the weight of a term t in the top X documents: tf_X is the term frequency in the top X documents, p_tx is the probability of the term t in the collection X (being x the number of documents in the collection X), p_tc is the probability of the term t in the collection c, and tf_N is the frequency of the term t in the collection c (being N the number of documents in the collection).

$$p_tx = \frac{tf_X}{x} \qquad (eq.\ 3)$$

$$weight(t) = \begin{cases} 0 & \text{if } p_tx < p_tc \\ p_tx * log_2(\frac{p_tx}{p_tc}) & \text{otherwise} \end{cases} \qquad (eq.\ 4)$$

The Terrier IR weighting models used in the experiments are the TF-IDF, BM25 and InL2.

- **TF-IDF**: The TF-IDF algorithm implemented in Terrier is slightly different from the original TF-IDF. The term frequency tf parameter is given by Robertson's tf. The *idf* parameter is given by the standard Sparck Jones' *idf* formula (Sparck-Jones, 1972). The current document length and the average document length in the collection are used also as a parameters. The original constants $k : 1$ and b are set by default at $k_1 = 1.2$ and $b = 0.75$ respectively. The $keyFrequency$ is the term frequency in the query.

$$Robertson_tf = k_1 * tf/(tf + k_1 * (1 - b + b * docLength/averageDocumentLength)) \quad (eq.\ 5)$$

[4] The values of X=10 (top returned documents) and T=40 (most informative terms) were used in the experiments

$$idf = log(numberOfDocuments/TermDocFrequency + 1) \qquad (eq.\ 6)$$

$$Score(t, q, d) = keyFrequency * Robertson_tf * idf \qquad (eq.\ 7)$$

- **BM25**: The BM25 Terrier's implementation of the Okapi BM25 weighting model was used. The default parameters used are: $k_1 = 1.2$, $k_3 = 8$, and $b = 0.75$.

$$K = k_1 * ((1 - b) + b * docLength/averageDocumentLength) + tf \qquad (eq.\ 8)$$

$$idf = log((numberOfDocuments - TermDocFrequency + 0.5)/(TermDocFrequency + 0.5)) \qquad (eq.\ 9)$$

$$Score(t, q, d) = (tf * (k_3 + 1) * keyFrequency/((k_3 + keyFrequency) * K)) * idf \qquad (eq.\ 10)$$

- **InL2**: Inverse document frequency model for randomness (Amati, 2003). The default parameters used are: $c = 1$.

$$TF = tf * log(1 + (c * avgDocLength)/docLength) \qquad (eq.\ 11)$$

$$NORM = 1d/(TF + 1d); \qquad (eq.\ 12)$$

$$idfDFR(tDF) = log((numberOfDocuments + 1)/(tDF + 0.5)) * (1/log(2))); \qquad (eq.\ 13)$$

$$Score(t, q, d) = TF * idfDFR(TermDocFrequency) * keyFrequency * NORM; \qquad (eq.\ 14)$$

4.3.2.3 Geographical Document Retrieval

The Geographical Document Retrieval can be performed with two approaches: Geographical Knowledge Based Retrieval and Border Filtering Retrieval. The first approaches used the Geographical Knowledge Base to retrieve geographically relevant documents using the following types of geographical terms from GIR queries: 1) toponyms, 2) feature types (e.g. "cities","countries", ...). The GeoKB uses a relaxed search policy method over toponyms and feature types that allows to retrieve all the documents that have a token that matches totally or partially the toponyms or the feature types appearing in the topic. As an example for the case of toponyms, the keyword *America@Northern_America@United_States* will retrieve U.S. places like Los Angeles, CA, USA and Baltimore, MD, USA (see Table 4.15). In addition, each geographical feature type in the query can be expanded using a set of feature type synonyns and related words that has been manually extracted from the GNIS feature types.

On the other hand the Border Filtering Retrieval uses the shape files toolbox to retrieve geographicall relevant documents. This process uses the Shape files toolbox of the GKB to create polygons of geographical points that enclose the geographical restriction described by the geographical terms of the topic. The documents that have at least one toponym that is included in one of these polygons are selected as relevant documents.

toponym	disambiguation (full or partial)
Los Angeles	*administrative_areas@populated_places@cities*
	America@Northern_America@United_States@California@Los_Angeles
Baltimore	*administrative_areas@populated_places@cities*
	America@Northern_America@Canada@Ontario@Baltimore *America@Northern_America@United_States@Maryland@Baltimore* *America@Northern_America@United_States@Ohio@Baltimore*

Table 4.15: Example of full and partial disambiguation.

4.3.2.4 Geographical Knowledge Re-Ranking

This component re-ranks the documents retrieved by the Textual Document Retrieval with Terrier using the set of geographically relevant documents detected by the Geographical Document Retrieval module and returns a set of 1,000 documents. First, the top-scored documents retrieved by Terrier that appear in the document set retrieved by the Geographical Document Retrieval module are selected. Then, if the set of selected documents is less than 1,000, the top-scored documents retrieved by Terrier that don't appear in the document set of Geographically Relevant documents are used to complete the retrieved set (changing its ranking and score).

4.3.2.5 Geographical Border Filtering Re-Ranking

This component is an alternative to the Geographical Knowledge Re-Ranking that uses as a set of geographically relevant documents the documents retrieved by the Geographical Border Filtering Retrieval. The fusion of results is described in the previous subsection.

4.3.3 Initial Tuning with the GeoCLEF 2006 Test Collection

A set of experiments with the GeoCLEF 2006 topics was performed in order to determine the top performing options for the Terrier IR platform. The best options were a TF-IDF schema over a lemmatized collection with Porter Stemmer and Query Expansion (docs=10;terms=40) with Bose-Einstein model 1 (Bo1) scheme. The previous configuration achieved a MAP of 0.3457 in the GeoCLEF 2006. Outperforming the BM25 and the DFR (Divergence From Randomness) schemas with MAPs of 0.3394 and 0.2862.

4.3.4 Experiments at GeoCLEF 2007

For the GeoCLEF 2007 evaluation a set of five experiments (over 25 topics) was designed with the aim of applying Geographical Knowledge Re-Ranking, Relevance Feedback, and different topic tags to an automatic state-of-the-art IR system (see Table 4.16). Basically, these experiments can be divided in two groups depending on the retrieval engines used:

- **Only Terrier**. Two baseline experiments have been done in this group: the runs *TALPGeoIRTD1* and *TALPGeoIRTDN1*. These runs differ uniquely in the use of the Narrative tag in the second one. Both runs use the Terrier IR system without GKBs over a lemmatized collection and applying TFIDF with Porter Stemmer and Query Expansion (docs=10;terms=40) with the Bo1 model in order to retrieve a max of 10,000 docs per topic (but only the top-ranked 1,000 were be used in the evaluation).

- **Terrier & GeoKB Border Filtering**. The runs *TALPGeoIRTD2* and *TALPGeoIRTDN2* use the same Terrier configuration than the previous runs for textual Document Retrieval and a GKB for geographical Document Retrieval. A process of Document Filtering based on a Geographical Document Retrieval re-ranks the textually retrieved docs. The experiment *TALPGeoIRTDN3* is similar to the previous experiments but uses Border Filtering and omits Query Expansion with Relevance Feedback. The GeoKB Border Filtering approach was only applied to the topics that have a geographical relation that implies “close” or “near” and some regions.

Table 4.16: Description of the Experiments at GeoCLEF 2007.

Automatic Runs	Tags	IR System	Relevance Feedback	Border Filtering
TALPGeoIRTD1	TD	Terrier	yes	-
TALPGeoIRTD2	TD	Terrier & GeoKB	yes	-
TALPGeoIRTDN1	TDN	Terrier	yes	-
TALPGeoIRTDN2	TDN	Terrier & GeoKB	yes	-
TALPGeoIRTDN3	TDN	Terrier & GeoKB	-	yes

4.3.4.1 Results at GeoCLEF 2007 Evaluation Benchmark

The results of the TALPGeoIR system at the GeoCLEF 2007 Monolingual English task are summarized in Table 4.18. This table has the following IR measures for each run: *MAP*, *R-Precision*, and *Recall*.

The runs that use Terrier and the GeoKB have a better *MAP*, *R-Precision* than the ones that use only Terrier. The run with the best *MAP* is *TALPGeoIRTD2* with 0.2850. The best *Recall* measure is obtained by the run *TALPGeoIRTDN1* with a 93.23% of the relevant documents retrieved. This run has the same configuration of the*TALPGeoIRTD1* run but uses the Narrative tag. The run *TALPGeoIRTDN3*, that used Border Filtering without Relevance Feedback, shows an slightly improvement of MAP compared with the results of the other runs that use the Narrative tag: *TALPGeoIRTDN1* and *TALPGeoIRTDN2*.

The results show that applying Gegraphical Knowledge Re-ranking can improve the MAP and R-Prec effectiveness measures of an state-of-the-art IR system. On the other hand, the Border Filtering approach applied without Relevance Feedback improved slightly the results in MAP but an analysis per topics indicate that the approaches do not perform a general improvement and only improves some topics.

The global results of our runs are good. Four of our five runs are ranked as the first four runs in the GeoCLEF 2007 evaluation task (consult Mandl et al. (2007) for more details) both considering Mean Average Precision (ranging from 28.50% to 27.11%, next system was scored 26.42%) and R-Precision (ranging from 31.70% to 28.47%, next system was scored 27.23%). See the official results of the GeoCLEF 2007 GIR evaluation benchmark in Table 4.19 and the graphs of the 5 top-ranked gropus at Figures 4.6 and 4.7.

Topic ID	Topic Title
GC-51	Oil and gas extraction found between the UK and the Continent
GC-52	Crime near St Andrews
GC-53	Scientific research at east coast Scottish Universities
GC-54	Damage from acid rain in northern Europe
GC-55	Deaths caused by avalanches occurring in Europe, but not in the Alps
GC-56	Lakes with monsters
GC-57	Whisky making in the Scottlsh Islands
GC-58	Travel problems at major airports near to London
GC-59	Meetings of the Andean Community of Nations (CAN)
GC-60	Casualties in fights in Nagorno-Karabakh
GC-61	Airplane crashes close to Russian cities
GC-62	OSCE meetings in Eastern Europe
GC-63	Water quality along coastlines of the Mediterranean Sea
GC-64	Sport events in the french speaking part of Switzerland
GC-65	Free elections in Africa
GC-66	Economy at the Bosphorus
GC-67	F1 circuits where Ayrton Senna competed in 1994
GC-68	Rivers with floods
GC-69	Death on the Himalaya
GC-70	Tourist attractions in Northern Italy
GC-71	Social problems in greater Lisbon
GC-72	Beaches with sharks
GC-73	Events at St. Paul's Cathedral
GC-74	Ship traffic around the Portuguese islands
GC-75	Violation of human rights in Burma

Table 4.17: Topic titles of GeoCLEF 2007 in English.

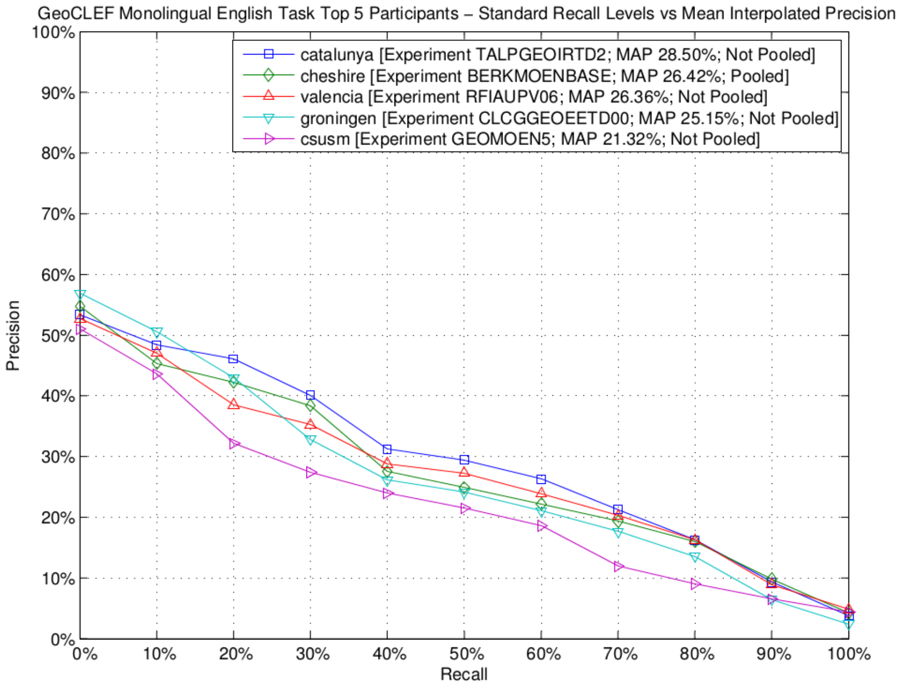


Figure 4.6: Interpolated Recall-Precision graphs of the 5 official top ranked systems at GeoCLEF 2007.

Table 4.18: TALPGeoIR results at GeoCLEF 2007.

Run	Tags	IR System	MAP	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	Terrier	0.2711	0.2847	91.23%	593/650
TALPGeoIRTD2	TD	Terrier & GeoKB	**0.2850**	**0.3170**	90.30%	587/650
TALPGeoIRTDN1	TDN	Terrier	0.2625	0.2526	**93.23%**	**606/650**
TALPGeoIRTDN2	TDN	Terrier & GeoKB	0.2754	0.2895	90.46%	588/650
TALPGeoIRTDN3	TDN	Terrier & GeoKB	0.2787	0.2890	92.61%	602/650

Approach		Best MAP
TALP-U.Politècnica Catalunya	(Ferrés and Rodríguez, 2007b)	0.2850
U.C. Berkeley	(R. R. Larson, 2007)	0.2642
U.Politècnica Valencia	(Buscaldi and Rosso, 2007)	0.2636
U. Groningen	(Andogah and Bouma, 2007)	0.2515
Cal State U.- San marcos	(Guillén, 2007)	0.2132
U.Lisbon	(Cardoso et al., 2007)	0.2180
ICL	(S. E. Overell et al., 2007)	0.1850
Moscow State Univ.	0.1761	
linguit Ltd	(Leidner, n.d.)	0.1612
U.Hildesheim	(Kölle et al., 2007)	0.1535
Microsoft Asia	(Z. Li et al., 2007a)	0.1519

Table 4.19: GIR approaches in the context of the official GeoCLEF 2007 evaluation ordered by MAP. The results of the TALPGeoIR approach are colored with light grey.

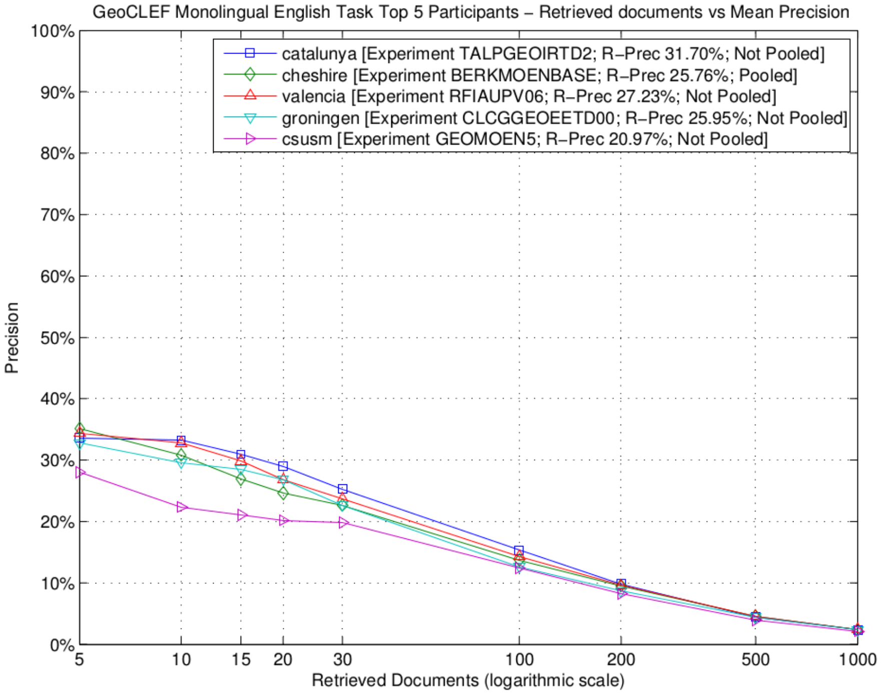


Figure 4.7: Precision at N graphs of the 5 official top ranked systems at GeoCLEF 2007.

In order to analyze the source of errors the less reliable topics have been examined, i.e. 1) all having a score clearly under the MAP for any of our runs (topics 4, 11, 16 and 17) and 2) all having a score close to the MAP for more than one run (topics 2, 9, 10, 12, 13, 14, 20, 21, 23 and 24). See the title of these topics in figure Table 4.20.

Table 4.20: Less reliable topics GeoCLEF 2007.

Num	Topic Title
2	Crime near St Andrews.
4	Damage from acid rain in northern Europe.
9	Meetings of the Andean Community of Nations (CAN).
10	Casualties in fights in Nagorno-Karabakh.
11	Airplane crashes close to Russian cities.
12	OSCE meetings in Eastern Europe.
13	Water quality along coastlines of the Mediterranean Sea.
14	Sport events in the french speaking part of Switzerland.
16	Economy at the Bosphorus.
17	F1 circuits where Ayrton Senna competed in 1994.
20	Tourist attractions in Northern Italy.
21	Social problems in greater Lisbon.
23	Events at St. Paul's Cathedral.
24	Ship traffic around the Portuguese islands.

The main sources of errors detected were:

1. Failing on properly recognizing toponyms.

   (a) Sometimes the location term has not been located in our gazetteers due to lack of coverage or different spelling, e.g. "Nagorno-Karabakh" in topic 10.

   (b) Sometimes there is a problem of segmentation. For instance, "Mediterranean Sea" (13) has been considered a multiword term by our NER and has not been located as so in our gazetteers.

   (c) Errors from the NERC classifying incorrectly toponyms as persons. For instance "Vila Franca de Xira" has been recognized as a person by our NERC system.

   (d) Our gazetteers have not recognized "St Paul's Cathedral". There is a lack of important facilities in our gazetteers.

   (e) Our NERC uses to perform correctly but failed to classify "CAN" (9) as an organization and classified it as a locative and "CAN" was found as a synonym of "Canada" in our gazetteers.

2. Failing on properly disambiguating toponyms. In some cases the toponyms have been correctly recovered from the gazetteers but the disambiguation process was wrong. This was the case of "Columbia" (narrative of 9), a typo in the text, that has been located in USA.

3. Some acronyms have not been expanded for refining queries. For instance, "OSCE" (12) has not been expanded.

4. The system did not refined the query with hyponyms. This limited in some cases the coverage. Neither “Crime” (2) nor “Economy” (16) have been refined beyond the examples included in the narrative.

5. GEO relations (as in 20) have been properly extracted but are used only in TDN3 run to apply a border filtering algorithm that has been used in 6 topics.

6. Sometimes as in (”F1 circuits”, 17) no locative has been found.

7. We have failed to attach complementary locative descriptors to the geographic term as in “Russian cities” (11) or “coastlines” (13).

The border filtering algorithm has been used in the following topics of the run TDN3: 2, 8, 16, 19, 21, and 25, applying a configuration of the Terrier IR without query expansion. Compared with the run TDN2 the MAP improves slightly in topics 8, and 25 but drops in topics 2, 16, and 19. The use of border filtering without query expansion seems not providing a general improvement neither in MAP nor in recall. On the other side, analyzing the topics that do not use border filtering without query expansion (19 topics) in run TDN3 and comparing them with the same topics in run TDN2, seems that at least in three topics avoiding query expansion has supposed a great improvement in recall and MAP (topics 1,3, and 7), only there is a sligthly drop in MAP in topics 10 and 20 and a noticeable drop in recall in topic 23.

4.3.5 Experiments with the GeoCLEF Test Collections

Several experiments with the full collection of GeoCLEF[5] (100 topics) have been designed to evaluate the relative impact of different features (alone and in combination among them) in the GIR over some state-of-the-art effectiveness measures (Ferrés and Rodríguez, 2015a). These experiments will be evaluated with the binary relevance assessments collected with pooling during the GeoCLEF forums (see Table 4.21 for details about the relevance assesments).

	2005	2006	2007	2008
#topics	25	25	25	25
#relevant_documents	1,028	378	650	747
#judged_documents	14,546	17,964	15,637	14,528

Table 4.21: Relevance assesment information about GeoCLEF evaluations

The baselines to compare are the IR algorithms TF-IDF, BM25, and InL2 with word forms in the indexed collection and the set of queries (topics). These experiments have been performed with three possible uses of the topics metadata: a) title (T), b) title and description (TD), c) title, description and narrative (TDN) . Several experiments have been performed with the full GeoCLEF collection (100 topics) to evaluate the following system components alone or in combination:

1. Linguistic Processing features evaluated in isolation or in combination: a) Lemmatization, b) Stemming, c) Lemmatization + stemming,

[5] The GeoCLEF test topics, relevance assesments and the official experiments performed at GeoCLEF from 2005 to 2008 can be downloaded at `http://direct.dei.unipd.it/`

Topic ID	Topic Title
GC-76	Riots in South American prisons
GC-77	Nobel prize winners from Northern European countries
GC-78	Sport events in the Sahara
GC-79	Invasion of Eastern Timor's capital by Indonesia
GC-80	Politicians in exile in German
GC-81	G7 summits in Mediterranean countries
GC-82	Agriculture in the Iberian Peninsula
GC-83	Demonstrations against terrorism in Northern Africa
GC-84	Bombings in Northern Ireland
GC-85	Nuclear tests in the South Pacific
GC-86	Most visited sights in the capital of France and its vicinity
GC-87	Unemployment in the OECD countries
GC-88	Portuguese immigrant communities in the world
GC-89	Trade fairs in Lower Saxony
GC-90	Environmental pollution in European waters
GC-91	Forest fires on Spanish islands
GC-92	Islamic fundamentalists in Western Europe
GC-93	Attacks in Japanese subways
GC-94	Demonstrations in German cities
GC-95	American troops in the Persian Gulf
GC-96	Economic boom in Southeast Asia
GC-97	Foreign aid in Sub-Saharan Africa
GC-98	Tibetan people in the Indian subcontinent
GC-99	Floods in European cities
GC-100	Natural disasters in the Western USA

Table 4.22: Topic titles of GeoCLEF 2008 in English.

2. Automatic Query Expansion: the Bose-Einstein (Bo1) and Kullback-Leibler QE term weighting models.

3. Geographical Knowledge Reranking (GeoKR) using a Geographical Knowledge Base (GeoKB).

4. Linguistic Processing combined with GeoKR.

5. Linguistic Processing combined with Query Expansion in the first retrieval and then applying GeoKR.

The effectiveness measures chosen to evaluate the full collection experiments have been MAP and R-Precision. Moreover, Interpolated Recall-Precision plots and Precision at N(5,10,15,20,30,100,200,500,100) plots have been used to show a more detailed evaluation of the respective improvement of the different features with respect to the baselines. All these measures have been applied over the 1,000 top-ranked retrieved documents. Significance testing has been performed using the following tests: two-tailed t-test (Sakai, 2014), and Fisher's two-sided, paired randomization test[6] (Smucker et al., 2007). Finally, a set of experiments has been done with the individual GeoCLEF collections of years 2005, 2006,

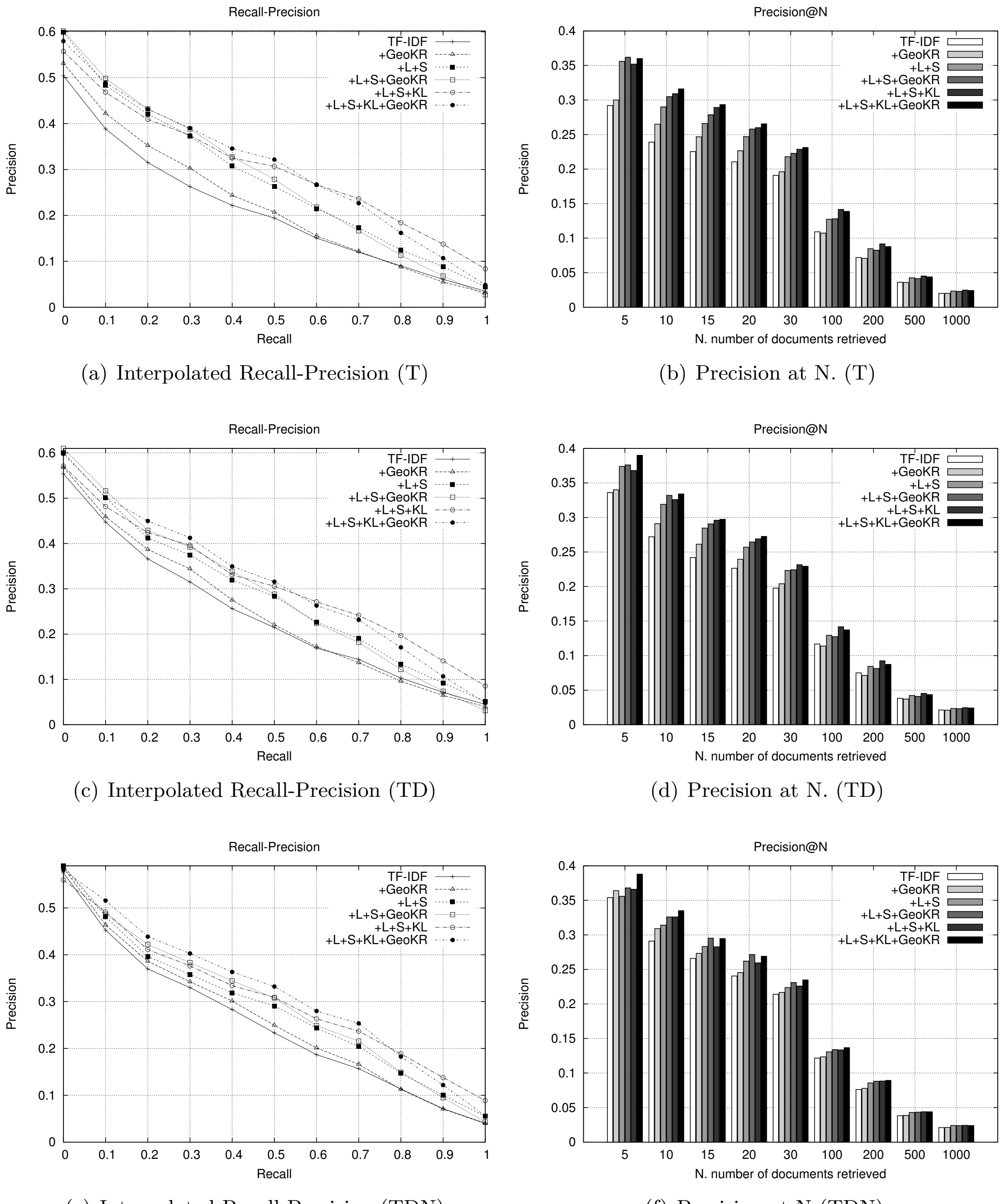
[6] http://www.mansci.uwaterloo.ca/~msmucker/software.html

2007 and 2008 to compute the performance in MAP of the best configurations with the full collection experiments. These experiments will be compared with the best official run of each GeoCLEF task.

The results of the full GeoCLEF collection experiments are shown in Table 4.23. The results of evaluating separately Geographical Knowledge Ranking, Linguistic Processing (lemmatization, stemming, and the combination of both), and Query Expansion show that all these processes improve the Mean Average Precision (MAP) and R-Precision in all the experiments and show statistical significance over the baselines in most of them. All the experiments that use only the title (T) field show statistical significance (p-value < 0.01) over baselines in MAP and R-Precision. The experiments with title and description (TD) obtained statistical significance (p-value < 0.01) in MAP (including R-Prec statistical significance with the ones that used the TF-IDF). MAP and RPrecision also show statistical significance (p-value < 0.01) in all the experiments that combine Lemmatization with stemming, GeoKB and Query Expansion. The best results in MAP (0.3116) and R-Precision (0.3142) are obtained with the InL2 algorithm with Title and Description, and using the following techniques: GeoKR, Lemmatization with Stemming, and Kullback-Leibler Query Expansion. This configuration and each method tested alone with respect the baseline show improvements in Precision at @(5,10,15,20,30,100,200,500,1000) in the majority of the experiments (see Figures 4.8, 4.9, and 4.10).

Table 4.23: Results in MAP and R-Precision with the 100 topics of all GeoCLEF collections using the Title (T), the Title and Description (TD), and the Title, Description, and Narrative (TDN) fields of the topics. Results in bold font mark the best results by field tag for each IR algorithm. Underlined results mark the best ones of each kind of field tag. Results in dark grey mark the best effectiveness measure among all field types and IR algorithms. The results marked with '*' and '**' have statistical significance for t-test and randomization tests with p-values < 0.05 and p-values <0.01 respectively.

Configuration	MAP			RPrec		
	T	TD	TDN	T	TD	TDN
TF-IDF (baseline)	0.1938	0.2238	0.2386	0.2040	0.2335	0.2444
+Stemming (S)	0.2642**	0.2740**	0.2742**	0.2678**	0.2811**	0.2707*
+Lemmatization (L)	0.2333**	0.2573**	0.2619*	0.2379**	0.2621**	0.2630
+L+S	0.2631**	0.2726**	0.2728**	0.2680**	0.2792**	0.2712 *
+Bo1	0.2372**	0.2541**	0.2692*	0.2462**	0.2647**	0.2644
+KL	0.2339**	0.2531**	0.2723**	0.2430**	0.2620**	0.2638
+GeoKB	0.2088**	0.2307**	0.2485**	0.2313*	0.2520*	0.2553**
+S+Bo1	0.2926**	**0.3007****	0.2908**	0.2942**	0.3030**	0.2779*
+L+S+Bo1	0.2869**	0.2977**	0.2959**	0.2865**	0.2997**	0.2845**
+L+S+Bo1+GeoKB	0.2899**	0.2988**	**0.3082****	**0.2957****	**0.3066****	0.3050**
+L+S+GeoKB	0.2647**	0.2735**	0.2833**	0.2700**	0.2881**	0.2877*
+S+KL	**0.2954****	0.3001**	0.2906**	0.2900**	0.3018**	0.2780*
+L+S+KL	0.2893**	0.2987**	0.2936**	0.2836**	0.2967**	0.2902**
+L+S+KL+GeoKB	0.2898**	0.2978**	0.3066**	0.2922**	0.3055**	**0.3092****
BM25 (baseline)	0.1935	0.2237	0.2390	0.2030	0.2360	0.24632
+Stemming (S)	0.2653**	0.2756**	0.2748**	0.2678**	0.2835**	0.2767**
+Lemmatization (L)	0.2353**	0.2589**	0.2624*	0.2383**	0.2626*	0.2655
+L+S	0.2643**	0.2752**	0.2744**	0.2702**	0.2800**	0.2755**
+Bo1	0.2384**	0.2635**	0.2718**	0.2405**	0.2640*	0.2650*
+KL	0.2399**	0.2676**	0.2743**	0.2403**	0.2709**	0.2630*
+GeoKB	0.2086**	0.2312**	0.2481**	0.2320	0.2534**	0.2571**
+S+Bo1	0.2898**	0.2997**	0.2908**	0.2933**	0.2962**	0.2836*
+L+S+Bo1	0.2854**	0.2951**	0.2943**	0.2850**	0.2908**	0.2880**
+L+S+Bo1+GeoKB	0.2906**	0.2983**	**0.3062****	**0.2995****	0.3037**	0.3084**
+L+S+GeoKB	0.2661**	0.2755**	0.2826**	0.2715**	0.2875**	0.2943
+S+KL	**0.2940****	0.2991**	0.2907**	0.2949**	0.2986**	0.2853
+L+S+KL	0.2899**	0.2962**	0.2916**	0.2861**	0.2930**	0.2910**
+L+S+KL+GeoKB	0.2939**	**0.3002****	0.3044**	0.2993**	**0.3084****	**0.3115** **
InL2 (baseline)	0.1939	0.2240	0.2387	0.2002	0.2348	0.2466
+Stemming (S)	0.2649**	0.2745**	0.2753**	0.2698**	0.2829**	0.2739**
+Lemmatization (L)	0.2370**	0.2612**	0.2613 *	0.2406**	0.2741**	0.2607
+L+S	0.2646**	0.2749**	0.2750**	0.2705**	0.2789**	0.2724*
+Bo1	0.2388**	0.2595**	0.2732**	0.2469**	0.2612*	0.2682
+KL	0.2384**	0.2592**	0.2764**	0.2454**	0.2658**	0.2698*
+GeoKB	0.2078**	0.2307**	0.2478**	0.2310**	0.2538*	0.2536**
+S+Bo1	0.2969**	0.3052**	0.2947**	0.2948**	0.2995**	0.2835*
+L+S+Bo1	0.2949**	**0.3067****	0.2967**	0.2933**	0.3010**	0.2884**
+L+S+Bo1+GeoKB	0.2974**	0.3052**	0.3092**	0.3029**	0.3106**	0.3060**
+L+S+GeoKB	0.2663**	0.2745**	0.2830**	0.2701**	0.2875**	0.2893
+S+KL	**0.3001****	0.3041**	0.2973**	0.2948**	0.3029**	0.2882**
+L+S+KL	0.2978**	0.3061**	0.2987**	0.2988**	0.3109**	0.2904**
+L+S+KL+GeoKB	0.2976**	0.3047**	**0.3116****	**0.3037****	**0.3142****	**0.3085****



(a) Interpolated Recall-Precision (T)

(b) Precision at N. (T)

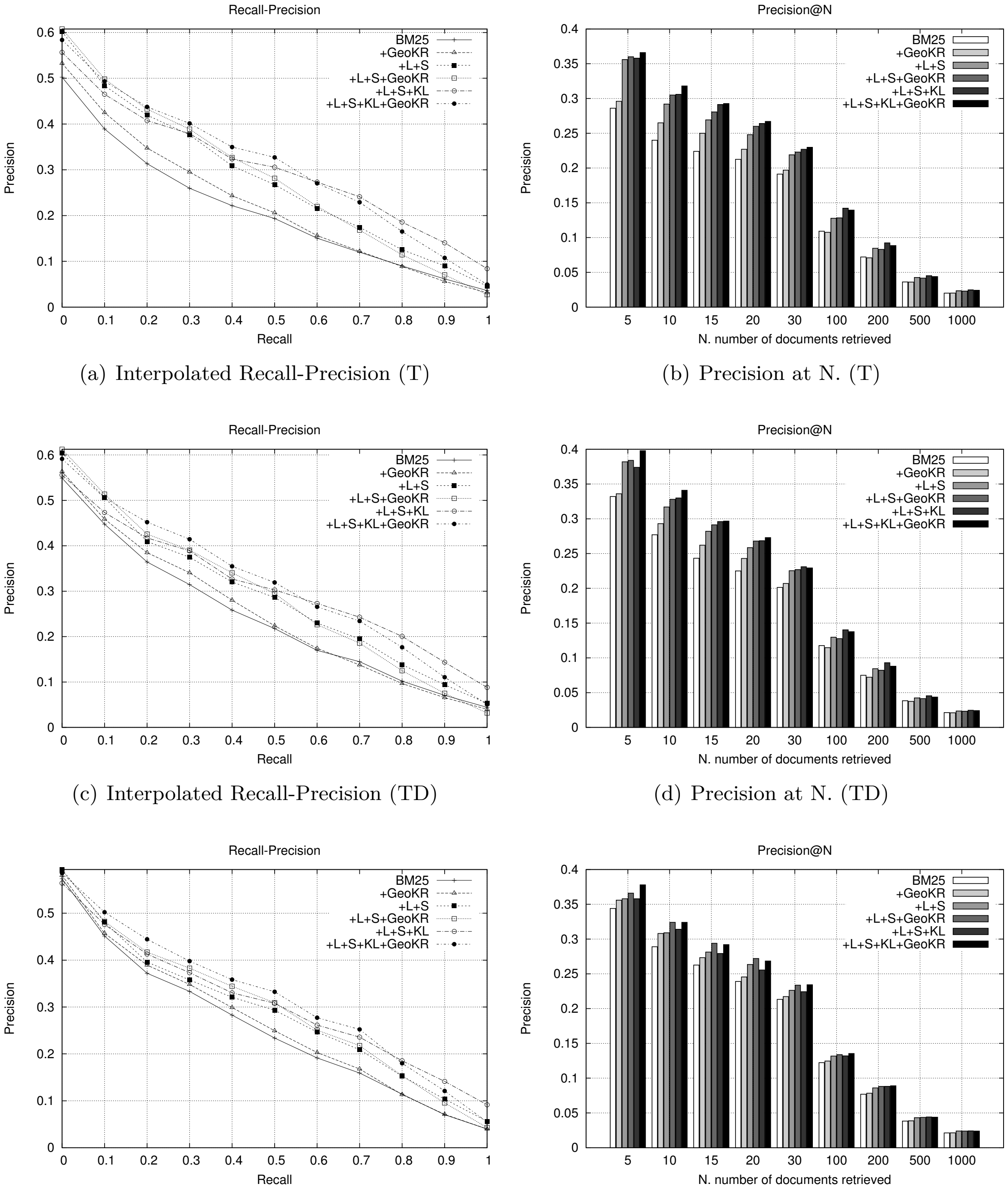
(c) Interpolated Recall-Precision (TD)

(d) Precision at N. (TD)

(e) Interpolated Recall-Precision (TDN)

(f) Precision at N (TDN).

Figure 4.8: Recall-Precision and Precision at N plots of the TF-IDF Terrier IR algorithm with different sets of features and the GeoCLEF collection (100 topics) using the Title (T), the Title and Description (TD), ant the Title, Description, and Narrative (TDN) field tags of the topics.



(a) Interpolated Recall-Precision (T)

(b) Precision at N. (T)

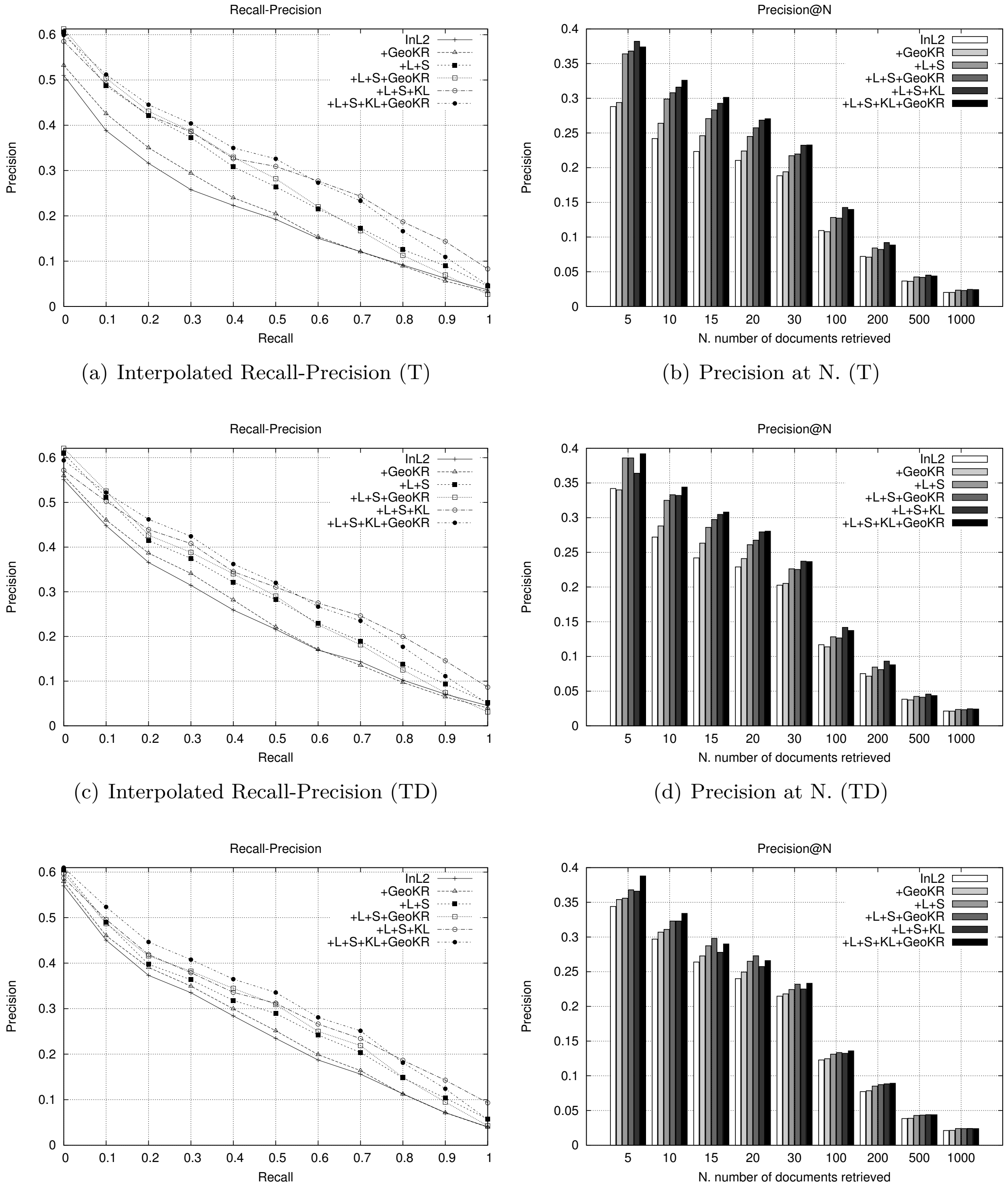
(c) Interpolated Recall-Precision (TD)

(d) Precision at N. (TD)

(e) Interpolated Recall-Precision (TDN)

(f) Precision at N (TDN).

Figure 4.9: Recall-Precision and Precision at N plots of the BM25 IR algorithm with different sets of features and the GeoCLEF collection (100 topics) using the Title (T), the Title and Description (TD), ant the Title, Description, and Narrative (TDN) field tags of the topics.



(a) Interpolated Recall-Precision (T)

(b) Precision at N. (T)

(c) Interpolated Recall-Precision (TD)

(d) Precision at N. (TD)

(e) Interpolated Recall-Precision (TDN)

(f) Precision at N (TDN).

Figure 4.10: Recall-Precision and Precision at N plots of the InL2 IR algorithm with different sets of features and the GeoCLEF collection (100 topics) using the Title (T), the Title and Description (TD), ant the Title, Description, and Narrative (TDN) field tags of the topics.

Some configurations with GeoKR, Linguistic Processing and Query Expansion have improved the MAP of the best official results at GeoCLEF evaluations of 2005, 2006, and

2007 (see Table 4.24). In the evaluation with the GeoCLEF 2008 topics, a huge drop in MAP (with respect to the use of only TD) is found when using the TDN tags. The results with the GeoCLEF 2008 dataset show MAPs with TDN of 0.2208 and 0.2178 with Bo1 and KL QE techniques respectively which are significantly lower than with T (0.22624 and 0.2616) or TD (0.2710 and 0.2697). The narrative terms of the GeoCLEF 2008 topics do not help to improve the MAP with respect the T and TD experiments while the use of TD and T is not affected. New experiments have been performed using TD for textual retrieval and TDN for GeoKR. This new configuration improved the MAP and R-Precision of the best MAP experiment in Table 4.23 from 0.3116 to 0.3198 (MAP) and from 0.3095 to 0.3236 (R-Precision).

Table 4.24: MAP at 1,000 documents with the best configurations for the full collection applied to each GeoCLEF Monolingual English task. Includes the best official results (in MAP) at GeoCLEF evaluations.

Base Configuration	MAP			
InL2+S+L+GeoKR	GeoCLEF 2005	GeoCLEF 2006	GeoCLEF2007	GeoCLEF2008
best official results	0.3936[7]	0.3034[8]	0.2850[9]	**0.3037**[10]
+Bo1(T)	0.3823	0.2573	0.2875	0.2624
+KL(T)	0.3881	0.2555	0.2853	0.2616
+Bo1(TD)	0.3863	0.2797	0.2843	0.2710
+KL(TD)	0.3898	0.2781	0.2809	0.2697
+Bo1(TDN)	0.3921	0.3303	**0.2937**	0.2208
+KL(TDN)	**0.3974**	**0.3390**	**0.2924**	0.2178

[7] (R. Larson et al., 2006)

[8] (Martins et al., 2007b)

[9] (Ferrés and Rodríguez, 2008a)

[10] (R. Wang and Neumann, 2009)

4.4 Conclusions

This chapter described three approaches to the GIR task in English and its evaluation in the context of the GeoCLEF tasks and posterior experiments with the GeoCLEF test collections in English. The first approach, GeoTALP-IR, the second one TALPGeoIR 2007, and the third one, TALPGeoIR 2007 used successfully Geographical Knowledge to improve the results of state-of-the-art IR algorithms by using a Geographical Re-Ranking process. This improvement has been shown with MAP and R-Precision effectiveness measures. These three approaches have been presented and evaluated in the official GIR GeoCLEF 2005, 2006 and 2007 evaluation benchmarks. The results of the GeoTALP-IR 2005 at GeoCLEF 2005 in comparison with the top ranked systems in MAP are low (0.2231). The three top ranked approaches achieved a MAP of 0.3936 (F. Gey et al., 2005), 0.3613 (Guillén, 2005) and 0.3495 (Ferrández et al., 2005) respectively. The TALPGeoIR 2006 approach achieved a very low MAP (0.1342) at the official GeoCLEF 2006 evaluation. This fact can be due to several reasons: 1) the *JIRS* PR was originally designed for the Question Answering task and maybe it was not used appropriately for GIR or is not suitable for GIR, 2) our system is not dealing with geographical ambiguities, 3) the lack of textual query expansion methods, 4) the need of Relevance Feedback methods, and 5) errors in the Topic Analysis phase. The second approach, TALPGeoIR 2006, evaluated officially at GeoCLEF 2006 had failed to use the Geographical Knowledge to improve the IR results. This fact was due to that this approach used the non-geographical keywords for the textual IR baselines instead of using all the keywords as in the other two approaches. The experiments with the same system after GeoCLEF 2006 have proved this fact, and showed that when using all the keywords for the textual approach the Geographical Re-Ranking process outperforms the baselines. The TALPGeoIR 2007 approach achieved the top-ranked results at the official GeoCLEF 2007 evaluation. Four of the five runs were ranked as the first four runs in the GeoCLEF 2007 evaluation task (consult Mandl et al. (2007) for more details) both considering MAP (ranging from 28.50% to 27.11%, next system was scored 26.42%) and R-Precision (ranging from 31.70% to 28.47%, next system was scored 27.23%). The reason for these competitive results at GeoCLEF 2007 are due to the use of: 1) the TF-IDF algorithm version (which uses a different TF), and 2) a combination of Geographical Knowledge Re-Ranking, Language Processing and Query Expansion with Relevance Feedback. The TALPGeoIR 2007 approach introduced a Border Filtering approach applied without Relevance Feedback that uses a shape files toolbox. This Border Filtering approach does not perform a general improvement of the results in MAP and Recall. In posterior experiments after GeoCLEF 2007 (in 2014 and 2015), the TALPGeoIR 2007 approach has been evaluated with the use of Geographical Knowledge Re-Ranking, Linguistic Processing, and Query Expansion techniques over the full GeoCLEF test collections (100 topics) (Ferrés and Rodríguez, 2015a). Evaluated separately and in combination each one of these methods has improved the MAP and R-Precision showing statistical significance with respect of the standard IR baselines (concretely TF-IDF, BM25 and InL2) in most of the experiments. The best results in MAP and R-Precision are obtained with the InL2 algorithm using the following techniques: Geographical Knowledge Re-Ranking, Lemmatization with Stemming, and Kullback-Leibler Query Expansion. Some configurations with Geographical Knowledge Re-Ranking, Linguistic Processing and Query Expansion have improved the MAP of the best official results at GeoCLEF evaluations of 2005, 2006, and 2007 with improvements of 0.9%, 11.73%, and 3.05% without statistical significance (with p-values < 0.05 detected).

CHAPTER 5

Geographical Question Answering Approaches

This Chapter describes three heterogeneous approaches to the Geographical Question Answering task in texts for Spanish and English and its evaluation in different contexts (see Table 5.1 for some details and differences between these approaches). The approaches are the following:

- *GeoTALP-QA*: a scope-based Geographical Question Answering approach. This approach has two execution modes: Knowledge-Based and Data-Driven
  - GeoTALP-QA Knowledge-Based. This approach uses Geographical Knowledge to deal with Geographical Question Answering. The GeoTALP-QA has been adapted from an existing Open-Domain Question Answering system with geographically oriented resources and specific domain adaptation. Due its language processing, answer extraction, and ontological reasoning requirements this approach is more oriented to answer questions contained in closed collections of documents that can be linguistically pre-processed.
  - *GeoTALP-QA Data-Driven.* This approach is similar to the GeoTALP-QA Knowledge-Based but uses a frequency-based approach to rank the answers. This approach is more oriented to answer questions contained in huge collections in order to exploit answer redundancy.
- *GikiTALP*: this approach uses a Data-Driven algorithm with limited Natural Language Processing and without Geographical Knowledge and it has been evaluated at GikiCLEF 2009 GeoQA evaluation benchmark.
- *GeoQuery2007 Parsing Approach*: this is an approach to the Geographical Query Parsing task that has been evaluated in the context of the official GeoCLEF 2007 GeoQuery task.

	GeoTALP-QA		GikiTALP	GeoQuery
	Knowledge-Based	Data-Driven		Parsing
Pos-Tagging	yes	yes	only queries	yes
Lemmatization	yes	yes	only queries	yes
Toponym Recognition	Yes	Yes	No	yes
Toponym Disambiguation	Partial	Partial	No	No
Question Classification	Yes	Yes	No	Yes
Answer Type Detection	Yes	Yes	No	Yes
IR system	Lucene API	Google search	Sphinx	-
Passage Retrieval	yes	snippets	no	-
Stopwords filtering	yes	yes	only queries	-
Answer Extraction	Yes	Yes	No	-
Answer Extraction Type	Reasoning	Data-Driven	-	-
Languages	en-es	en-es	en-es	en
Languages Evaluated	es	es	en-es	multilingual

Table 5.1: Description of the proposed and implemented approaches to Geographical QA.

5.1 GeoTALP-QA Geographical Question Answering Approach

This section describes GeoTALP-QA, an approach to Geographical Question Answering that uses both Knowledge-Based and Data-Driven techniques for Passage Retrieval and Answer Extraction (Ferrés and Rodríguez, 2006a; Luque et al., 2006). GeoTALP-QA is an adaptation of TALP-QA, an existing multilingual Open-Domain Question Answering (ODQA) system for factoid questions, to a Restricted Domain (RD), the geographical domain. The adaptation of the ODQA system to the geographical domain involved the modification of some components of our system such as: Question Processing, Passage Retrieval and Answer Extraction. The new system uses external resources like GNS Gazetteer for Named Entity (NE) Classification and Wikipedia or Google in order to obtain relevant documents for this domain. As pointed out in Benamara (2004), the Geographical Domain (GD) can be considered a middle way between real Restricted Domains and open ones because many open domain texts contain a high density of geographical terms. The system focuses on a Geographical Scope: given a region, or country, and a language the system can semi-automatically obtain multilingual geographical resources (e.g. gazetteers, trigger words, groups of place names, etc.) of this scope. This Restricted Domain Question Answering (RDQA) system has been built over an existing ODQA system, TALP-QA (see the architecture of this system in Figure 5.1). This system has been trained and evaluated for Spanish in the scope of the Spanish Geography.

In this section the overall architecture of GeoTALP-QA is presented and its main components are described, focusing on those components that have been adapted from an ODQA to a GDQA. Then, the Scope-Based Resources needed for the experimentation and the experiments and results obtained over a Geographical Domain corpus are presented (see the GeoTALP-QA architecture in Figure 5.2).

5.1.1 System Description

The GeoTALP-QA system architecture uses a common classical Question Answering architecture with three phases that are performed sequentially without feedback: Question Processing (QP), Passage Retrieval (PR) and Answer Extraction (AE). More details about the original TALP-QA architecture can be found in Ferrés et al. (2005c) and Ferrés et al. (2004a).

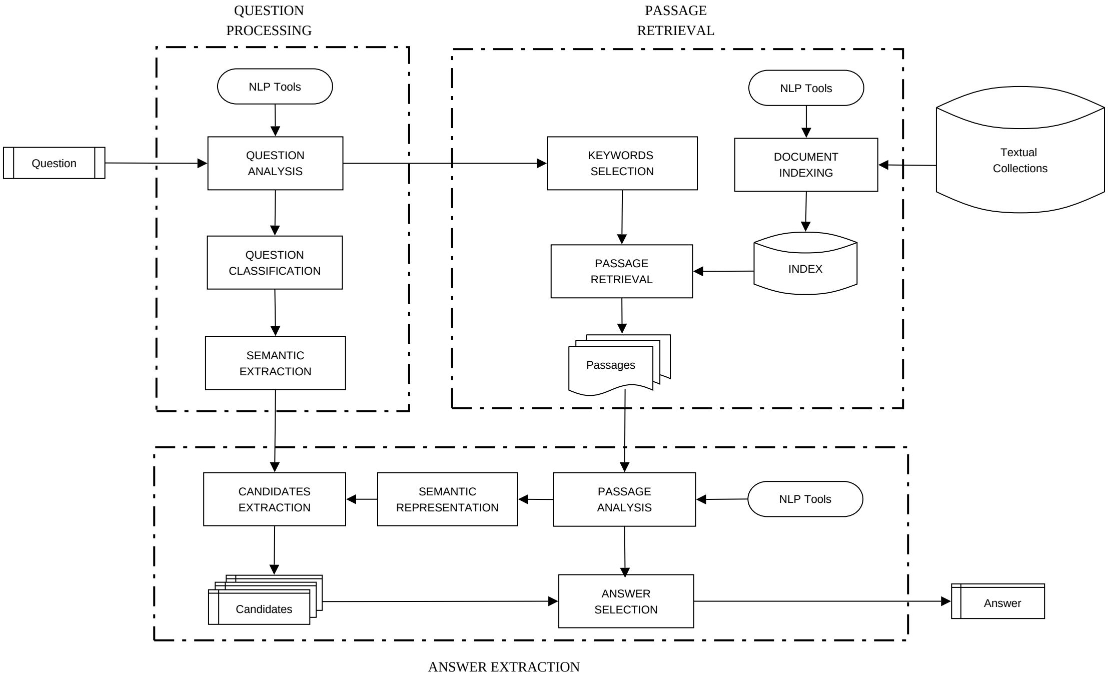


Figure 5.1: Original architecture of TALP-QA system.

Before describing these subsystems, some additional knowledge sources are described. These knowledge sources have been added to our system for dealing with the geographic domain and some language-dependent NLP tools for English and Spanish. The aim of this approach is to develop a language independent system (at least able to work with English and Spanish). Language dependent components are only included in the Question Pre-processing and Passage Pre-processing components, and can be easily substituted by components for other languages.

5.1.1.1 Additional Knowledge Sources

One of the most important task to deal with the problem of GeoQA is to detect and classify NEs with its correct geographical feature types. Geographical scope based Knowledge Bases (KB) are used to solve this problem. These KBs can be built using these resources:

- **GEOnet Names Server (GNS)**. A worldwide gazetteer, excluding the USA and Antarctica, with 5.3 million entries.
- **Geographic Names Information System** (GNIS). A gazetteer with 2.0 million entries about geographic features of the USA.

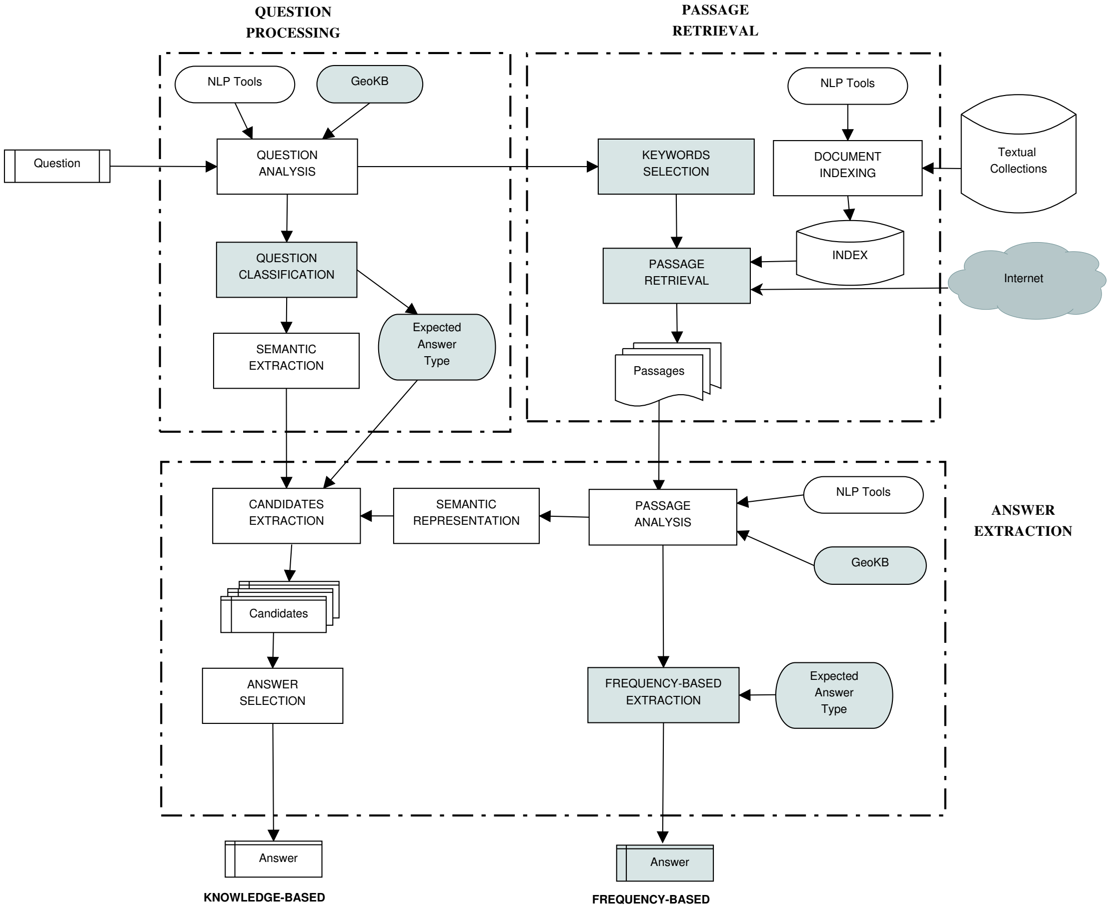


Figure 5.2: GeoTALP-QA system architecture. In grey colour the modules or data that have changed or have been added with respect to the original TALP-QA architecture.

- **Grammars for creating NE aliases**. Geographic NEs tend to occur in a great variety of forms. It is important to take this into account to avoid losing occurrences. A set of patterns for expanding have been created. (e.g. <toponym>_Mountains, <toponym>_Range, <toponym>_Chain).

- **Trigger Words Lexicon**. A lexicon containing trigger words (including multi-word terms) is used for allowing local disambiguation of ambiguous NE, both in the questions and in the retrieved passages.

Working with geographical scopes avoids many ambiguity problems, but even in a scope these problems occur:

- **Referent ambiguity problem.** This problem occurs when the same name is used for several locations (of the same or different class). In a question, sometimes it is impossible to solve this ambiguity, and, in this case, it should be accepted as correct all of the possible interpretations (or a superclass of them). Otherwise, a trigger phrase pattern can be used to resolve the ambiguity (e.g. “Madrid” is an ambiguous NE, but in the phrase, “comunidad de Madrid” (State of Madrid), ambiguity is solved). Given a scope, the system can semi-automatically obtain the most common trigger phrase patterns of the scope from the GNS gazetteer.

- **Reference ambiguity problem.** This problem occurs when the same location can have more than one name (in Spanish texts this frequently occurs as many place names occur in languages other than Spanish, as Basque, Catalan or Galician). Our approach to solve this problem is to group together all the geographical names that refer to the same location. All the occurrences of the geographical NEs in both questions and passages are substituted by the identifier of the group they belong to. The geographical knowledge available in the GNS gazetteer is used to obtain this geographical NEs groups. First, for each place name in the scope-based GNS gazetteer, all the NEs that have the same feature designation code, latitude and longitude are obtained. For each group, is selected an identifier choosing one of the NE included in it using the following heuristics: the information of the GNS field ”native” tells if a place name is native, conventional, a variant, or, is not verified. So the group representative is decided assigning the following order of priorities to the names: native, conventional name, variant name, unverified name. If there is more than one place name in the group with the same name type, then the additional length gives more priority to be cluster representative. It is necessary to establish a set of priorities among the different place names of the group because in some retrieval engines (e.g. web search engines) is not possible to do long queries.

5.1.1.2 Language-Dependent Processing Tools

A set of general purpose NLP tools are used for Spanish and English. The same tools are used for the linguistic processing of both the questions and the passages (see Ferrés et al. (2005c) and Ferrés et al. (2004a) for a more detailed description of these tools). The tools used for Spanish are:

- *FreeLing*, which performs tokenization, morphological analysis, POS tagging, lemmatization, and partial parsing.
- *ABIONET*, a NE Recognizer and Classifier (NERC) on basic categories.
- *EuroWordNet*, used to obtain a list of synsets, a list of hypernyms of each synset, and the Top Concept Ontology class.

The following tools were used to process English:

- *TnT*, a statistical POS tagger.
- *WordNet lemmatizer.*
- *ABIONET.*
- *WordNet* .
- *Spear.* A modified version of the Collins parser for dealing with questions.
- *Alembic*, a NERC with MUC classes.

5.1.1.3 Question Processing

The main goal of this subsystem is to detect the Question Type (QT), the Expected Answer Type (EAT), and the question analysis. This information is needed for the other subsystems. A language-independent formalism is used to represent this information. The processes described above are applied to the the question and passages to obtain the following information:

- Lexical and semantic information for each word: form, lemma, POS tag (Eagles or PTB tag-set), semantic class and subclass (feature type) of NE, and a list of EWN synsets.
- Syntactic information: syntactic constituent structure of the sentence and the information of dependencies and other relations between these components.

The information obtained is using in the following tasks:

- **Environment Building.** The *Environment* of a question is the set of semantic relations that hold between the different components identified in the question text. The ontology has been adapted for the Geographical Domain (see below the classes related with this domain).

```
 ENTITY
     ENTITY_PROPER_PLACE
         GEOLOGICAL_REGION
             ARCHIPELAGO
             ISLAND
             LAND_FORM
                 MOUNTAIN
             SEA_FORM
                 CAPE
                 GULF
                 SEA
             WATER_FORM
                 RIVER
         POLITICAL_REGION
             CITY
             CONTINENT
             COUNTY
             COUNTRY
             STATE
ENTITY_QUANTITY
          NUMERIC
MAGNITUDE
          AREA
          LENGTH
          FLOW
          WEIGHT
```

- **Question Classification**. The original ODQA system uses 25 QTs. For the GD only 10 Question Types are used (see Table 5.2). Only 5 QTs are common with the ODQA QTs, 5 QTs have been specially created for this domain.

  In order to determine the QT our system uses a Prolog DCG Parser. This parser uses the following features: word form, word position in the question, lemma and

Question Type	Expected Answer Type	Example
Count_objects	NUMBER	How many tributaries has the Nile River?
How_many_people	NUMBER	How many inhabitants has the state of California?
What_area	MEASURE_AREA	What is the extension of Russia?
What_flow	MEASURE_FLOW	What is the flow of the Nile?
What_height	MEASURE_HEIGHT	What is the height of the Mount Everest?
What_length	MEASURE_LENGTH	What is the length of the Amazonas River?
Where_action	LOCATION_FEATURE_TYPE	In which city does the Turia River flow through?
Where_location	LOCATION_FEATURE_TYPE	In which state is located Seattle?
Where_quality	LOCATION_FEATURE_TYPE	What is the capital of U.S.A.?
Default_class	LOCATION	-

Table 5.2: Question Types and Expected Answer Types.

part-of-speech (POS). A set of DCG rules was manually configured in order to ensure a sufficient coverage.

The parser uses external information: geographical NE feature types, trigger words for each Geographical feature type (e.g. “poblado” (*ville*)), semantically related words of each feature type (e.g. “water” related with *sea* and *river*), and introductory phrases for each Question Type (e.g. “which extension” is a phrase of the QT *What_area*).

- **Semantic Constraints Extraction.** The Semantic Constrains Set is the set of Mandatory (MC) and Optional (OC) constraints extracted from the question. MC have to be satisfied in the AE phase. OC just constrains the search for a more accurate selection. An example of the constraints extracted from an environment is shown in Table 5.3. This example shows the question type predicted, the initial predicates extracted from the question, the Environment predicates, the MCs and the OCs. MCs are *entity(4)* and *i_en_city(6)*. The first predicate refers to token number 4 (”autonomia” (*state*)) and the last predicate refers to token number 6 (”Barcelona”).

Question	1 2 3 4 5 6 7 ¿ A qué autonomía pertenece Barcelona ? (Which state Barcelona pertains to?)
Q. Type	*where_location*
Predicates	*city('Barcelona'),state(X), pertains('Barcelona',X)*
Environment	*action(5), participant_in_event(5,4), theme_jones turpin mizzano collection IRof_event(5,6),prep(4,2),entity(4), i_en_proper_place(6),det(4,3),qu(3)*
Mandatory Constraints	*entity(4),i_en_city(6)*
Optional Constraints	*action(5),theme_of_event(5,6), participant_in_event(5,4),prep(4,2), type_of_location(5,5,i_en_state), property(5,5,pertenecer,3,6)*

Table 5.3: Question Analysis example.

5.1.1.4 Passage Retrieval

Two different approaches were used for Passage Retrieval. The first one used a pre-processed corpus as a document collection. The second one used the web as document collection.

5.1.1.4.1 Knowledge-Based Off-line Corpus Retrieval This approach used a pre-processed and indexed corpus with Scope-related Geographical Information as a document collection for Passage Retrieval. The processed information was used for indexing the documents. Storing this information allowed to avoid the pre-processing step after retrieval. The Passage Retrieval algorithm used is the same used in the TALP-QA ODQA system and used also in the GeoCLEF2005 approach for GIR: a data-driven query relaxation technique with dynamic passages implemented using Lucene IR engine API (See Chapter 4.1.3 and Ferrés et al. (2005c) for more details).

5.1.1.4.2 Data-Driven Online Web Snippet Retrieval The other approach used a search-engine to get snippets with relevant information. It was expected to get a high recall with few snippets. In our experiments, Google was chosen as the search-engine using a boolean retrieval schema that takes advantage of its phrase search option and the Geographical KB to create queries that can retrieve highly relevant snippets. This approach tried to maximize the number of relevant sentences with only one query per question.

The algorithm used to build the queries is simple. First, some expansion methods described below can be applied over the keywords. Then, stop-words (including normal stop-words and some trigger words) are removed. Finally, only the nouns and verbs are extracted from the keywords list. The expansion methods used are:

- **Trigger Words Joining (TWJ).** Uses the trigger words list and the trigger phrase pattern list (automatically generated from GNS) to join trigger phrases (e.g. "isla Conejera" o "Sierra de los Pirineos").

- **Trigger Words Expansion (TWE).** This expansion is applied to the NEs that were not detected as a trigger phrase. The expansion uses its location subclass (feature type) to create a keyword with the pattern: *TRIGGER* + *NE* (e.g. "Conejera" is expanded to: ("isla Conejera" OR "Conejera")).

- **GNS Grouping Expansion (CE).** Noun Phrase expansion based on the groups generated from GNS Gazetteer.

- **Question-based Expansion (QBE).** This method appends keywords or expands the query depending on the question type. As an example, in the case of a question classified as *What_length*, trigger words and units associated to the question class like "*longitud*" (*length*) and "*kilómetros*" (*kilometers*) are appended to the query.

5.1.1.5 Answer Extraction

The system can use two different sub-systems for Answer Extraction: our ODQA system (adapted for the GD) and a frequency based system.

5.1.1.5.1 Knowledge-Based ODQA Extraction The Candidates Extraction phase is based on a relaxation process of the set of semantic constraints that is performed by means of structural or semantic relaxation rules, using the semantic ontology (see Ferrés et al. (2005c)). Then an extraction process applies a set of extraction rules on the set of sentences that have satisfied the Mandatory Constraints. In order to select the answer from the set of candidates, the following scores are computed for each candidate sentence: i) the rule score (which uses factors such as the confidence of the rule used, the relevance of the OC satisfied in the matching, and the similarity between NEs occurring in the candidate sentence and the question), ii) the passage score, iii) the semantic score , iv) the relaxation score (which takes into account the level of rule relaxation in which the candidate has been extracted). For each candidate the values of these scores are normalized and accumulated in a global score. The answer to the question is the candidate with the best global score.

5.1.1.5.2 Data-Driven Frequency-Based Extraction This extraction algorithm is quite simple. First, all snippets are pre-processed. Then, a ranked list of all the tokens satisfying the expected answer type of the question is created. The score of each token in the snippets is computed using the following formula:

$$Score(tk_i) = \sum_{o \in Occurrence(tk_i)} \frac{1}{snippet\_rank(o)}$$

Finally, the top-ranked token is extracted.

5.1.2 Resources for Scope-Based Experiments

This subsection describes how we obtained the resources needed to carry out experiments in the Spanish Geography domain using Spanish language. These resources were: the question corpus (validation and test), the document collection required by the Knowledge-Based off-line ODQA Passage Retrieval, and the geographical scope-based resources. Finally, the experiments performed are described.

5.1.2.1 Language and Scope Based Geographical Question Corpus

A corpus of Geographical questions was obtained from Albayzin, a speech corpus (Diaz et al., 1998) that contains a geographical subcorpus with utterances of questions about the geography of Spain in Spanish. A set of 6,887 question patterns were obtained from Albayzin. This corpus were analyzed and the following type of questions were extracted: Partial Direct, Partial Indirect, and Imperative Interrogative factoid questions with a simple level of difficulty (e.g. questions without nested questions). A set of 2,287 question patterns was selected. To create the question corpus a random process selected a set of 177 question patterns from the previous selection (see Table 5.4). These patterns have been randomly instantiated with Geographical NEs of the Albayzin corpus. Then, the answers were searched in the Web and the Spanish Wikipedia (SW). The results of this process were: 123 questions with answer in the SW and the Web, 33 questions without answer in the SW but with answer using the Web, and finally, 21 questions without answer (due to the fact that some questions when instantiated cannot be answered (e.g. which sea bathes the coast of Madrid?)). The 123 questions with answer in the SW were divided in two sets: 61 questions for development (setting thresholds and other parameters) and 62 for test (see this questions in Table 5.5.

¿A qué comunidad autónoma pertenece el <PICO>?

At which state pertains <PEAK>?

¿Cuál es el capital de <COMUNIDAD>?

Which is the capital of <STATE>?

¿Cuál es la comunidad en la que desemboca el <RíO>?

What is the state in which <RIVER> flows into?

¿Cuál es la extensión de <COMUNIDAD>?

Which is the extension of <STATE>?

Longitud del río <RíO>.

Length of river <RIVER>.

¿Cuántos habitantes tiene la <COMUNIDAD>?

How many people does <STATE> has?

Table 5.4: Some question patterns from Albayzin.

1 ¿A qué comunidad autónoma pertenece el Puigcampana?

To what autonomous community does the Puigcampana belongs?

2 ¿A qué comunidad pertenece El Ferrol?

To what autonomous community does El Ferrol belongs?

3 ¿A qué comunidad pertenece la isla La Gomera?

To what community belongs the island La Gomera?

4 ¿A qué mar desemboca la ría de Betanzos?

To what sea leads the ria of Betanzos?

5 ¿Cuál es el sistema de la comunidad autónoma Canaria?

What is the mountain range of tha Canary autonomous community?

6 ¿Cuál es el capital de Andalucía?

What is the capital of Andalusia?

7 ¿Cuál es el nombre de la comunidad autónoma en la que se encuentra Cullera?

What is the name of the autonomous community in which Cullera is located?

8 ¿Cuál es la capital Navarra?

What is the capital of Navarre?

9 ¿Cuál es la capital de las islas Las Canarias?

What is the capital of the Canary Islands?

10 ¿Cuál es la comunidad en la que desemboca el Guadalentín?

What is the community in which Guadalentín ends?

11 ¿Cuál es la extensión de la comunidad Madrileña?

What is the extension of the Madrid community?

12 ¿Cuál es la extensión de la comunidad autónoma donde está el golfo de Vizcaya?

What is the extension of the autonomous community in which the Bay of Biscay is located?

13 ¿Cuál es la extensión de la comunidad de Castilla y León?

What is the extension of the community of Castilla y León?

14 ¿Cuántos habitantes tiene la comunidad autónoma de Castilla?

How many inhabitants has the autonomous community of Castile?

15 ¿Cómo se llama la capital de la comunidad autónoma de La Rioja?

What is the capital of the autonomous community of La Rioja called?

16 ¿Cómo se nombra el río que pasa por Granada?

How is the river that passes through Granada named?

17 Dime a qué sistema pertenece el pico Teide?

Tell me which system belongs the peak Teide?

18 Dime a qué comunidad pertenece el cabo de La Nao?

Tell me which community is the Cape of La Nao?

19 Dime a qué comunidad pertenece la ría de Vigo?

Tell me which community is the Vigo estuary?

20 Dime el mar en que desemboca el Llobregat?

Tell me the sea where the Llobregat flows?

21 Dime el mar que baña las islas Canarias?

Tell me the sea that bathes the Canary Islands?

22 Dime en qué sistema nace el río Aragón?

Tell me in what system is the river Aragón born?

23 Dime en qué comunidad autónoma se encuentra Manacor?

Tell me in what autonomous community is Manacor?

24 Dime en qué comunidad autónoma se encuentra la ciudad de Barbastro?

Tell me in what autonomous community is the city of Barbastro?

25 Dime en qué comunidad desemboca el Llobregat?

Tell me in what community does the Llobregat ends?

26 Dime en qué mar está la isla de Conejera?

Tell me in what sea is the island of Conejera?

27 Dime la población de la comunidad autónoma de Murcia?

Tell me the population of the autonomous community of Murcia?

28 Dime qué extensión tiene la isla de Hierro?

Tell me the extent of the island of Hierro?

29 ¿Dónde está la isla de Gran Canaria?

Where is the island of Gran Canaria?

30 ¿Dónde está la ría Ribadeo?

Where is the Ribadeo estuary?

31 ¿En qué archipiélago se encuentra Mallorca?

In what archipelago is Mallorca located?

32 ¿En qué ciudad desemboca el río Segura?

In what city does the Segura River ends?

33 ¿En qué comunidad autónoma está el Cantábrico?

In what autonomous community is the Cantabrian Sea located?

34 ¿En qué comunidad autónoma está el Mulhacén?

In what autonomous community is the Mulhacen located?

35 ¿En qué comunidad autónoma está el cabo Tarifa?

In what autonomous community is the Tarifa Cape located?

36 ¿En qué comunidad autónoma está situada la Sierra de Gũdar?

In what autonomous community is the Sierra of Gúdar located?

37 ¿En qué comunidad autónoma están los Picos de Europa?

In what autonomous community are the Picos the Europa located?

38 ¿En qué comunidad autónoma se encuentra la isla de La Gomera?

In what autonomous community is the island of La Gomera located?

39 ¿En qué comunidad autónoma se encuentra la Sierra del Maestrazgo?

In what autonomous community is the Sierra of Maestrazgo located?

40 ¿En qué comunidad está la sierra de Somosierra?

In what autonomous community is the sierra of Somosierra located?

41 ¿En qué comunidad nace el río Guadarrama?

In what autonomous community is the Guadarrama river located?

42 ¿En qué comunidad se encuentra el cabo San Adrián?

In what autonomous community is the San Adrián Cape located?

43 ¿En qué comunidad se encuentran los Pirineos?

In what autonomous community are the Pyrenees located?

44 ¿En qué mar está situado el golfo de Cádiz?

In what sea is the Gulf of Cádiz the located?

45 ¿En qué mar se encuentra la ría de Camariñas?

In what sea is the Camariñas estuary?

46 La comunidad en la que nace el río Guadalbullón?

The community in which the river Guadalbullón is born?

47 Me gustaría saber la extensión de la comunidad Vasca?

I would like to know the extension of the Basque community?

48 Nombre de la capital de Andalucía?

Name of the capital of Andalusia?

49 Nombre de la capital de la comunidad autónoma de Andalucía?

Name of the capital of the Autonomous Community of Andalusia?

50 Nombre de la comunidad donde nace el río Eresma?

Name of the community where the river Eresma is born?

51 Podría decirme el nũmero de habitantes de Figueras?

Can you tell me the number of inhabitants of Figueras?

52 Quiero que me digas la capital de la comunidad autónoma de Canarias?

I want you to tell me the capital of the autonomous community of the Canary Islands?

53 Quisiera saber el mar en donde está situada La Gomera?

I would like to know the sea where La Gomera is located?

54 ¿Qué capital tiene Castilla?

What capital does Castilla have?

55 ¿Qué extensión tiene La Gomera?

What is La Gomera extension?

56 ¿Qué extensión tiene la comunidad autónoma Asturiana?
What is the extension of the Asturian Autonomous Community?
57 ¿Qué mar baña el golfo de Onteniente?
What sea bathes the Gulf of Onteniente?
58 ¿Qué mar baña la comunidad autónoma Murciana?
What sea bathes the Murcian autonomous community?
59 ¿Qué mar es el que baña a la comunidad de Murcia?
What sea bathes the Murcian community?
60 ¿Qué nũmero de habitantes tiene Castilla la Mancha? *What is the number of inhabitants of Castilla la Mancha?*
61 ¿Qué nũmero de habitantes tiene Astorga?
What is the number of inhabitants of Astorga?
62 ¿Qué río pasa por Salamanca?
Which river passes through Salamanca?

Table 5.5: Test set of 62 instantiated questions patterns from Albayzin (in Spanish).

5.1.2.2 Document Collection for the Knowledge-Based ODQA Passage Retrieval

In order to test our ODQA Passage Retrieval system we need a document collection with enough geographical information to solve the questions of Albayzin corpus. We used the filtered Spanish Wikipedia[1]. First, we obtained the original set of documents (26,235 files). Then, we selected two sets of 120 documents about the Spanish geography domain and the non-Spanish geography domain. Using these sets we obtained a set of Topic Signatures (TS) (C.-Y. Lin and E. Hovy, 2000) for the Spanish geography domain and another set of TS for the non-Spanish geography domain. Then, we used these TS to filter the documents from Wikipedia, and we obtained a set of 8,851 documents belonging to the Spanish geography domain. These documents were pre-processed and indexed.

5.1.2.3 Geographical Scope-Based Resources

A Knowledge Base (KB) of Spanish Geography has been built using four resources:

- GNS: A set of 32,222 non-ambiguous place names of Spain.
- Albayzin Gazetteer: a set of 758 places.
- A Grammar for creating NE aliases. We created patterns for the summit and state classes (the ones with more variety of forms), and we expanded this patterns using the entries of Albayzin.
- A lexicon of 462 trigger words.

A set of 7,632 groups of place names were obtained using the grouping process over GNS. These groups contain a total of 17,617 place names, with an average of 2.51 place names per group. See in Figure 5.3 an example of a group where the canonical term appears underlined.

[1] **Spanish Wikipedia**. http://es.wikipedia.org

{*Cordillera Pirenaica*, *Pireneus*, *Pirineos*, *Pyrenaei Montes*, *Pyrénées*, *Pyrene*, *Pyrenees*}

Figure 5.3: Example of a group obtained from GNS.

In addition, a set of the most common trigger phrases in the domain has been obtained from the GNS gazetteer (see Table 5.6).

	Geographical Scope	
	Spain	UK
Top-ranked Trigger Phrases	*TRIGGER* de *NE*	*NE TRIGGER*
	TRIGGER NE	*TRIGGER NE*
	TRIGGER del *NE*	*TRIGGER* of *NE*
	TRIGGER de la *NE*	*TRIGGER* a' *NE*
	TRIGGER de las *NE*	*TRIGGER* na *NE*

Table 5.6: Sample of the top-ranked trigger phrases automatically obtained from GNS gazetteer for the geography of Spain and UK.

5.1.3 Experiments

The experiments to evaluate the accuracy of different subsystems and operational modes were performed over the 62 instantiated test questions patterns from albayzin corpus. The accuracy measures of the GDQA system and its subsystems (Question Processing , Passage Retrieval, and Answer Extraction) with its two execution modes (Knowledge-Based and Data-Driven) were calculated. For the Data-Driven operational mode of the system, the Passage Retrieval phase and the global system accuracy were evaluated over the web-based snippet retrieval with queries to Google API with some variants of query expansions. For the Knowledge-Based operational mode of the system, the Passage Retrieval phase was evaluated over the dataset extracted from the filtered Spanish Wikipedia; and the Answer Extraction phase and the global accuracy were evaluated over tw the web-based snippet retrieval an the set of filtered geographically relevant documents from Spanish Wikipedia.

5.1.4 Results

This section evaluates the behavior of our GDQA system over a test corpus of 62 questions and reports the errors detected on the best run. The whole system and its three main components are evaluated.

- **Question Processing**. The Question Classification task has been manually evaluated. This subsystem has an accuracy of 96.77%.
- **Passage Retrieval**. The results of the evaluation of this subsystem for both the two kinds of Retrieval: ODQA+Wiki and google snippets with query expansions.

  are shown in Table 5.7. The *answer accuracy at N passages/snippets* measure computes the ratio of questions that have a correct answer in its set of passages or snippets.

Retrieval Mode	Accuracy at N passages/snippets			
	N=10	N=20	N=50	N=100
Google	0.6612	0.6935	0.7903	0.8225
+TWJ	0.6612	0.6774	0.7419	0.7580
+TWJ+TWE	0.6612	0.6774	0.7419	0.7580
+CE	0.6612	0.6774	0.7741	0.8064
+QBE	0.8064	0.8387	0.9032	0.9354
+TWJ+QB+CE	0.7903	0.8064	0.8548	0.8870
Google+All	0.7903	0.8064	0.8548	0.8870
ODQA+Wiki	0.4354	0.4516	0.4677	0.5000

Table 5.7: Passage Retrieval results.

- **Answer Extraction**. The evaluation of the ODQA Answer Extractor subsystem is shown in Table 5.8. The accuracy was evaluated taking into account the number of correct and supported answers by the passages divided by the total number of questions that have a supported answer in its set of passages. This evaluation was done using the results of the top-ranked retrieval configuration over the development set: the *Google+TWJ+QB+CE* configuration of the snippet retriever.

Accuracy at N Snippets		
N=10	N=20	N=50
0.2439 (10/41)	0.3255 (14/43)	0.3333 (16/48)

Table 5.8: Results of the ODQA Answer Extraction subsystem (accuracy).

In Table 5.9 are shown the global results in accuracy of the two QA Answer Extractors used (ODQA and Frequency-Based). The passages retrieved by the *Google+TWJ+QB+CE* configuration of the snippet retriever were used.

	Accuracy	
Num. Snippets	ODQA	Freq-based
10	0.1774 (11/62)	0.5645 (35/62)
20	0.2580 (16/62)	0.5967 (37/62)
50	0.3387 (21/62)	0.6290 (39/62)

Table 5.9: QA results over the test set.

The analysis of the 23 questions that fail in the best run detected that 10 questions had no answer in its set of passages. In 5 of these questions it is due to have a non common question or location. The other 5 questions have problems with ambiguous trigger words (e.g. *capital*) that confuse the web-search engine. On the other hand, 13 questions had the answer in its set of passages, but were incorrectly answered. The reasons are mainly due to the lack of passages with the answer (8), answer validation and spatial-reasoning (3), toponym normalization error (1), and the need of more context in the snippets (1).

Out of 62 questions, our system provided the correct answer to 39 questions in the experiment with the best results.

The Passage Retrieval for ODQA offers less attractive results when using the SW corpus. The problem of using a corpus of the Spanish Wikipedia to extract the answers is that it gives few documents with the correct answer, and, it is difficult to extract the answer because the documents contain tables, lists, ill-formed sentences, etc. The ODQA AE needs a grammatically well-structured text to extract correctly the answers. The QA system offers a low performance (33% of accuracy) when using this AE over the web-based retrieved passages. In some cases, the snippets are cut and better performance will be expected by retrieving the whole documents from Google.

On the other hand, web-based snippet retrieval, with only one query per question, gives good results in Passage Retrieval. The QA system with the Frequency-Based Answer Extractor obtained better results than the system witht the ODQA Answer Extractor (62.9% versus 33.87% of accuracy using a set of 50 snippets).

5.2 Geographical QA Approach over the Wikipedia

In this section the overall architecture of the gikiTALP Geographical Question Answering system is presented (Ferrés and Rodríguez, 2010a). The experiments, results, and initial conclusions in the context of the GikiCLEF 2009 Monolingual English and Spanish task (D. Santos and L. M. Cabral, 2009) (D. Santos and L. Cabral, 2010) are described.

GikiCLEF 2009 is an evaluation task under the scope of CLEF. Its aim is to evaluate systems which find Wikipedia entries/documents that answer a particular information need, which requires geographical reasoning of some sort. GikiCLEF is the successor of the GikiP 2008 (D. Santos et al., 2008) pilot task which ran in 2008 under GeoCLEF.

For GikiCLEF, systems will need to answer or address geographically challenging topics, on the Wikipedia collections, returning Wikipedia document titles as list of answers in all languages it can find answers (see an example of two GikiCLEF topics in Table 5.4).

```
<topic id="GC-2009-07">
What capitals of Dutch provinces received their town privileges before the fourteenth century?
<topic id="GC-2009-08">
Which authors were born in and write about the Bohemian Forest?
```

Figure 5.4: Example of two topics of the GikiCLEF 2009 evaluation.

The Wikipedia collections for all GikiCLEF languages are available in three formats, HTML dump, SQL dump, and XML version. We used the SQL dump version of the English and Spanish collections (see details about these collections summarized in Table 5.10).

Table 5.10: Description of the Wikipedia collections used by the GikiTALP approach at GikiCLEF 2009.

Language	#Total	#Pages	#Templates	#Categories	#Images
en	6,587,912	5,255,077	154,788	365,210	812,837
es	714,294	641,852	11,885	60,556	1

5.2.1 System Description

The system architecture has three phases that are performed sequentially: Collection Indexing, Topic Analysis, and Information Retrieval. The textual Collection Indexing has been applied over the textual collections with MySQL and the open-source full-text engine Sphinx using the Wikipedia SQL dumps.

Sphinx[2] is a full-text search engine that provides fast, size-efficient and relevant full-text search functions to other applications. The indexes created with Sphinx do not have any language processing. Sphinx has two types of weighting functions: Phrase rank and Statistical rank. Phrase rank is based on a length of longest common subsequence (LCS) of search words between document body and query phrase. Statistical rank is based on classic BM25 function which only takes word frequencies into account. Two types of search

[2] http://www.sphinxsearch.com/

modes in Sphinx were used (see [3] for more information about the search mode and weighting schemes used):

- MATCH ALL: the final weight is a sum of weighted phrase ranks.
- MATCH EXTENDED: the final weight is a sum of weighted phrase ranks and BM25 weight, multiplied by 1000 and rounded to integer.

The Topic Analysis phase extracts some relevant keywords (with its analysis) from the topics. These keywords are then used by the Document Retrieval phases. This process extracts lexico-semantic information using the following set of Natural Language Processing tools: *TnT* POS tagger (Brants, 2000), *WordNet lemmatizer* (version 2.0) for English, and *Freeling* (Atserias et al., 2006) for Spanish. Some of these NLP tools were used also for the GIR task in GeoCLEF 2007 (Ferrés and Rodríguez, 2007b) as described in Chapter 4. The language processing with these NLP tools is applied only in the queries. The Wikipedia collection is indexed without applying the stemming and stopword filtering options of Sphinx. The retrieval is done with Sphinx and then the final results are filtered. The Wikipedia entries without Categories are discarded.

5.2.2 Experiments at GikiCLEF 2009

For the GikiCLEF 2009 evaluation a set of three experiments were designed (Ferrés and Rodríguez, 2010a). These experiments consisted in applying different baseline configurations (see Table 5.11) to retrieve Wikipedia entries (answers) of 50 geographically challenging topics.

The three baseline runs were designed changing two parameters of the system: the IR Sphinx search mode and the Natural Language Processing techniques applied over the query. The first run (gikiTALP1) do not uses any NLP processing technique over the query and the Sphinx match mode used is MATCH_ALL. The second run (gikiTALP2) uses stopwords filtering and the lemmas of the remaining words as a query and the Sphinx match mode used is MATCH_ALL. The third run (gikiTALP3) uses stopwords filtering and the lemmas of the remaining words as a query and the Sphinx match mode used is MATCH_EXTENDED.

Table 5.11: Description of the experiments at GikiCLEF 2009.

Automatic Runs	NLP in Query	Sphinx Match
gikiTALP1	-	MATCH_ALL (phrase rank)
gikiTALP2	lemma + stopwords filtering	MATCH_ALL (phrase rank)
gikiTALP3	lemma + stopwords filtering	MATCH_EXTENDED (BM25)

[3] http://www.sphinxsearch.com/docs/current.html. Sphinx 0.9.9 documentation.

5.2.3 Results at GikiCLEF 2009

The results of the gikiTALP system at the GikiCLEF 2009 Monolingual English and Spanish task are summarized in Table 5.12. This table has the following IR measures for each run: number of correct answers (*#Correct Answers*), *Precision*, and *Score*. The run gikiTALP1 obtained the following scores for English, Spanish and Global: 0.6684, 0.0280, and 0.6964. Due to an unexpected error we did not produced answers for the Spanish topics in run 2 (gikiTALP2), then the results for English and global were 1,3559. The results of the scores of the run gikiTALP3 for English, Spanish and Global were 1.635, 0.2667, and 1.9018 respectively.

Table 5.12: TALP GikiTALP Results

run	**Measures**	**English (EN)**	**Spanish (ES)**	**Total**
run 1	#Answers	383	143	526
	#Correct answers	16	2	18
	Precision	0.0418	0.0140	0.0342
	Score	0.6684	0.0280	0.6964
run 2	#Answers	295	–	295
	#Correct answers	20	–	20
	Precision	0.0678	–	0.0678
	Score	1.3559	–	1.3559
run 3	#Answers	296	60	356
	#Correct answers	22	4	26
	Precision	0.0743	0.0667	0.0730
	Score	1.6351	0.2667	1.9018

The best results were obtained with the NLP techniques (lemmas in the queries and stopwords filtered) and the Sphinx mode MATCH_EXTENDED without Geographical Knowledge as baseline algorithms. In comparison with other approaches at GikiCLEF this approach was not so good.

5.3 Geographical Query Parsing Approach

This section describes the overall architecture of our Geographical Query Parsing system and its main components (Ferrés and Rodríguez, 2008b). This system uses some modules of a Geographical Information Retrieval system presented at GeoCLEF 2006 (Ferrés and Rodríguez, 2007a) and modified for GeoCLEF 2007 GeoQuery task (Ferrés and Rodríguez, 2007b). The experiments, results, analysis of the results and conclusions in the context of the GeoCLEF's 2007 GeoQuery pilot task are also presented.

The Query Parsing task (GeoQuery) was a pilot task proposed in GeoCLEF 2007 (Z. Li et al., 2007c). This task was dedicated to identifying geographic queries within a log file from the MSN search engine (see in Figure 5.5 some records of the GeoCLEF2007 queries log example). This task was organized by Microsoft Research Asia (Mandl et al., 2007).

```
Discount Airline Tickets To Brazil.
doctors hospital augusta ga.
minibus trips in cyprus.
niagara day tours from toronto.
plumbers in manhattan ny new york.
```

Figure 5.5: Example of some queries extracted from the GeoQuery2007 .

The GeoQuery task consisted on five subtasks:

- Detect whether the query is geographic or no.
- Extract the WHERE component of the query.
- Extract the GEO-RELATION (from a set of predefined types) if present.
- Extract the WHAT component of the query and classify it as MAP, YELLOW PAGE or INFORMATION types.
- extract the coordinates (LAT-LONG) of the WHERE component. This process involves sometimes a disambiguation task.

As an example, see in Table 5.13 the information that has to be extracted from the query "Discount Airline Tickets to Brazil".

Field	Content
LOCAL	YES
WHAT	Discount Airline Tickets
WHAT-TYPE	INFORMATION
WHERE	Brazil
GEO-RELATION	TO
LAT-LONG	-10.0_-55.0

Table 5.13: Example of information extracted from a query.

5.3.1 System Description

The system architecture has two main phases that are performed sequentially: Topic Analysis and Question Classification.

5.3.1.1 Topic Analysis

The Topic Analysis phase has two main components: a Linguistic Analysis and a Geographical Analysis.

5.3.1.2 Linguistic Analysis

This process extracts lexico-semantic and syntactic information using the following set of Natural Language Processing tools (explained previously in Chapter 4): i) **TnT** an statistical POS tagger (Brants, 2000), ii) **WordNet lemmatizer** (version 2.0), iii) **A Maximum Entropy based NERC** trained with the CONLL-2003 shared task English data set, iv) **Spear** a modified version of the Collins parser (Collins, 1999).

A dataset of 800.000 queries in English from a web search-engine was pre-processed with linguistic tools to obtain the following data structures:

- **Sent**, which provides lexical information for each word: form, lemma, POS tag (Penn-Tree-Bank (PTB) tag-set for English), semantic class of NE, list of EWN synsets and, finally, whenever possible the verbs associated with the actor and the relations between some locations (specially countries) and their gentiles (e.g. nationality).

- **Sint**, composed of two lists, one recording the syntactic constituent structure of the question (basically nominal, prepositional and verbal phrases) and the other collecting the information of dependencies and other relations between these components.

- **Environment.** The environment represents the semantic relations that hold between the different components identified in the question text. These relations are organized into an ontology of about 100 semantic classes and 25 relations (mostly binary) between them. Both classes and relations are related by taxonomic links. The ontology tries to reflect what is needed for an appropriate representation of the semantic environment of the question (and the expected answer). The environment of the question is obtained from *Sint* and *Sent*. A set of about 150 rules was built to perform this task. Refer to Ferrés et al. (2004a) for details.

5.3.1.3 Geographical Analysis

The Geographical Analysis is applied to the Named Entities from the queries that have been classified as LOCATION or ORGANIZATION by the NERC module. A Geographical Thesaurus is used to extract geographical information about these Name Entities. This component has been built joining four gazetteers that contain entries with places and their geographical class, coordinates, and other information (this thesaurus has been described in Chapter 4). This thesaurus contains subsets of data from: GNS, GNIS, GeoWorldMap and World Gazetteer.

A subset of the most important features from this thesaurus has been manually set using 46.132 places (including all kind of geographical features: countries, cities, rivers, states,...).

This subset of important features has been used to decide if the query is geographical or not geographical.

5.3.1.4 Question Classification

The query classification task is performed through the following steps:

- The query is linguistically preprocessed (as described in the previous subsection) for getting its lexical, syntactic and semantic content. See in Table 5.14 the results of the process for the former example. What is relevant in the example is the fine grained classification of 'Brazil' as country, the existence of taxonomic information, both of location type (administrative_areas@@political_areas@@countries) and location content (America@@South_America@@ Brazil), and coordinates (-10.0_-55.0, useful for disambiguating the location and for restricting the search area) and the existence of a shallow syntactic tree consisting on simple tokens and chunks, in this case built by the composition of two chunks, a nominal chunk ('Discount Airline Tickets') and a prepositional one ('to Brazil').

Query: “Discount Airline Tickets to Brazil”
Semantic: [entity(3),mod(3,1),quality(1),mod(3,2),entity(2),i_en_proper_country(5)]
Linguistic: Brazil Brazil NNP LOCATION
Geographical: America@@South_America@@Brazil@@-10.0_-55.0
Feature type: administrative_areas@@political_areas@@countries

Table 5.14: Semantic and Geographical Content of GQ-38.

- Over the sint structure, a DCG like grammar consisting of about 30 rules developed manually from the sample of GeoQuery and the set of queries of GeoCLEF 2006, is applied for obtaining the list of topics (each topic represented by its initial and final positions) represented by a triple <geo-relation, initial position, final position>). A set of features (consultive operations over chunks or tokens and predicates on the corresponding sent structures) is used by the grammar. The following features were available:
  - **chunk features**: category, inferior, superior, descendents.
  - **token features**: num, POS, word form, lemma, NE 1 (general), NE 2 specific.
  - **token semantics**: synsets, concrete and generic Named Entity type predicates (Named Entity types include: location, person, organization, date, entity, property, magnitude, unit, cardinal point, and geographical relation.
  - **head of the chunk features**: num, POS, word, lemma, first NE, second NE.
  - **head of the chunk semantic features**.
  - **left corner of the chunk**: num, POS, word form, lemma, NE 1 (general), NE 2 (specific)
  - **left corner of the chunk semantics**: WordNet synsets.

- Finally from the result of step 2 several rule-sets are in charge of extracting: i) LOCAL, ii) WHAT and WHAT-TYPE, iii) WHERE and GEO-RELATION, and iv) LAT-LONG data. So, there are four rule sets with a total of 25 rules.

5.3.2 Experiments and Results at GeoQuery 2007

Only one experiment for the GeoQuery 2007 data set was performed for this evaluation benchmark. The experiment consisted in to extracting the requested data for the GeoQuery from a set of 800.000 queries.

The results of the TALP system and the other approaches at the GeoCLEF's 2007 GeoQuery Geographical parsing task for English are summarized in Table 5.15. This table has the following IR measures for each run: *Precision*, *Recall*, and *F1*.

In the evaluation data set, a set of 500 queries had been labeled which are chosen to represent the whole query set (800.000). The submitted results have been manually evaluated using a strict criterion where a correct results should have all <local>, <what>, <what-type> and <where> fields correct (the <lat-long> field was ignored in the evaluation).

Our run achieved the following results: 0.2222 of Precision, 0.249 of Recall, and 0.235 of F1.

Table 5.15: Official Results at GeoQuery 2007.

Team Name	Precision	Recall	F1
Ask (Z. Li et al., 2007c)	**0.625**	0.258	0.365
CSUSM (Guillén, 2008a)	0.201	0.197	0.199
Linguit (Z. Li et al., 2007c)	0.112	0.038	0.057
Miracle (Lana-Serrano et al., 2008)	0.428	**0.566**	**0.488**
TALP (Ferrés and Rodríguez, 2008b)	0.222	0.249	0.235
XLDB (Z. Li et al., 2007c)	0.096	0.08	0.088

The global results of our run for the local query were 0.222 Precision, 0.249 Recall, and 0.235 of F1. Our system was ranked the third from 6 participants, being the groups of Miracle and ASK the best ones in terms of F1-Scores, 0.488 and 0.365 respectively.

In order to analyse the source of errors the evaluation criteria described in Z. Li et al. (2007b) was used. The confusion matrices for LOCAL and WHAT-TYPE for the 500 queries evaluated are presented in tables 1 and 2. The number of errors have been 99 for LOCAL, 126 for WHAT, 245 for WHAT-TYPE, 41 for GEO-RELATION, 122 for WHERE, giving a total of 315 queries with one or more errors and 185 correctly answered.

Table 5.16: Confusion matrix for LOCAL

	NO	YES
NO	122	31
YES	68	279

Here are described the most problematic figures:

1. Queries not recognized as LOCAL (31) by our system. Clearly this case corresponds to the different coverage of our gazetteers and those used by the evaluators. Some frequent errors can be classified as follows:

Table 5.17: Confusion matrix for WHAT-TYPE

	Map	Information	Yellow Pages
Map	4	4	10
Information	44	39	121
Yellow Pages	38	28	212

- Some errors simply correspond to lack of coverage of the gazetteer subset used for this task (as “cape may”).
- Some of these errors could be recovered using the context (as “Gila County”).
- Sometimes the query has been considered as LOCAL because it corresponds to an address (street, place and so). have not considered these kinds of locations (as “caribbean joe”).
- Some cases correspond to misspellings of Spanish words (as “Cercanías” considered erroneously as a toponym is Spain).

2. Queries that the approach has improperly considered as LOCAL (68): i) in some cases (30%) it seems that the gazetteers have a higher coverage. ii) Other queries correspond to Named Entities probably not present in the gazetteers but erroneously classified as location by our NERC (as “Hitachi” or “Sala”).

3. From Table 5.17 the following problematic cases arise: i) confusion between “Map” and “Information” or “Yellow Pages”. Most of the errors correspond to a lack of a rule that assigns “Map” to the queries consisting on only a locative (as “Coronado, San Diego”). Sometimes it is due to an only partial recognizing of the locative. ii) the confusion between “Information” and “Yellow Pages” is problematic as Z. Li et al. (2007b) point out. There is no clear trends on the typification of the errors. Besides, we have used “Yellow Pages” as our default class when no classification rule can be applied. Obviously a more precise classification is needed.

5.4 Conclusions

This Chapter describes three heterogeneous approaches for Geographical Question Answering in English and Spanish: 1) GeoTALP-QA system with the Knowledge-Based and Data-Driven operational modes, 2) GikiTALP, and 3) GeoQuery2007 Query Parsing.

GeoTALP-QA is an adaptation of an existing multilingual Open-Domain Question Answering (ODQA) system for factoid questions to a Restricted Domain, the Geographical Domain, that has two kind of operational modes for Answer Extraction (Knowledge-Based and Data-Driven), and two kind of sources of passages: indexed documents from a preprocessed corpora or a query-expansion based Passage-Retrieval that uses the Google search API. The system focuses on a Geographical Scope: given a region, or country, and a language the system can semi-automatically obtain multilingual geographical resources (e.g. gazetteers, trigger words, groups of place names, etc.) of this scope. The system has been trained and evaluated for Spanish in the scope of the Spanish Geography. A set of 62 questions about Spanish Geography has been used to evaluate the Geo-TALP-QA approaches. Out of 62 questions, our system provided the correct answer to 39 questions in the experiment with the best results using the GeoTALP-QA Data-Driven approach (62.9% of accuracy) and 21 questions (33% of accuracy) in the experiment with the Knowledge-Based approach.

The GikiTALP Geographical Question Answering approach uses a Data-Driven approach based on the Sphinx full-text search engine with limited NLP and without using Geographical Knowledge. This approach has been tested in the context of the GikiCLEF 2009 Geographical Question Answering over the Wikipedia. The best results were obtained with the NLP techniques (lemmas in the queries and stopwords filtered) and the Sphinx mode MATCH_EXTENDED algorithms. Although in comparison with other approaches at GikiCLEF this approach had poor results the experiments be useful as a baseline for further developments. These poor results are due to the lack of Geographical Knowledge applied in the task: including Toponym Resolution, Geographical Question Classification, and Answer Extraction.

This Chapter also describes a Geographical Query Parsing Approach that has been evaluated in the context of the official GeoCLEF 2007 GeoQuery task. This approach is based on a linguistic and geographical knowledge analysis of the queries. The global results of our run for the local query were 0.222 Precision, 0.249 Recall, and 0.235 of F1. Our system was ranked the third from 6 participants achieving the top ranked groups F1-Scores of 0.488 and 0.365. The analysis of the results show that the selection of a subset of the most important features to create a gazetteer of only the most important places implies a lost of coverage and thus missing geographical places and classifying queries as non-local.

CHAPTER 6

Textual Georeferencing Approaches

This Chapter describes four generic approaches for Textual Georeferencing of multilingual informal and formal documents and its evaluation in the context of international evaluation benchmarks and posterior experiments with other existing datasets. The georeferencing approaches described in this Chapter are the following ones:

1. Geographical Knowledge Approach.
2. Information Retrieval Approach.
3. Information Retrieval with Re-Ranking Approach.
4. GeoFusion Approach: combines predictions from the Geographical Knowledge-Based approach with predictions computed by one of the IR-based approaches (the IR approach or the IR with Re-Ranking approach). This approach has a variant that can use Georeferenced Wikipedia pages to predict.

The resources employed in the Textual Georeferencing approaches were the Geonames geographical gazetteer, the TFIDF and BM25 IR algorithms, the Hiemstra Language Modelling (HLM) algorithm for IR, the Haversine geographical distance measure, stopwords lists from several languages, and an electronic English dictionary. These approaches have been evaluated with informal and formal documents. The evaluation with informal documents have been performed within the context of the MediaEval Placing Task evaluations at 2010, 2011 and 2014 (M. Larson et al., 2010; Rae et al., 2011; Choi et al., 2014) and posterior experiments with these data sets. The Media Eval Placing tasks evaluated systems that can perform automatic geo-prediction of Flickr[1] photos and videos using visual, audio or textual features. The informal documents evaluated in this thesis are the textual annotations and tagsets associated with Flickr photos/videos. The evaluation of the formal documents has been performed with a set Wikipedia documents provided by the Wikipedia Spot corpus (Van Laere et al., 2014) as a test corpus and a Wikipedia Corpus from B. P. Wing and Baldridge (2011) as a training corpus for the IR approaches.

[1] Flickr is a popular image hosting and video hosting website. `www.flickr.com`

6.1 Geographical Knowledge Approach

The approach of using only Geographical Knowledge for georeferencing has two main phases: Toponym Recognition and Geographical Focus Detection (see the architecture of this approach in Figure 6.1). Some place name disambiguation techniques are applied in both phases: place name normalisation and geo/non-geo ambiguities solved in the first one and the geographical class ambiguities and reference ambiguity applied in the second one. A main difference from the text and gazetteer based approaches systems of Kelm et al. (2010) and J. Perea-Ortega et al. (2010) with respect to the Geoographical Knowledge approach is that this approach do not use Named Entity Recognizers and NLP processors.

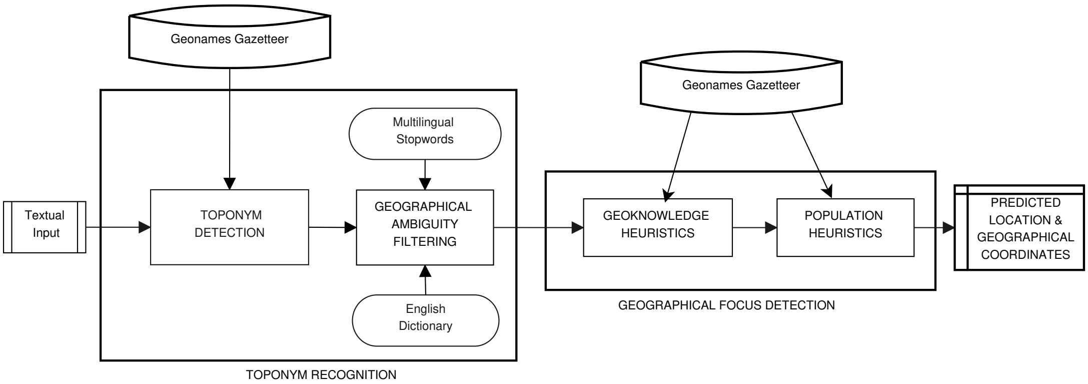


Figure 6.1: Geographical Knowledge-Based Approach.

6.1.1 Toponym Recognition

The Toponym Recognition phase uses the Geonames Gazetteer (explained in Section 2.1.1) for detecting the place names in the textual annotations. NERC was not applied for several reasons: 1) multilingual textual annotations and tags complicate the application of NLP tools such as POS-tagging and NERC because of lowercased entities (e.g. ”barcelona”) and joined named entities (e.g. ”riodejaneiro”) , 2) textual annotations and tags are not suitable for most NERC systems trained in news corpora, and 3) some NERC systems are not performing better than Toponym Recognition from gazzetteer lookup (Stokes et al., 2008). On the other hand is interesting to notice also some limitations of a Gazetteer lookup approach: highly irregular coverage (Popescu et al., 2008), and poor spatial inclusion defined (Popescu and Kanellos, 2009). The Geonames gazetteer has been used in GIR (Toral et al., 2007), Geographical Scope Resolution (Andogah et al., 2008) and textual georeferencing (Serdyukov et al., 2009). The information contained in this gazetteer is used by the toponym recognition phase to deal with the following issues: multilinguality (e.g ”Wien” and ”Viena” refer to the city ”Vienna” with German and Spanish respectively), acronyms (e.g. ”USA”), lowercased place names (e.g. ”lisboa, portugal”) , uppercased place names (e.g ”BARCELONA”), and joined place names (e.g. ”newyorkcity”). The following fields from each Geonames toponym entry are used: country, state, and continent of the place, feature type, coordinates, and population. In the initial experiments at Media Eval 2010 the Gazetteer employed for the recognition of place names was used limiting the n-grams to a maximum of five tokens (5-grams) (e.g. “Sierra Nevada de Santa Marta”), that were increased after the official Media Eval 2010 experiments to seven tokens in new experiments (without improvements noticed). A toponym disambiguation technique is applied to solve the geo/non-geo ambiguity in of toponyms. This ambiguity occurs because there are a high number of place names with meaning that lowercased could be a noun in English or other languages (e.g. ”dog” (noun) vs ”dog” (refering to Dog, a city in Guinea). This phase uses stopwords lists in several languages[2] (including English) and an English Dictionary of 71,348 words obtained from the Freeling[3] toolbox (v2.1) to filter out non-geographical words that could be erroneously tagged as place names. Obviously there are some cases in which a word could refer both to a common word and a geographical toponym (e.g. ”aurora” (noun) vs ”aurora” (city in USA)”, but this algorithm assumes that these cases would be generally rare and could only be solved with complex toponym disambiguation techniques.

6.1.2 Geographical Focus Detection

The Geographical Focus Detection phase uses some of the Toponym Disambiguation strategies presented in the GIR literature (Leidner, 2007). This phase has been designed with some heuristics described in Hauptmann et al. (1999) and Leidner (2007). The one reference per discourse hypothesis: one geographical place/coordinates per image/video is assumed; and if there are no detected place names in the textual annotations the georeference is unresolved. The approach uses topological knowledge and population knowledge. Using the information of all possible referents of all the place names detected by the Toponym Recognition the following sets of heuristics can be applied separately (each set) or in combination:

- **Geographical Knowledge heuristics**. This set of heuristics is similar to the toponym resolution algorithm applied by Hauptmann et al. (1999) to plot on a map locations mentioned in automatically transcribed news broadcasts.

  The system first resolves the geo-class ambiguities in the following way: 1) the ambiguity between country and city names is resolved by giving priority to the country names (e.g. “Brasil” (city in Colombia) versus “Brasil” (country)), 2) the ambiguity between state and city names is resolved by giving priority to city names.

  Once the set of different toponyms appearing in the text have been obtained and geo-class disambiguated then the following sets rules can be applied in priority order to select the scope (focus) and geographical coordinates assigned to the text:

  – H1) The top-priority rule selects the geographical coordinates of the most populated toponym that is not a state, country or continent and has its state appearing in the text as another toponym (e.g. if the toponyms ”San Francisco”, ”Sacramento” and ”California” appear in the text, then the focus will be ”San Francisco”).

  – H2) Otherwise, select the geographical coordinates of the most populated toponym that is not a state, country or continent and has its country appearing

[2] `http://search.cpan.org/dist/Lingua-StopWords`. Includes stopwords for Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Spanish, Swedish, and Russian.

[3] `http://nlp.lsi.upc.edu/freeling/`

in the text as another toponym (e.g. if the toponyms ”San Francisco”, ”Sacramento” and ”United States” appear in the text, then the focus will be ”San Francisco”).

- H3) Otherwise select the geographical coordinates of the most populated toponym that has its country appearing in the text as a toponym (e.g. if the toponyms ”California”, ”Nevada” and ”United States” appear in the text, then the focus will be ”California”).

- **Population heuristics**. These heuristics use only population information to disambiguate between all the possible toponyms. The following rules are used:

  - P1) Select the geographical coordinates associated to the most populated toponym appearing in the text that is not a country, state or a continent.
  - P2) Otherwise select the geographical coordinates associated to the most populated state toponym appearing in the text.
  - P3) Otherwise select the geographical coordinates associated to the most populated country toponym appearing in the text.
  - P4) Otherwise select the geographical coordinates associated to the most populated continent toponym appearing in the text.

- **Geographical Knowledge and Population heuristics**. This method combines the Geographical Knowledge and the Population heuristics presented above in the following way: 1) if it is possible apply the geographical knowledge heuristics H1, H2, H3 in priority order, 2) otherwise apply the population heuristics presented above in priority order.

In case that any heuristic has been activated then the system can leave the query unresolved or it can return a default pair of coordinates.

6.2 Information Retrieval Approach

This approach treats the input text to georeference as an IR query and uses existing state-of-the-art IR models to retrieve from an IR index a set of weighted coordinates relevant to the query (see the architecture of this approach in Figure 6.3). The indexing approach is a point-based approach (modelling individual coordinates) instead of grid-based indexing approaches (modelling spatial regions) commonly used in Serdyukov et al. (2009), Van Laere et al. (2013), Kordopatis-Zilos et al. (2014), and Popescu et al. (2014) among others. The indexing process is performed by the following steps: 1) a filtering out step can be applied in some cases: documents with repeated content and documents without content, 2) then for each unique coordinate pair in the training corpus a new document was created joining the content of all the documents associated to this coordinate pair (see an example of document created in Fig 6.2), 3) this new document is then indexed with its coordinates pair associated as the document name.

The Terrier[4] IR software (version 4.0) was used for indexing with its default settings for each IR model used. The indexing process uses the default Terrier stopwords list to

[4] http://terrier.org

```
<DOCNO>40.783149_-73.95895</DOCNO>
<TEXT>
Guggenheim  2003,guggenheim,manhattan,museum,newyork,nyc
</TEXT>
</DOC>
```

Figure 6.2: Example of a document created for the coordinates pair <40.783149 (latitude),-73.958952 (longitude)> corresponding to a point at the Guggenheim museum of New York.

filter out irrelevant tokens to be indexed. The IR weighting models used for retrieval are the TF-IDF, BM25 and Hiemstra Language Model (HLM) (Hiemstra, 2001) implemented in the Terrier IR system[5][6] Ounis et al., 2006. The retrieval phase obtains the prediction: from the list of ranked documents the geographical coordinates of the top-ranked one are chosen as the final prediction.

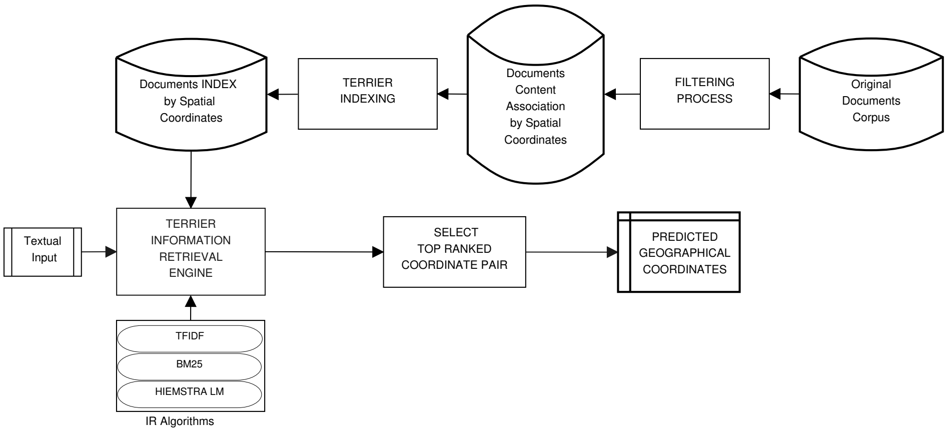

Documents INDEX by Spatial Coordinates
TERRIER INDEXING
Documents Content Association by Spatial Coordinates
FILTERING PROCESS
Original Documents Corpus
Textual Input
TERRIER INFORMATION RETRIEVAL ENGINE
SELECT TOP RANKED COORDINATE PAIR
PREDICTED GEOGRAPHICAL COORDINATES
TFIDF
BM25
HIEMSTRA LM
IR Algorithms

Figure 6.3: Information Retrieval Approach.

The specific details and formulas of the IR weighting models TF-IDF and BM25 implemented by the Terrier IR can be found at Chapter 4 - Section 4.3.2.2. On the other hand, the Hiemstra Language Model (HLM) weighting model Hiemstra, 2001 implemented in Terrier is shown in equation *eq.11*. The equation *eq.11* describes the Terrier implementation of the HLM Weighting model (version 1 Hiemstra, 2001) score of a term t in document d ; where $tf_{t,d}$ is the term frequency in the document, cf_t is the collection frequency of the term, $\sum_i cf_i$ is the number of tokens in the collection, and $\sum_i tf_{i,d}$ is the document length.

$$Score(t,d) = \log{(1 + \frac{\lambda * tf_{t,d} * \sum_i cf_i}{(1-\lambda) * cf_t * \sum_i tf_{i,d}})} \qquad (eq.11)$$

[5] Terrier IR engine. `http://www.terrier.org`

[6] The HLM was used with the default $\lambda = 0.15$ and the BM25 was used with $k_1 = 1.2d$, $k_3 = 8$, and $b = 0.75d$

6.3 Information Retrieval with Re-Ranking Approach

This approach has the same architecture of the Information Retrieval approach for indexing, but for Retrieval uses a different technique. Given an informal text to georeference, this approach treats this text as an IR query and uses existing state-of-the-art IR models to retrieve a set of weighted coordinates relevant to the query and re-rank them with the Haversine geographical distance function (see the architecture of this approach in Figure 6.5). For each unique coordinate pair in the training corpus a document was created with some of the textual metadata fields (specifically the fields: title, description and user tags) content of all the photos/videos that pertain to this coordinate pair. A Re-Ranking process is applied after the IR process. For each query their first top-ranked 1,000 retrieved documents (with its associated coordinate pairs) from the IR software are used. From them we selected the subset of coordinate pairs with a score equal or greater than the two-thirds (threshold 66.66%)[7] of the top-ranked coordinate pair(s) (see an example of a query in Figure 6.4). Then for each geographical coordinate pair of the subset we sum its associated score (provided by the IR software) and the score of their neighbours in the subset at a threshold distance (e.g. 100km) below their Haversine distance. Then we select the one with the maximum weighted sum as the final predicted coordinate pair.

```
16:9,2013,city,flickr,limburg,luxtonnerre,maastricht,shopping,sightseeing,street,
the netherlands,the netherlands/nl,vacation,wallpaper,Men At Work
```

Figure 6.4: Example of a MEPT2004 query (includes keywords present at the Title, Description and Keywords metadata fields).

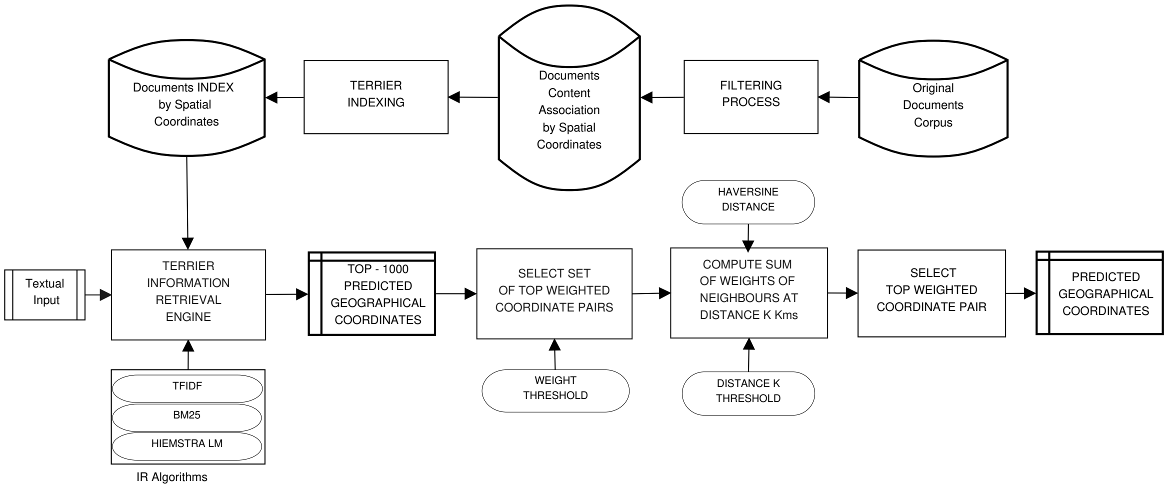

Documents INDEX by Spatial Coordinates
TERRIER INDEXING
Documents Content Association by Spatial Coordinates
FILTERING PROCESS
Original Documents Corpus
HAVERSINE DISTANCE
Textual Input
TERRIER INFORMATION RETRIEVAL ENGINE
TOP - 1000 PREDICTED GEOGRAPHICAL COORDINATES
SELECT SET OF TOP WEIGHTED COORDINATE PAIRS
COMPUTE SUM OF WEIGHTS OF NEIGHBOURS AT DISTANCE K Kms
SELECT TOP WEIGHTED COORDINATE PAIR
PREDICTED GEOGRAPHICAL COORDINATES
TFIDF
BM25
HIEMSTRA LM
IR Algorithms
WEIGHT THRESHOLD
DISTANCE K THRESHOLD

Figure 6.5: Information Retrieval with Re-Ranking approach for Textual Georeferencing.

[7] This threshold has been chosen tuning with the Media Eval 2011 dataset.

6.4 GeoFusion: Knowledge-Based and Data-Driven Georeferencing

The GeoFusion approach combines predictions that come from the GeoKB approach and predictions from the IR or the IR with Re-Ranking approaches (see the architecture of the GeoFusion approach in Figure 6.6) in the following way:

- Step 1. The GeoKB approach is executed with the heuristics H1, H2 and H3. The first heuristic that is activated will return the predicted coordinates.
- Step 2. Otherwise, if the heuristics are not activated, then the prediction will be calculated by one of the IR approaches (with or without Re-Ranking).

Moreover, another variant of the GeoFusion approach called GeoFusion+Wiki (see the architecture of this variant of the approach in Figure 6.7) was implemented. This variant is executed with the same steps of the GeoFusion plus an additional step:

- Step 3: This step is activated when, in the Step 2 the IR or the IR with Re-Ranking approaches return a coordinates pair with a score lower than a threshold[8], then the prediction will be calculated with an IR approach that uses a set of georeferenced Wikipedia pages. This dataset has 857,574 Wikipedia georeferenced pages[9] and is used with the purpose of covering cases of keywords that where not covered by Flickr datasets but could be covered by the Wikipedia (e.g. places that are rarely or never photographed). The coordinates of the top ranked georeferenced Wikipedia page after the IR process are used as a prediction.

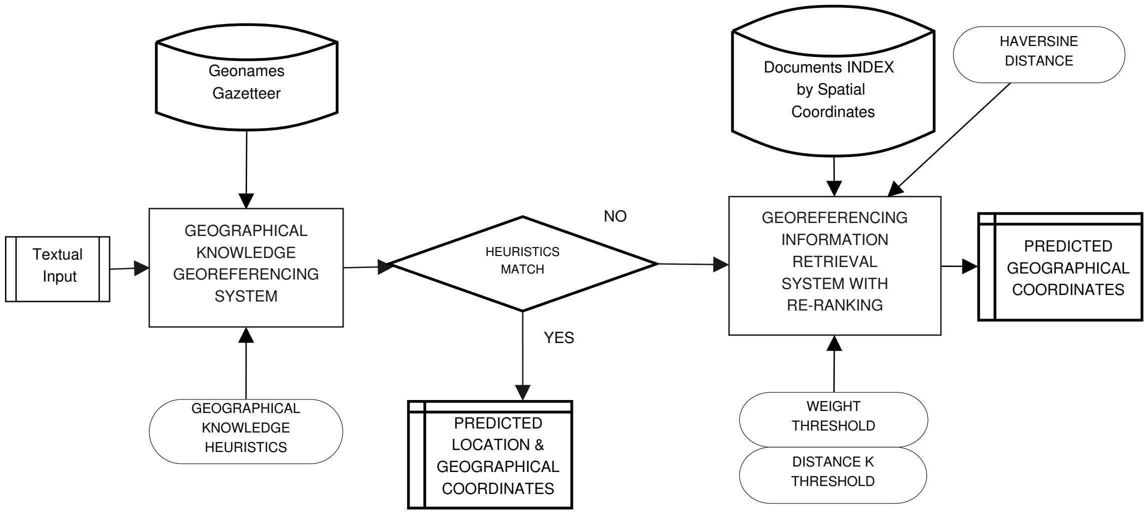

Geonames Gazetteer
Documents INDEX by Spatial Coordinates
HAVERSINE DISTANCE
Textual Input
GEOGRAPHICAL KNOWLEDGE GEOREFERENCING SYSTEM
HEURISTICS MATCH
NO
YES
GEOREFERENCING INFORMATION RETRIEVAL SYSTEM WITH RE-RANKING
PREDICTED GEOGRAPHICAL COORDINATES
GEOGRAPHICAL KNOWLEDGE HEURISTICS
PREDICTED LOCATION & GEOGRAPHICAL COORDINATES
WEIGHT THRESHOLD
DISTANCE K THRESHOLD

Figure 6.6: GeoFusion Approach: Combining Geographical Knowledge Heuristics with Information Retrieval with Re-Ranking Approach for Textual Georeferencing.

[8] A threshold with value 7.0 was found empirically training with the MediaEval 2011 test set.

[9] http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Hauptseite/Wikipedia-World/en

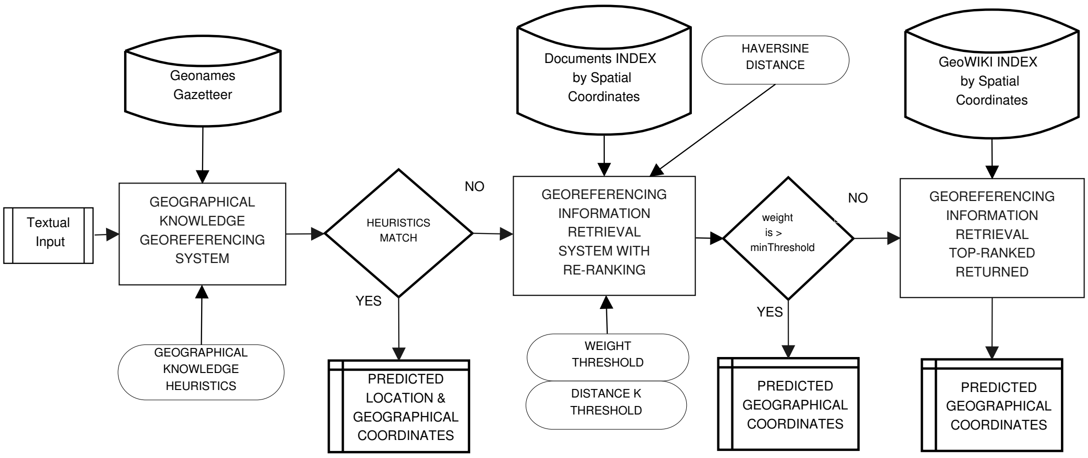

Geonames Gazetteer
Documents INDEX by Spatial Coordinates
HAVERSINE DISTANCE
GeoWIKI INDEX by Spatial Coordinates
Textual Input
GEOGRAPHICAL KNOWLEDGE GEOREFERENCING SYSTEM
HEURISTICS MATCH
NO
YES
GEOREFERENCING INFORMATION RETRIEVAL SYSTEM WITH RE-RANKING
weight is > minThreshold
NO
YES
GEOREFERENCING INFORMATION RETRIEVAL TOP-RANKED RETURNED
GEOGRAPHICAL KNOWLEDGE HEURISTICS
PREDICTED LOCATION & GEOGRAPHICAL COORDINATES
WEIGHT THRESHOLD
DISTANCE K THRESHOLD
PREDICTED GEOGRAPHICAL COORDINATES
PREDICTED GEOGRAPHICAL COORDINATES

Figure 6.7: GeoFusion+Wiki Approach: Combining Geographical Knowledge Heuristics with Information Retrieval with Re-Ranking Approach, and predictions based on Georeferenced Wikipedia pages for Textual Georeferencing.

6.5 Experiments Georeferencing Informal Documents

The experiments for georeferencing informal documents were performed with the MediaEval 2010, 2011, and 2014 Placing task data sets during the official benchmark and posterior experimentation (see in Table 6.1 the experiments performed within the evaluation benchmarks or a posteriori with several different approaches). The MediaEval Placing task required that participants automatically assign geographical coordinates (latitude and longitude) to Flickr videos/images using one or more of: Flickr metadata, visual content, audio content, and social information. In the experiments presented in this thesis only textual content from three metadata fields (title, description, and keywords) from the Flickr videos/images was used to perform the task. The evaluation is performed by calculating the distance (haversine formula) from the actual point (assigned by a Flickr user) to the predicted point (assigned by a georeferencing approach). Runs are evaluated finding how many videos were placed at least within some threshold distances (e.g. 1 km, 5km, 10km, 50km, 100km, 1000km, 5000km). These evaluated threshold distances can change slightly depending on the benchmark.

The following issues were detected for the task of recognizing the place names in textual annotations and tagsets from Flickr:

- lowercased Named Entities. (e.g. “brazil”)
- joined place names (e.g. “riodejaneiro”, “buenosaires”)
- acronyms (e.g. “L.A.”, “NY”, “MN”),
- parts of a toponym instead of using the full place name (e.g. “rio” or “paulo” instead of ”Rio de Janeiro” or ”São Paulo”)

- place names with affixes (“halloweenbrazil”, “brazilguides”, “inbraziltours”)
- multilingual place names (e.g. “Cataratas de Iguaçu”, “iguazufalls”, “iguaçufalls”)
- place name plus a feature name (e.g. “iguazufalls”, “newyorkcity”),
- orthographic errors (e.g. “Rio da Janeiro”, “sao Paulo”).
- geo/non-geo names ambiguity problems: nouns that could be tagged as toponyms and viceversa (e.g. “aurora” (name), “aurora” (place name)),
- referent ambiguity problems with toponyms (e.g. “Barcelona” (Spain) or “Barcelona” (Colombia)).
- geo-class ambiguity problems (e.g. “Madrid” (city) or “Madrid” (state)).

Table 6.1: Experiments for Textual Georeferencing of Informal Documents at official (OF) Media Eval Placing Tasks (MEPT) and posterior experiments (PS).

	Georeferencing Approaches Used				
Experiments	GeoKB	IR	IR+RR	GeoFusion	GeoFWiki
MEPT2010 (OF) (Ferrés and Rodríguez, 2010b)	✓	-	-	-	-
MEPT2010 (PS) (Ferrés and Rodríguez, 2011a)	✓	✓	✓	✓	-
MEPT2011 (OF) (Ferrés and Rodríguez, 2011b)	✓	-	✓	✓	-
MEPT2011 (PS)	-	-	✓	-	-
MEPT2014 (OF (Ferrés and Rodríguez, 2014)	✓	-	✓	✓	✓
MEPT2014 (PS) (Ferrés and Rodríguez, 2015b)	-	-	✓	✓	-

6.5.1 Experiments at MediaEval Placing Task 2010

The MediaEval Placing Task 2010 (MEPT2010) data sets are composed by 5,125 and 5,091 videos (and their metadata) for the development and test sets respectively. See in Table 6.2 a sample of Textual annotations and tagset, its prediction, and the annotated groundtruth. For the MEPT2010 evaluation a set of experiments that consist in tagging the test set and applying different baseline configurations of the Geographical Knowledge Approach was designed (see Table 6.3). The following metadata fields from the Flickr videos were used to perfom the task: Title, Description, and Keywords. The Toponym Recognition system was allowed to recognize place names of a maximum of five tokens (e.g. “Sierra Nevada de Santa Marta”) from the Geonames Gazetteer. The best results (see Table 6.3 and Figure 6.8) are achieved by the TALP_2 run which has geographical knowledge heuristics and population heuristics combined with the use of stopwords and the English dictionary. The number of videos in which their geographical focus could not be predicted were 918 videos for runs TALP_1 and TALP_2, 454 for runs TALP_3 and TALP_4, and 410 for the run TALP_5. In those videos the latitude and longitude were set to 0.0 0.0 because there were no place names detected.

The results of the experiments using the Geographical Knowledge approach at MEPT2010 (see Ferrés and Rodríguez (2010b)) show that: 1) the approach that combine Geographical Knowledge heuristics with population heuristics obtains the best results, and 2) the use of stopwords lists and controlled dictionaries improves slightly the results.

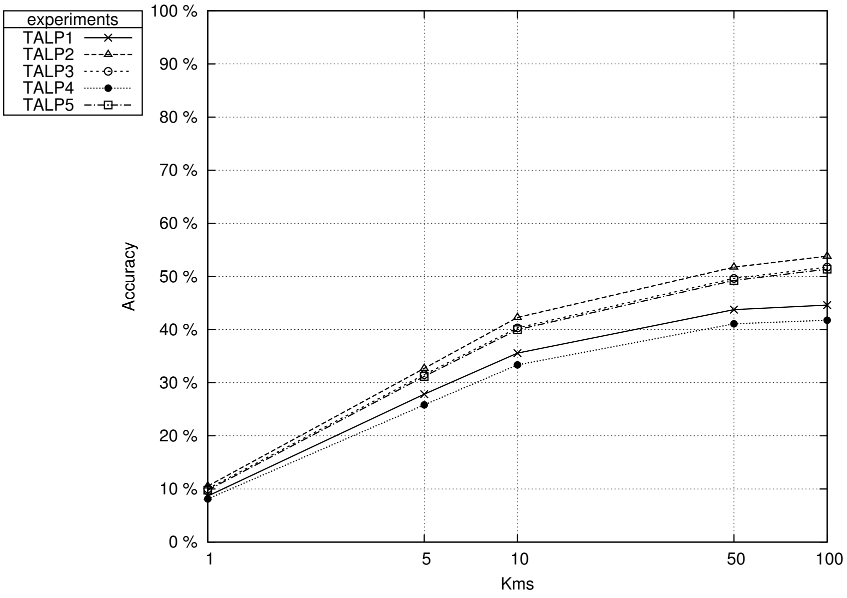
Table 6.2: Georeferencing example: includes the metadata to be predicted, the predicted coordinates, the GeoKB data, the groundtruth and its distance to the prediction.

Title	Lensbaby D90 Video.
Keywords	lonsdalequay, northvancouver, ocean, lensbaby, composer, d90, video, smartcookies
Predicted coordinates	49.31636 -123.06934
GeoKB data	toponym: ”northvancouver” (North Vancouver) feature class: P.PPL (populated place: city, town,...) country: CA (Canada) admin1 code: 02 (BC) latitude: 49.31636 longitude: -123.06934 population: 48000
Groundtruth	latitude: 49.309837 longitude: -123.082108 North Vancouver, British Columbia, Canada
Distance to Groundtruch	1.1760 Km

Table 6.3: Official MEPT2010 experiments with the GeoKB Approach.

	Parameters			Results				
run	Disambiguation	stopwords	dictionary	#videos_correctly_predicted				
TALP_1	population	yes	yes	441	1,417	1,811	2,227	2,271
TALP_2	know+population	yes	yes	**536**	**1,665**	**2,153**	**2,635**	**2,740**
TALP_3	know+population	yes	no	510	1,604	2,052	2,526	2635
TALP_4	population	yes	no	413	1,315	1,698	2,092	2,126
TALP_5	know+population	no	no	497	1,587	2,035	2,507	2,615

Figure 6.8: Official Results at MEPT2010. Accuracy against margin of error in kms.



6.5.2 Experiments after MediaEval Placing Task 2010

After Media Eval 2010 a set of experiments were performed in order to improve the Geographical Knowledge approach and test the other approaches. The experiments performed with the geographical knowledge approach include the following improvements with respect to the preliminar experiments presented in the MEPT2010: 1) filtering out weak geographical named Entities (e.g. the toponym *Porto Alegre* has the weak toponyms Porto and Alegre which could be erroneously matched as a topononym, 2) improving the focus detection phase for cases in which with several toponyms with the same state,country or continent appear in the medatada (in the original MEPT2010 submitted experiments the population sorting of these cases was not activated), 3) adding the Geonames Alternate Names file (with 2.9 million of features).

The experiments performed after MEPT2010 show that the improvements of adding a weak NE filter and the focus detection refinement can improve the results of the GeoKB georeferencing approach . On the other hand the addition of the Alternate Names data does not improve the results. The improvement with respect of the best results at MediaEval 2010 official evaluation is from 2740 videos (0.5382 of accuracy) to 2838 videos (0.5574 of accuracy) correctly georeferenced at a distance maximum of 100 Km.

After the improvement of the Geographical Knowledge Approach the georeferencing experiments done with this approach were the following:

1. *Experiments to detect the relative importance of the metadata fields (title, description, and keywords) for the Flickr georeferencing task:* the results show that the metadata field Keywords (tags) is the most important one, achieving results of 52% of accuracy at 100 Km (see in Table 6.4 the results of these experiments). The inclusion of the Title and/or the Description fields improves the results of the Keywords (tags) alone: Title and Keywords (53.7%), Keywords and Description (54.4%), Title and Keywords and Description (original TALP_2 configuration) achieves the 55.2% of accuracy at 100 Km. The use of only Title and Keywords achieves an accuracy of 21.7% at 100 Km. These results are slightly better than the results of J. Perea-Ortega et al. (2010), that used a Geo-NER for detecting Named Entities in Title and Description achieving an accuracy of 21.3%. Although is not clear how to compare the toponym disambiguation process between the two systems this may indicate that the performance of toponym recognition with Gazetteer lookup using Geonames is performing at state-of-the-art NERC level.

Table 6.4: Experiments with different metadata fields.

Metadata fields	Accuracy (over 5091 videos)				
	1km	5km	10km	50km	100km
Title (T)	0.034	0.076	0.090	0.104	0.109
Description (D)	0.023	0.075	0.098	0.131	0.132
Keywords (K)	0.106	0.331	0.412	0.503	0.520
T + K	0.103	0.336	0.421	0.517	0.537
T + D	0.051	0.135	0.168	0.210	0.217
K + D	**0.108**	0.343	0.429	0.526	0.544
T + K + D	0.105	**0.345**	**0.433**	**0.532**	**0.552**

2. *Experiments to detect the performance and precision of the geographical disambiguation heuristics*: these experiments show the performance and importance of geographical knowledge and population heuristics applied alone or in combination but priorizing first the geographical knowledge ones (see in Table 6.5 the configuration details of these experiments). The results (see Table 6.6) show that the heuristics that apply geographical knowledge without population heuristics obtain the best precision with a 86.36% of correctly predicted videos from the 2,215 predicted in the experiment EXP_4 and a 83.21% of correctly predicted videos from the 2,353 predicted in the experiment EXP_3. The difference between EXP_3 and EXP_4 is that the last one uses stopwords and the English dictionary to filter out ambiguous place names. On the other hand the combination of geographical knowledge and population heuristics in experiments EXP_4 and EXP_5 obtained a precision of 58.36% (with 4,681 predicted videos) and 68.03% (with 4,173 predicted videos). In order to know the relative performance in precision of each specific heuristic that pertains to the geographical knowledge set of heuristics we computed the precision of each rules (applied in priority order) in the context of the experiment EXP_4 (see Table 6.7).

Table 6.5: Configuration of the georeferencing experiments after MEPT2010.

experiment	Geo. Heuristic	StopWords	Dictionary
EXP_1	population	no	no
EXP_2	population	yes	yes
EXP_3	knowledge	no	no
EXP_4	knowledge	yes	yes
EXP_5	knowledge+population	no	no
EXP_6	knowledge+population	yes	yes

Table 6.6: Results of the georeferencing experiments after MEPT2010.

	100km (margin of error)			
experiment	#predictedOK	#predicted	Accuracy	Precision
EXP_1	2,185	4,681	0.4291	0.4667
EXP_2	2,337	4,173	0.459	0.5600
EXP_3	1,958	2,353	0.3846	0.8321
EXP_4	1,919	2,215	0.3769	**0.8636**
EXP_5	2,732	4,681	0.5366	0.5836
EXP_6	2,839	4,173	**0.5576**	0.6803

Table 6.7: Relative performance in precision of each geographical knowledge heuristic data set georefercing up to 100 Km with the experiment EXP_4.

Heuristic	Measures		
Feature (Superordinate)	#predictedOK	#predicted	Precision
H1_city/spot (state)	1,351	1,546	0.8738
H2_city/spot (country)	515	609	0.8456
H3_state (country)	53	60	0.8833

6.5.2.1 Experiments with the IR Approach

The IR indexes were created with a metadata corpus of Flickr photos provided in the MediaEval 2010 for development purposes. The corpus consists of 3,185,258 Flickr photos uniformly sampled from all parts of the world. The photos are georeferenced with geotags with 16 zoom accuracy levels. The accuracy shows the zoom level used by the user when placing the photo on the map ((e.g., 6 - region level, 12 - city level, 16 - street level). The medatada of the corpus is represented by the following information: *UserID*, *PhotoID*, *HTMLLinkToPhoto*, *GeoData* (includes longitude, latitude, and zoom accuracy level), *tags*, *date taken*, and *date uploaded*. From the metadata corpus of photos we filtered out some data: 1) if a user has several photos metadata with the same tagset then only one photo metadata of them is kept, 2) metadata without existing tags is filtered. After this filtering steps a set of 1,723,090 metadata entries for each photo was obtained. Then, from the filtered corpus we selected four subsets depending on the values of the zoom level accuracy: 1) level 16 (715,318 photos), 2) levels from 14 to 16 (1,140,031 photos), 3) levels from 12 to 16 (1,570,771 photos), 4) levels from 6 to 16 (1,723,090 photos). Moreover, for each unique coordinate pair in the corpora all the tagsets associated to the same coordinate pair were joined resulting of: 1) level 16 (511,222 coordinate pairs), 2) levels from 14 to 16 (756,916 coordinate pairs), 3) levels from 12 to 16 (965,904 coordinate pairs), 4) levels from 6 to 16 (1,026,993 coordinate pairs). The indexing of the metadata subsets was done with the coordinates as a document number and their associated tagsets the document text. Indexing was performed by filtering out tokens that match a multilingual stopwords list and without stemming. The retrieval experiments have been done with the metadata of the videos as queries to the IR system. The following metadata fields were used for the query: Keywords (tags), Title and Description. The metadata fields Title and Description were lowercased for the query. The experiments shown in Table 6.8 show that BM25 achieves the best results in accuracies from 10 to 100 Km and the Hiemstra Language Model IR algorithm achieves the best results in accuracies georeferencing up to 1 and 5 km.

6.5.2.2 Experiments with the IR Re-Ranking and GeoFusion Approaches

The GeoFusion approach is applied by combining the results of the Geographical Knowledge approach and the IR approach with Re-Ranking (see the results of this approach in Table 6.9). The results are combined in the following way: from the set of Geographical based experiments we selected the experiment with best precision (EXP_4). From the Geographical Knowledge-based experiment with highest precision the system selects the predicted coordinates, and the ones that are not predicted because the geographical rules do not match are selected from the Information Retrieval approaches with Re-Ranking. This means that from the EXP_4 were selected 2,215 predictions and the rest (2,876 predictions) were selected from the IR with RR approaches. The results of the IR Re-Ranking and the GeoFusion approaches (see Table 6.9) show that both approaches outperform the Geographical and the IR approaches and the baselines. The baselines presented in Table 6.9 are three: 1) the best results obtained at the MEPT2010 with the test set (Van Laere et al., 2010a), 2) the experiment with BM25 trained with a corpus with accuracies from 6 to 16 levels, and 3) the Hiemstra LM trained with accuracies from 14 to 16 levels. These last two baselines were the ones that obtained the best results in accuracies compared to the other IR and corpus training models.

Table 6.8: Results of the georefencing experiments with the Information Retrieval Approach.

Model	Accuracy				
	1km	5km	10km	50km	100km
	annotation accuracy=16				
BM25	0.4236	0.5055	0.5395	0.5951	0.6091
TFIDF	0.4227	0.5028	0.5362	0.5912	0.6059
HLM	0.4309	0.5054	0.5356	0.5989	0.6130
	annotation accuracy=14-16				
BM25	0.4236	0.5063	0.5446	0.5990	0.6120
TFIDF	0.4227	0.5044	0.5417	0.5939	0.6065
HLM	**0.4364**	0.5124	0.5474	0.6079	0.6218
	annotation accuracy=12-16				
BM25	0.4203	0.5044	0.5515	**0.6065**	0.6216
TFIDF	0.4201	0.5040	0.5494	0.6028	0.6179
HLM	0.4350	**0.5146**	0.5515	0.6042	0.6201
	annotation accuracy=6-16				
BM25	0.4142	0.5016	**0.5527**	0.6063	**0.6244**
TFIDF	0.4136	0.5012	0.5505	0.6028	0.6201
HLM	0.4284	0.5107	0.5494	0.6049	0.6220

The experiments show that stopwords lists and controlled dictionaries can help the disambiguation of placing names and the focus detection. The experiments also show that geographical knowledge heuristics can achieve a high precision in georeferencing: up to a 86.36%. This fact is very interesting for establishing high confidence rules that could allow a high precision georeferencing detection in textual annotations and tags. The strategy that combines geographical knowledge and population heuristics for geographical focus detection achieves the best results in the experiments with the Geographical approach with the MEPT2010 data set. The Information Retrieval approaches outperformed the Geographical one, but the fusion of both is achieving the best results. The best approach georeferencing up to 1, 5 and 10 km is achieved with the Information Retrieval Re-ranking approach with the Hiemstra LM. The best results in accuracy up to 50 and 100Km are achieved with the fourth strategy: a fusion of Information Retrieval Re-ranking with Geographical Knowledge approaches. These strategies outperformed the best results in accuracy reported by the state-of-the art systems participating at MEPT2010. The best results of accuracy georeferencing up to a distance of 100 Km are 68.53% and obtained with the GeoFusion approach with IR Re-Ranking at a distance of 100km. The approaches of Van Laere et al. (2010a) and Kelm et al. (2010) obtained a 67,23% and 60,46% of accuracy with the same test set at the MEPT2010.

Table 6.9: Results of the experiments with IR Re-Ranking and with GeoFusion with the MEPT2010 data (in bold the results that improve the MEPT2010 best results). Note that the IR Re-Ranking experiments are specified with the following syntax: 1) first the name of the IR algorithm (e.g. BM25, HLM, or TFIDF), 2) then a '@' symbol, 3) finally the clustering threshold in kms. The GeoFusion experiments follow the same syntax of the IR Re-Ranking plus the following string: "+GeoKB" (indicating that combines predictions from the GeoKB and the IR Re-Ranking approaches). The "+GeoKB" string also indicates that the Geographical Knowledge-Based approach has been applied only with the H1,H2,H3 heuristics.

Experiments	Accuracy				
	1km	5km	10km	50km	100km
	Baselines				
Best MEPT2010 (Van Laere et al., 2010a)	0.4329	0.5425	0.5879	0.6509	0.6723
BM25 (annotation accuracy 6-16)	0.4142	0.5016	0.5527	0.6063	0.6244
HLM (annotation accuracy 14-16)	0.4364	0.5124	0.5474	0.6079	0.6218
	Experiments at different Re-Ranking distances				
BM25@1km	0.4331	0.5134	0.5507	0.6057	0.6230
BM25@1km+GeoKB	0.2598	0.4549	0.5246	0.6307	0.6552
HLM@1km	**0.4535**	0.5336	0.5690	0.6338	0.6491
HLM@1km+GeoKB	0.2728	0.4670	0.5391	0.6491	0.6733
BM25@5km	0.3698	0.5266	0.5631	0.6216	0.6375
BM25@5km+GeoKB	0.2427	0.4633	0.5332	0.6389	0.6643
HLM@5km	0.4030	**0.5433**	0.5739	0.6468	0.6595
HLM@5km+GeoKB	0.2541	0.4761	0.5470	0.6590	0.6823
BM25@10km	0.3688	0.5055	0.5772	0.6256	0.6399
BM25@10km+GeoKB	0.2429	0.4568	0.5389	0.6409	0.6660
HLM@10km	0.4030	0.5275	**0.5894**	0.6485	0.6611
HLM@10km+GeoKB	0.2563	0.4704	0.5523	**0.6611**	0.6847
BM25@50km	0.3496	0.4680	0.5124	0.6340	0.6470
BM25@50km+GeoKB	0.2304	0.4378	0.5148	0.6438	0.6682
HLM@50km	0.3834	0.4928	0.5427	0.6482	0.6599
HLM@50km +GeoKB	0.2439	0.4553	0.5346	0.6590	0.6831
BM25@100km	0.3500	0.4635	0.5008	0.5957	0.6485
BM25@100km+GeoKB	0.2309	0.4399	0.5116	0.6318	0.6702
HLM@100km	0.3838	0.4902	0.5358	0.6187	0.6609
HLM@100km+GeoKB	0.2433	0.4539	0.5299	0.6464	**0.6853**

6.5.3 Experiments at MediaEval Placing Task 2011

The Georeferencing experiments at MediaEval Placing Task 2011 (MEPT2011) evaluation benchmark (see Rae et al. (2011) for more details about this evaluation) were performed with the three following approaches: 1) the Geographical Knowledge approach, 2) the Information Retrieval based approach with Re-Ranking, and 3) the GeoFusion Approach.

Two corpus were used for training the IR system for MEPT2011: 1) the MEPT2011 Flickr corpus (3,185,258 photos) and 2) the union of the MediaEval corpus with the CoPhIR[10] image collection (Bolettieri et al., 2009) (106 million processed images). From the MediaEval corpus we filtered and extracted 1,026,993 coordinates (accuracies between 6 and 16 zoom levels) with their associated tagsets. From CoPhIR we selected the photos with geographical referencing with accuracies between 6 and 16 zoom levels (8,428,065 photos). Then we filtered repeated content and null content (7,601,117 photos). The union of the extracted data from CoPhIR and MediEval gives a total of 2,488,965 unique coordinates with associated tagsets.

A set of four experiments (see Table 6.10) was designed for the MEPT2011 test set of 5347 Flickr videos. The experiment *TALP1* used the IR approach with Re-Ranking up to 100 km and the MEPT2011 photos corpus as a training data. The experiment *TALP2* used the GeoKB approach. The experiment *TALP3* used the GeoFusion approach with the MediaEval training corpora. The experiment *TALP5* used the GeoFusion approach with the MediaEval and the CoPhIR corpora of photos for training. In all the experiments that use the IR with Re-Ranking approaches (*TALP1*, *TALP*3 and *TALP5* the HLM IR algorithm with the default parameters: 1) weight clustering threshold (66.66% of the top-ranked document), and 2) clustering distance threshold (100km). The results are shown in Table 6.11 and Figure 6.9.

Table 6.10: MediaEval Placing Task 2011 Experiments.

run	Approach	Training Corpus
TALP1	IR (HLM) Re-Rank (100km)	MediaEval (Flickr)
TALP2	GeoKB	-
TALP3	GeoFusion: IR (HLM) Re-Rank (100km)+GeoKB	MediaEval (Flickr)
TALP5	GeoFusion: IR (HLM) Re-Rank (100km)+GeoKB	MediaEval (Flickr)+ CoPhIR

The GeoFusion approach achieved the best results within the margin of errors from 10km to 10,000km. and the IR with Re-Ranking achieved the best results in accuracy in the margin of error of 1km. The GeoFusion clearly outperformed the other two approaches the IR with Re-Ranking and the Geographical Knowledge Based approach. This approach achieves the best results because combines high precision rules based on Toponym Disambiguation heuristics and predictions that come from a data driven IR Re-Ranking approach. The GeoKB rules H1,H2 and H3 were activated in 2,231 videos from a 5,347 of total videos. These rules achieved achieved 80.18% of accuracy (1,789 correctly predicted videos of 2,231 videos predicted) predicting up to 100km.

[10] **CoPhIR**. `http://cophir.isti.cnr.it`

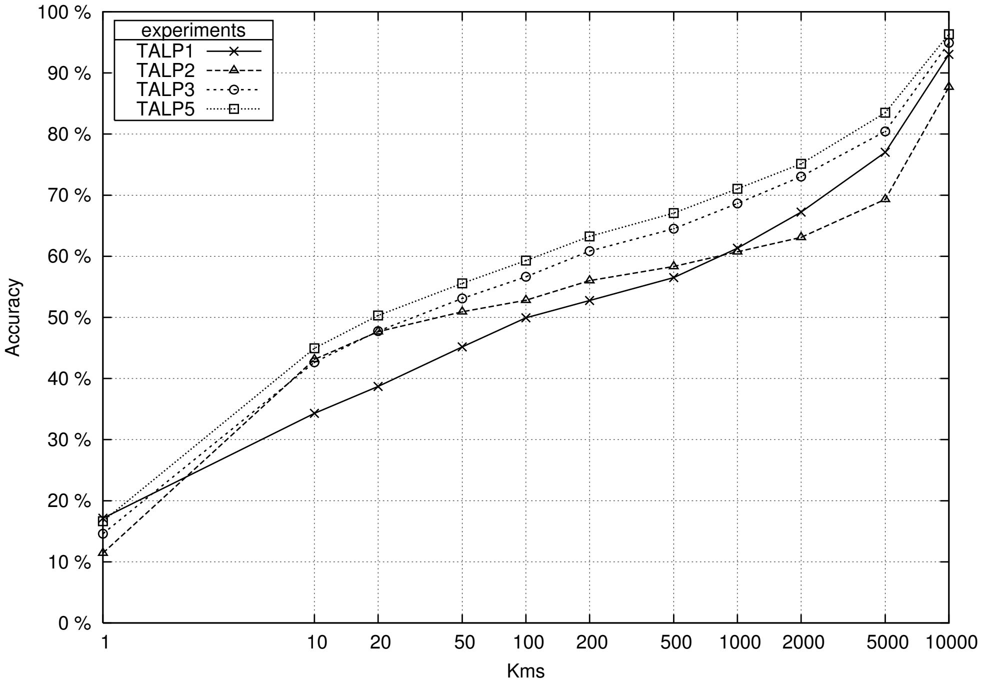
Table 6.11: Official results of the TALP-UPC approach at the MediaEval Placing Task 2011 (predicting 5,347 videos). Number of correctly classified videos and accuracy.

Margin	TALP1	TALP2	TALP3	TALP5
1km	**916 (0.1713)**	611 (0.1147)	781 (0.1461)	890 (0.1664)
10km	1,834 (0.3430)	2,306 (0.4313)	2,281 (0.4266)	**2,403 (0.4494)**
20km	2,070 (0.3871)	2,549 (0.4767)	2,553 (0.4774)	**2,690 (0.5031)**
50km	2,415 (0.4517)	2,723 (0.5093)	2,840 (0.5311)	**2,971 (0.5557)**
100km	2,670 (0.4993)	2,823 (0.5280)	3,029 (0.5665)	**3,171 (0.5930)**
200km	2,821 (0.5226)	2,995 (0.5601)	3,253 (0.6084)	**3,382 (0.6325)**
500km	3,022 (0.5652)	3,119 (0.5833)	3,450 (0.6452)	**3,587 (0.6708)**
1000km	3,278 (0.6130)	3,247 (0.6073)	3,670 (0.6864)	**3,799 (0.7105)**
2000km	3,594 (0.6722)	3,374 (0.6310)	3,906 (0.7305)	**4,017 (0.7513)**
5000km	4,119 (0.7703)	3,706 (0.6931)	4,301 (0.8044)	**4,465 (0.8350)**
10000km	4,975 (0.9304)	4,688 (0.8768)	5,076 (0.9493)	**5,151 (0.9633)**

Table 6.12: Results of the official Media Eval Placing Task 2011, evaluated over the 5347 test videos for 2011. The data from this table was extracted from Van Laere (2013).

Group	accuracy (% of video correctly predicted)				
	1km	10km	100km	1000km	10000km
UNICAMP (L. T. Li et al., 2011)	0.21%	1.12%	2.71%	12.16%	79.45%
CUT (Krippner et al., 2011)	9.86%	21.49%	29.79%	43.26%	84.16%
TALP-UPC (Ferrés and Rodríguez, 2011b)	14.61%	42.66%	56.65%	68.64%	94.93%
ICSI (Choi et al., 2012)	20.00%	38.20%	52.60%	66.30%	94.20%
WISTUD (Hauff and Houben, 2011)	17.20%	50.76%	**70.77%**	82.61%	97.21%
Ghent (Van Laere et al., 2011b)	**24.20%**	**51.49%**	63.27%	**85.62%**	**97.85%**

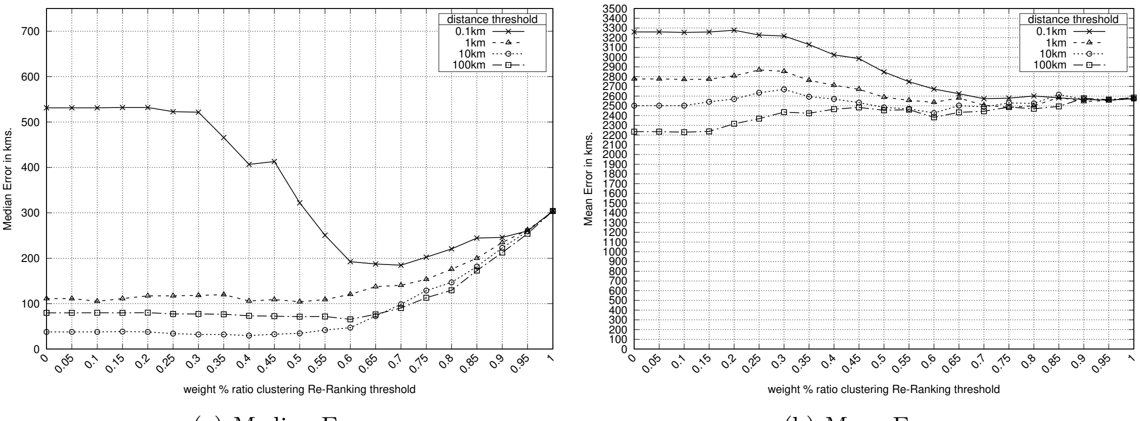
Figure 6.9: Official results of the TALP-UPC approach at the MediaEval Placing Task 2011. Plot of the accuracy against margin of error in kms.



6.5.4 Experiments after MediaEval Placing Task 2011

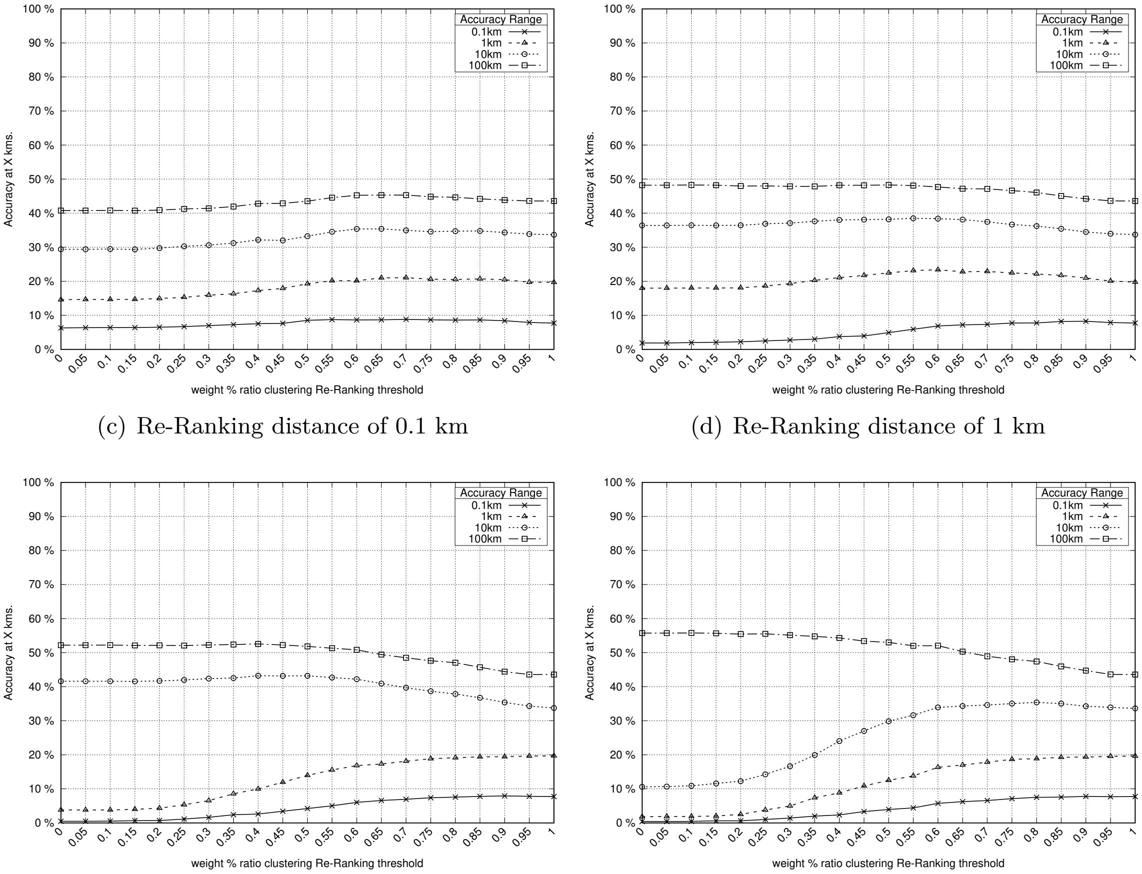
A set of development experiments has been done with the MEPT2011 training and test set data. These experiments were done to detect appropiate thresholds for the HLM Re-Ranking algorithm. The Re-Ranking algorithm has two main thresholds: 1) the 1,000 top-ranked documents weight threshold, 2) clustering distance threshold. The top ranked documents weight threshold indicates at which percentage of the weight of the top-ranked document is established the threshold to create the set of documents that will be used in the reranking process (i.e. a threshold of 100% indicates that only the documents with the same weight of the top-ranked document will be selected for the clustering process, and a threshold of 0% indicates that all 1,000 top-ranked documents will be used). Note that at lower thresholds the clustering processing time is increased. The distance clustering threshold indicates at which distance between geographical coordinates associated to documents (from the set of documents selected for clustering) their weight will be re-calculated by adding the weight of their neighbour at certain distance. The MediaEval 2011 Flickr corpus has 3,185,258 photos for training. From the MediaEval corpus we filtered and extracted 1,026,993 coordinates (accuracies between 6 and 16 zoom levels) with their associated tagsets. The test set consists of 5,347 Flickr videos. Four experiments were performed in order to test all the threshold ranges of the following clustering Re-Ranking distances: 0.1 km, 1 km, 10 km and 100km. The experiments show the Mean error in kms, Median error in kms, and the accuracies of prediction at different Kms (0,1, 1, 10 and 100 kms respectively) of the different thresholds (clustering Re-Ranking distance threshold and top-ranked documents weight threshold). See these results in Figures 6.10 and 6.11. The results indicate that both re-rankings at distance 10km and 100km outperform the 0.1km and 1km distances in median and mean error rates in documents weight thresholds below 70% (of the weight of the top-ranked document).

Figure 6.10: Median and Mean error in Kms at different clustering Re-Ranking distances and threshold values.



(a) Median Error

(b) Mean Error



(c) Re-Ranking distance of 0.1 km

(d) Re-Ranking distance of 1 km

(e) Re-Ranking distance of 10 kms

(f) Re-Ranking distance of 100 kms

Figure 6.11: Accuracies at different Kms. for clustering Re-Ranking at 0.1km, 1km, 10km, and 100kms at different clustering threshold values.

6.5.5 Experiments at MediaEval Placing Task 2014 and Posterior Results

The MediaEval Placing Task 2014 (ME2014PT) required that participants use systems that automatically assign geographical coordinates (latitude and longitude) to Flickr photos and videos using one or more of the following data: Flickr metadata, visual content, audio content, and social information (see Choi et al. (2014) for more details about this evaluation). The ME2014PT training data consisted of 5,000,000 geotagged photos and 25,000 geotagged videos, and the test data consists of 500,000 photos and 10,000 videos. This data has been extracted from the YFCC100M[11] dataset (Yahoo Flickr Creative Commons 100M) (Thomee et al., 2015). This resource has 99.3 million images and 0.7 million videos.

6.5.5.1 Official Experiments at MediaEval Placing Task 2014

A set of four experiments was designed for the MEPT2014 (Main Task) test set of 510,000 Flickr photos and videos (see the description of the experiments in Table 6.13 and the results in Figure 6.12 and Table 6.14):

1. The experiment *run1* used the HLM approach with Re-Ranking up to 100 km and the MediaEval 2014 training set metadata as a training data. From a set of 5,050,000 photos and videos of the MediaEval 2014 training set, a set of 3,057,718 coordinate pairs with related metadata info were created as textual documents and then indexed with Terrier.

2. The experiment *run3* used the GeoKB approach.

3. The experiment *run4* used the GeoFusion approach with the MediaEval training corpora.

4. The experiment *run5* used the GeoFusion approach with the MediaEval training corpora in combination with the English Wikipedia georeferenced pages HLM model.

Table 6.13: MediaEval Placing Task 2014 Experiments.

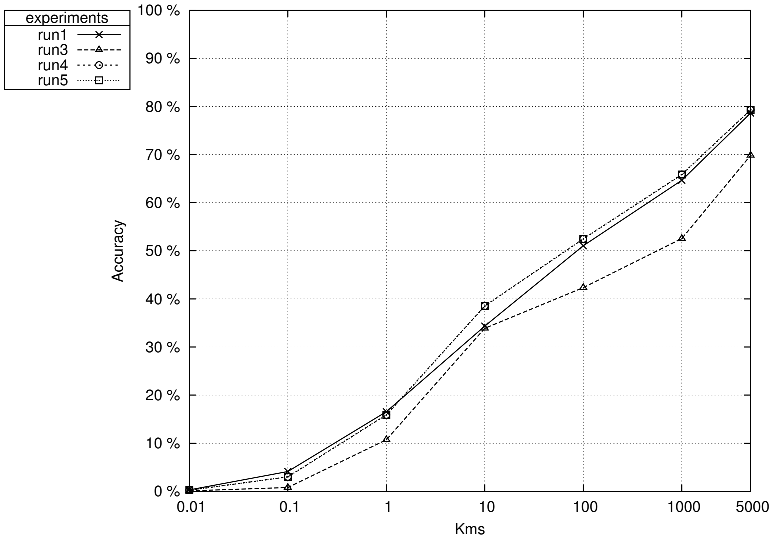
run	Approach
run1	IR (HLM) Re-Rank (100km)
run3	GeoKB
run4	GeoFusion: IR (HLM) Re-Rank (100km) + GeoKB
run5	GeoFusionWiki: R(HLM) Re-Rank (100km) + GeoKB+ GeoWiki

[11] http://www.yli-corpus.org/

Table 6.14: Official TALP-UPC Results at Media Eval Placing Task 2014. Percentage of correctly georeferenced photos/videos within certain amount of kilometers and median error for each run.

Margin	run1	run3	run4	run5
10m	**0.29**	0.08	0.23	0.23
100m	**4.12**	0.80	3.00	3.00
1km	**16.54**	10.71	15.90	15.90
10km	34.34	33.89	38.52	**38.53**
100km	51.06	42.35	**52.47**	**52.47**
1000km	64.67	52.54	**65.87**	65.86
5000km	78.63	69.84	**79.29**	79.28
Median Error (kms)	83.98	602.21	64.36	64.41

Figure 6.12: Official TALP-UPC Results at Media Eval Placing Task 2014. Accuracy against margin of error in kms



6.5.5.2 Experiments after MediaEval Placing Task 2014

The approaches were tested with three corpora for training (see Table 6.15): 1) the ME2014 Training dataset, 2) YFCC100M_A, the YFCC100M geotagged dataset (47,959,829 geotagged items) with items that are not contained in the test set, and 3) YFCC100M_B, the YFCC100M geotagged dataset (47,959,829 geotagged items) with items that are not contained in the test set and items that do not pertain to any user of the test set. From the ME2014PT training and the YFCC100M geotagged datasets we extracted all the unique coordinates with associated text: about 2,741,717, 11,382,289, and 11,253,099 coordinates respectively.

Table 6.15: Features of the training corpus in official and posterior experiments.

Training Corpus	#items	#unique_coordinates	#coordinates_with_text	#users
MEPT2014	5,025,000	3,057,718	2,741,717	172,024
YFCC100M_A	47,959,829	12,578,450	11,382,289	212,877
YFCC100M_B	44,000,224	11,619,425	11,253,099	205,988

Two sets of experiments were performed:

1. **Official experiments with the ME2014PT dataset and posterior experiments with and without gazetteer use.** The results of this experiments are shown in Table 6.16. The official run1 at the benchmark was done with the HLM model and a distance threshold of 100km for Re-Ranking and it achieved the best official results in accuracies at high distances (1,000km and 5,000km). It is worth noting that in the benchmark there is not a system performing well in all distances. The GeoFusion approaches achieved the best results in the experiments at ranges from 10 km to 5,000 km with the ME2014PT Training dataset, clearly outperforming the GeoKB, IR, and IR with Re-Ranked approaches. The GeoFusion approaches achieved the best results at these evaluation ranges because this approach combines high precision rules based on Toponym Disambiguation heuristics and predictions that come from an IR model when these rules are not activated. When these rules are activated (144,074 cases of 510,000), they achieve accuracy percentages of 87.37% (125,878 of 144,074 items) predicting up to 100 km. By contrast, the HLM IR model trained with the ME2014PT training set with Re-Ranking achieved a 78.34% of accuracy at 100 km when evaluated over this subset (144,074 cases). The HLM approach with Re-Ranking obtained the best results in distance ranges from 10m to 1 km because it captures non-geographical highly descriptive and unique keywords and place names appearing in the geographical coordinates' associated metadata that are not present in the gazetteer. The approach that uses the English Wikipedia georeferenced pages to handle difficult cases does not generally offer better performance than the original GeoFusion approach.

Table 6.16: Results of Run1 at ME2014PT (use provided training dataset only) and posterior experiments (without and with gazetteers used).

		accuracy percentage						
	System	10m	100m	1km	10km	100km	1000km	5000km
Benchmark Results	CEALIST(Popescu, et al, 2014)	0.01	0.61	22.62	40.00	47.36	61.17	74.94
	RECOD(L. Li et al., 2014)	0.55	**6.06**	21.04	37.59	46.14	61.69	76.76
	SonSensCERTH(Kordopatis et al 2014)	0.50	5.85	23.02	39.92	46.87	60.11	74.80
	UQ-DKE(Cao et al., 2014)	**1.07**	4.98	19.57	**41.71**	**52.46**	63.61	77.28
	USEMP(Popescu, et al, 2014)	0.78	1.61	**23.48**	40.77	48.11	61.79	75.30
	ICSI/TUDelft (Choi and X. Li, 2014)	0.24	3.15	16.65	34.70	45.58	60.67	75.03
	TALP-UPC [12] (Ferrés, et al, 2014)	0.29	4.12	16.54	34.34	51.06	**64.67**	**78.63**
Post-Evaluation Experiments	GeoKB	0.07	0.89	11.31	34.44	42.26	48.45	58.32
	HLM top-ranked	0.46	5.58	20.07	37.17	46.34	60.40	75.59
	HLM@10km	0.29	4.18	17.35	**41.99**	50.97	63.38	77.91
	HLM@1km	0.30	4.65	**24.03**	41.10	49.53	62.20	75.79
	HLM@0.1km	0.46	**7.20**	22.29	38.37	46.86	60.10	74.59
	TFIDF@100km	0.29	4.21	16.84	34.32	50.15	63.52	77.69
	BM25@100km	0.29	4.24	17.01	34.63	50.60	63.88	77.93
	HLM@100km+GeoKB	0.25	3.25	16.82	39.71	**53.61**	**66.78**	**80.06**
	HLM@10km+GeoKB	0.26	3.32	17.30	**43.48**	**53.47**	**65.67**	**79.47**
	HLM@1km+GeoKB	0.25	3.56	20.74	**42.80**	52.36	**64.76**	77.48
	HLM@0.1km+GeoKB	0.35	5.03	19.69	40.95	50.53	63.22	76.58
	TFIDF@100km+GeoKB	0.25	3.19	16.72	39.34	**53.07**	**66.10**	**79.39**
	BM25@100km+GeoKB	0.25	3.21	16.83	39.53	**53.31**	**66.30**	**79.52**
	HLM@100km+GeoKB+Wiki	0.25	3.25	16.82	39.72	**53.61**	**66.77**	**80.05**

2. **Official experiments with the use of external data and gazetters allowed and posterior experiments with the YFCC100M geotagged dataset.** The results and details of these experiments are shown in Table 6.17. In these experiments the official results obtained were not so good and achieved only the median (of all participants) in distances higher than 10km. In this case the CEALIST and USEMP (Popescu et al., 2014) systems[13] got the best results. On the other hand, the GeoFusion approaches trained with the YFCC100M_A only improve slightly the IR models in accuracy ranges from 1,000 km to 5,000 km. The results with the YFCC100M_A geotagged dataset as a training data lead to the following conclusions: 1) with YFCC100M_A data, the accuracy of the Data-Driven approach outperforms the GeoKB approach, 2) although the YFCC100M_A geotagged dataset used in this study had filtered out the items appearing in the test set, some users with items in the test set could have also items in the train set, and this fact could lead the IR model to have a gain by modeling user's particular way of tagging (M. Larson et al., 2015).
In comparison with the results of the other participants, the IR with Re-Ranking and GeoFusion approaches achieved state-of-the-art results at ME2014PT evaluation. The HLM with Re-Ranking approach obtained the best results for accuracies at distances of 1,000 km and 5,000 km in the task where only the official training data can be used to predict. In posterior experiments using the YFCC100M_A geotagged dataset, the IR with Re-Ranking and GeoFusion approaches outperformed the best results for

[12] This run used the HLM@100km approach (Re-ranking at 100km).

[13] In these official experiments CEALIST and USEMP systems were trained with the YFCC100M_A geotagged dataset.

accuracies from 10m to 100m with accuracy percentages of 20.63% and 26.64%.

A final experiment has been done to assess the effects of filtering out all those users of the training set (YFCC100M) that have items that appear also in the test set. The experiment used the IR HLM with Re-Ranking approach at 100km with the YFCC100M_B geotagged dataset which has those users filtered as a training set. The results show an important difference compared with the same algorithm using the YFCC100M_A dataset. Thus seems that this results confirm that the observation that using the YFCC100M_A dataset could lead to model user's particular way of tagging done by M. Larson et al. (2015).

Table 6.17: Overall official best results at ME2014 runs (anything allowed except crawling the exact items of the test set) and posterior experiments (training with YFCC100M geotagged).

		accuracy percentage						
	System	10m	100m	1km	10km	100km	1000km	5000km
Benchmark Results	CEALIST(Popescu, et al, 2014)	0.01	1.22	40.25	55.98	62.26	72.14	81.95
	RECOD(L. Li et al., 2014)[14]	0.59	**6.26**	21.15	37.50	46.03	61.41	75.07
	SonSensCERTH(Kordopatis et al 2014)	0.50	5.85	23.02	39.92	46.87	60.11	74.80
	UQ-DKE(Cao et al., 2014)	1.08	5.05	20.23	43.68	56.03	69.08	81.14
	USEMP(Popescu, et al, 2014)	**2.56**	4.33	**44.14**	**61.34**	**69.10**	**78.69**	**86.52**
	ICSI/ TUDelft(Choi and X. Li, 2014)	0.32	3.41	12.13	19.95	22.82	33.79	53.06
	TALP-UPC[15](Ferrés, et al, 2014)	0.23	3.00	15.90	38.52	52.47	65.87	79.29
Post-Evaluation	training with YFCC100M_A geotagged photos/videos							
	HLM@100km	**20.63**	**26.64**	40.65	56.13	68.52	76.60	84.76
	BM25@100km	**19.96**	**26.10**	40.30	55.80	68.30	76.72	85.69
	TFIDF@100km	**19.84**	**25.97**	40.11	55.57	68.06	76.54	85.56
	HLM@100km+GeoKB	**13.72**	**18.14**	32.62	54.53	67.49	77.05	86.10
	BM25@100km+GeoKB	**13.20**	**17.64**	32.16	54.05	67.09	76.83	85.97
	TFIDF@100km+GeoKB	**13.12**	**17.55**	32.03	53.88	66.91	76.69	85.87
	training with YFCC100M_B (geotagged without users in test set)							
	HLM@100km	0.36	4.53	17.27	34.10	51.31	64.95	78.52

[14] In this run they used both textual and visual features.

[15] This run used the GeoFusion approach with the HLM model with Re-Ranking at 100km.

6.6 Experiments Georeferencing Informal Texts in Emergency Scenarios

A set of experiments with a subset of the MEPT2014 test set has been performed to evaluate the TG algorithms in emergency scenarios over informal texts from social networks. The experiments done in this article have been inspired by the ones described in Kordopatis-Zilos et al., 2015, in which a set of items (about 6,000) from the MEPT2014 test set with keywords related with emergency scenarios (6 keywords) have been extracted and evaluated compared with the predictions of the full test set. Kordopatis-Zilos et al., 2015 used the following keywords to create an ”emergency scenario” test set derived from MEPT2014 test set: “demonstration”, “earthquake”, “fire” , “flood”, “hurricane” and “riot”. The experiments presented in this thesis only used 4 of these keywords (“earthquake”, “flood”, “hurricane” and “riot’) because of the ambiguity of the “demonstration” and “fire” keywords (see in Table 6.18 some samples of the MEPT2014 test set that include these keywords).

Table 6.18: Sample of some metadata associated to Flickr photos that include at least one of these keywords: ”earthquake”, ”flood”, ”hurricane”, and ”riot”.

ID	Keywords
6091040955	View of the Raritan River’s South Branch from Downtown Clinton clinton **flood** hunterdon **hurricane** new jersey raritan river storm Sunday Morning Aug 28th as **Hurricane** Irene leaves the area
4119131675	Rapids bridge cork **flood** flooding ireland rain rapid rescue submerged water
3848366492	Roio Poggio 138 giorni dopo **earthquake** l’aquila poggio roio terremoto
7788469962	Untitled activists affairs brussels demonstration dog foreign ministry palestine police pro protest **riot** Brussels Ministry of Foreign affairs
4295769184	Larry King was recording downstairs c cnn hospital larry king live As we were checking in CNN was setting up for an interview between Larry King and a Haiti **earthquake** survivor that is being treated at the same hospital where C was eventually born
3846634179	Greater Baton Rouge Food Bank crisis corps **hurricane** katrina louisiana peace corps response
15487966	Mississippi River **Flood** May 2002 **flood** martin luther king bridge mississippi river riverfront saint louis st louis <a href=”http://www geobloggers com”>GeoTagged</a>
5352575983	Pay your tax london revolts **riot riots** student tax Taken during December 2010 student revolts

From the set of 510,000 items (photos/videos) of the MEPT2014 test set were extracted a total of 784 items that contain at least one of these keywords using the Keywords (user tags) metadata field. A total of 63 items contain the keyword “riot”, 229 items contain the keyword “hurricane”, 366 contain the keyword “flood”, and 144 items contain the keyword “earthquake”. The algorithms evaluated were the GeoKB, the IR with Re-Ranking at a clustering distance of 100km and a document weight threshold of 0.66 with the HLM model, and the GeoFusion approach. See the results of the approaches over the full MEPT2014 test set and the emergency scenario test set in Figure 6.19.

The evaluation of the emergency scenario data set shows that the GeoKB approach achieved the best results in the ranges of 0.01km, 0.1km, 1km and 10km in comparison with the other approaches over the same test set. On the other hand, the GeoFusion approach achieved the best results in the ranges of 100km, 1,000km and 5,000km. These results of

Table 6.19: Results of evaluating the emergency scenario data set (784 items) extracted from ME2014PT test set compared with the full MEPT2014 test set (510,000 items)

	accuracy percentage						
System	10m	100m	1km	10km	100km	1000km	5000km
MEPT2014 test set (510,000 items)							
GeoKB	0.07	0.89	11.31	34.44	42.26	48.45	58.32
HLM@100km	**0.29**	**4.12**	16.54	34.34	51.06	64.67	78.63
HLM@100km+GeoKB	0.25	3.25	**16.82**	**39.71**	**53.61**	**66.78**	**80.06**
MEPT2014 Emergency scenario test set (784 items)							
GeoKB	**0.51**	**1.79**	**21.17**	**53.70**	64.41	76.79	89.03
HLM@100km	0.00	1.28	12.76	43.11	67.22	86.10	93.24
HLM@100km+GeoKB	**0.51**	1.53	19.77	52.68	**71.81**	**88.27**	**94.39**

the approaches in comparison with the same approaches with the full test set can indicate that in emergency scenarios the annotation and description by the users of the location where the event happens is more present and accurate. As an example, the results of the accuracy at 10km and 100km for the Geofusion approach with the full test set compared with the ones of the emergency scenario test set were 39.71, 53.61 (full test set) and 52.68, 71.81 (emergency scenario test set) respectively. The results of these experiments seem to coincide with the findings of Kordopatis-Zilos et al., 2015, that emergency-related images seem to carry text metadata that is very useful for geo-location even without automatic GPS annotation.

6.7 Experiments Georeferencing Formal Documents

In order to test some of the georeferencing algorithms designed over a collection of formal documents several experiments[16] have been performed with test collections derived from the English Wikipedia. An assumption of one geolocation (georeference) per document is made. For these experiments two corpus were used: the Wikipedia UK Spot Test corpus was used as a test set, and the W&B corpus was used as a training set. The Wikipedia UK Spot Test Corpus is a set of 21,839 English Wikipedia documents that refer to a spot (railwaystation, landmarks, buildings, and schools) in the United Kingdom and it was created by Van Laere et al., 2014. The W&B Wikipedia training set[17] is a set of about 390,574 Wikipedia documents created by B. P. Wing and Baldridge, 2011. This set has been processed to filter out the Wikipedia UK Spot Test Corpus documents, resulting a total of 376,110 documents (Van Laere et al., 2014). The algorithms tested were: Geographical Knowledge (GeoKB), IR with HLM using the top-ranked predictions, IR with HLM with Re-Ranking, and GeoFusion using the IR HLM and GeoKB. The MediaEval Placing Task 2014 (MEPT2014) training data set has been used to train some IR with HLM models in some experiments. In Table 6.20 is shown the different experiments with the approaches and training corpora used in each one.

Table 6.20: Experiments to georeference formal documents from Wikipedia.

Experiment	Training Corpus
1 GeoKB	-
2 IR (HLM) Re-Rank (100km)	MEPT2014 training Set (Choi et al., 2014)
3 IR (HLM) Re-Rank (100km)	W&B corpus filtered (B. P. Wing and Baldridge, 2011)
4 IR (HLM) top-Ranked	W&B Corpus filtered (W&B)
5 GeoFusion: IR (HLM) Re-Rank (100km) +GeoKB	MEPT 2014 training set
6 GeoFusion: IR (HLM) Re-Rank (100km)+GeoKB	W&B corpus filtered (W&B)
7 GeoFusion: IR (HLM) top-ranked +GeoKB	W&B corpus filtered (W&B)

The results of the experiments are summarized in Table 6.21. The HLM (top-ranked) approach using lowercased tokens showed the best performance on 1m to 1km kms accuracies. The HLM@100km approach trained with the W&B training set with lowercased data showed the best performance on 100km and 1000 kms accuracies. On the other hand, the geofusion approach that uses the HLM top-ranked predictions (lowercased data) combined with the GeoKB obtained the best results in accuracies up to 10km. This approach also obtained the best median error (17.47 km). The geofusion approaches only improve slightly the results of accuracy at 10km when using the HLM top-ranked algorithm. This fact means that the GeoKB heuristics are not achieving enough precision to improve the Data-Driven results (HLM). Formal documents from Wikipedia are more complex and have

[16] Note that although some of the experiments presented in this Section were performed initially in 2015 and they were presented at the Text Speech and Dialogue 2015 conference (Ferrés and Rodríguez, 2015b), these experiments have been repeated because of some small differences in the training set corpus. The experiments and results explained in this Section were performed with the original training set of 376,110 documents created by Olivier Van Laere (Van Laere et al., 2014) . By contrast the experiments performed at TSD2015 (Ferrés and Rodríguez, 2015b) were performed with a total of 376,236 documents of the training set. This small difference of documents could explain also some small differences in results between these experiments.

[17] This dataset originates from the original English-language Wikipedia dump of September 4th, 2010.

Table 6.21: Percentage of correctly georeferenced documents within certain amount of kilometers for each approach and Median Error (ME) in kilometers. The best results among the experiments to georeference formal documents are marked in bold. The overall best results, including other state-of-the-art systems, are marked with light grey cells. Note that the (W&B) and (PT2014) acronyms indicate the training datasets: Wing & Baldridge Wikipedia filtered training set and Media Eval 2014 training set.

	accuracy percentage							
Experiments	1m	10m	100m	1km	10km	100km	1000km	ME km
case sensitive experiments								
GeoKB	0.00	0.02	0.40	9.36	38.21	65.15	83.77	26.58
HLM@100km (W&B)	0.29	0.33	1.22	8.56	27.09	69.50	95.70	49.22
HLM-top (W&B)	0.48	0.55	2.06	13.86	38.87	63.61	92.10	28.60
HLM@100km+GeoKB (W&B)	0.06	0.08	0.57	10.24	39.65	70.91	90.79	22.06
HLM-top+GeoKB (W&B)	0.09	0.13	0.78	11.23	42.28	70.59	90.42	18.10
lowercase experiments								
HLM@100km (PT2014)	0.01	0.19	1.70	6.80	18.74	46.52	83.10	126.37
HLM@100km (W&B)	0.29	0.33	1.21	8.64	28.48	**71.65**	**96.10**	44.69
HLM-top (W&B)	**0.49**	**0.56**	**2.13**	**14.67**	41.11	65.55	92.32	22.90
lowercased (HLM) and case sensitive (GeoKB)								
HLM@100km+GeoKB (PT2014)	0.01	0.05	0.71	9.83	37.95	66.74	89.41	27.14
HLM@100km+GeoKB (W&B)	0.05	0.08	0.58	10.25	40.00	71.42	90.80	21.57
HLM-top+GeoKB (W&B)	0.10	0.14	0.80	11.34	**42.75**	70.97	90.41	**17.47**
Comparison with state-of-the-art systems								
Roller et al. (2012) (W&B)	0.02	0.02	0.10	4.17	53.11	75.98	92.36	8.12
Van Laere et al. (2014) (W&B)	0.33	0.38	1.79	19.2	67.12	90.03	97.35	4.17
Geonames (Van Laere et al., 2014)	0.01	0.10	0.90	9.95	34.63	63.67	73.40	24.05
Placemaker (Van Laere et al., 2014)	0.00	0.03	0.27	4.14	27.57	73.48	97.80	30.17

more places compared with textual annotations and tagsets from Flicrk metadata (informal documents). For this reason the GeoKB heuristics should be improved to deal with the complexity of formal documents in order to improve both the GeoKB approach and the Geofusion approaches.

In comparison with the results reported by Van Laere et al. (2014), that used their own approaches and the approach of Roller et al. (2012) trained with the Wikipedia W&B training set corpus and tested with the Wikipedia Spot corpus, the results of the HLM Re-Ranking at 100km and the HLM with top-ranked outperformed their approaches in the ranges of 1m, 10m and 100m. The HLM with Re-Ranking approach outperformed the approach of Roller et al. (2012) in the ranges of 1km and 1,000km. The approach of Van Laere et al. (2014) trained with the W&B training set achieves the best results of all the approaches in the ranges from 1km to 100km. Regarding the comparison with the Geonames and Yahoo! Placemaker approaches described in Van Laere et al. (2014): 1) the GeoKB approach improves the performance with respect to the Geonames approach in the ranges from 10km to 1000km, 2) in comparison with the Yahoo! Placemaker the GeoKB achieves better results in the ranges of 100m to 10km.

6.8 Conclusions

This chapter presented a description and an evaluation of four generic georeferencing approaches that deal with formal and informal documents. Some of these approaches achieved state-of-the-art results in official georeferencing evaluations and posterior experiments.

These georeferencing approaches were the following ones:

1. **A Geographical Knowledge-Based Approach**. This approach uses only Geographical Knowledge for georeferencing. It uses Toponym Recognition and Geographical Focus Detection Heuristics to predict the georeferenced coordinates of the text. The Geographical Focus Detection Heuristics can use Geographical Knowledge Based Heuristics and/or Population Heuristics. In some experiments with the ME2010PT dataset (metadata of 5091 videos) the Geographical Knowledge heuristics achieved a high precision (when activated) in georeferencing (up to a 86.36%). In MEPT2014 when the GeoKB rules H1,H2, and H3 are activated (144,074 cases of 510,000), they achieve accuracy percentages of 87.37% (125,878 of 144,074 items) predicting up to 100 km. By contrast, the HLM IR model trained with the ME2014PT training set with Re-Ranking achieved a 78.34% of accuracy at 100 km when evaluated over this subset (144,074 cases). These results with informal documents were very useful for establishing high confidence rules that could allow a high precision georeferencing detection in textual annotations and tags.

2. **An Information Retrieval Approach**. This approach treats the document to georeference as an IR query and uses existing state-of-the-art IR models to retrieve the top-ranked coordinates pair as the final prediction for query. The IR models used are the TF-IDF, BM25 and Hiemstra Language Model (HLM) (Hiemstra, 2001).

3. **Information Retrieval with Re-Ranking Approach**. This approach is a variant of the IR approach that re-ranks the set of ranked predictions associated with a georeferenced document by means of a clustering process that uses a geographical distance function (Haversine distance). In some experiments after the MEPT2010 benchmark this approach achieved the best results (compared with the other approaches presented by the author) georeferencing at 1, 5 and 10 km of margin of errors distances using the Hiemstra Language Model algorithm for IR. At the MEPT2014 the HLM approach with Re-Ranking obtained the best results (compared with the other approaches presented by the author) in distance ranges from 10m to 1 km because it captures non-geographical highly descriptive and unique keywords and place names appearing in the geographical coordinates' associated metadata that are not present in the gazetteer. In comparison with the results of the other participants at MEPT2014, the IR with Re-Ranking and GeoFusion approaches achieved competitive state-of-the-art results. The HLM with Re-Ranking approach obtained the best official results for accuracies at distances of 1,000 km and 5,000 km in the task where only the official training data can be used to predict.

4. **GeoFusion approaches**. A combination of predictions from the Geographical Knowledge Heuristics H1,H2, and H2 and from the IR approaches (with and without Re-Ranking). In experiments after MEPT2010 the GeoFusion approach with Re-Ranking outperformed the best results in accuracy reported by the state-of-the art systems participating at MEPT2010, achieving a 68.53% of accuracy georeferencing up to a

distance of 100 Km. The approaches of Van Laere et al. (2010a) and Kelm et al. (2010) obtained a 67,23% and 60,46% of accuracy with the same test set at the official MEPT2010. In both official MEPT2014 evaluation and posterior experiments the GeoFusion approaches achieved the best results within the ranges of accuracies of 10km, 100km, 1,000km and 5,000km (compared with the other approaches presented by the author). Finally, for the task of Textual Georeferencing of formal documents, the GeoFusion approaches only achieved the best results predicting in the accuracies margin of 10km when trained with the W&B Wikipedia training set and without using Re-Ranking. It has to be noted that in this task the GeoFusion results are not usually outperforming the IR and IR with Re-Ranking approaches. For this reason, to improve the GeoKB results for this kind of documents is needed a more sophisticated use of the Geographical Toponym Focus Detection.

In conclusion, geographical knowledge can be useful in some cases in the Geographical Information Access task of Textual Georeferencing. In Table 7.5 are shown some improvements in accuracy of the algorithms that use geographical knowledge combined with data-driven techniques (the approaches HLM@100km and HLM@100km+GeoKB) over the ones that use only data-driven techniques (HLM top-ranked) in different tasks and contexts. The HLM@100km uses the Haversine distance to re-rank IR predictions and the HLM@100km+GeoKB uses both the Haversine distance for re-ranking and the GeoKB H1,H2,H3 heuristics with topological and population knowledge to predict when activated. The use of the GeoKB rules can be useful in some cases to establish high precision rules that could be choosen with priority over data-driven techniques. The use of these heuristics can also help the systems to give a confidence to the predictions.

	TG (informal documents)			TG (formal documents) (case sensitive experiments)		
Training Corpus	MEPT2014 Training			W&B filtered		
Test Corpus	MEPT2014 Test			Wikipedia Spot		
Baseline	HLM top-ranked					
Approach	HLM@100km					
Accuracy Measure	Acc@1km	Acc@10km	Acc@100km	Acc@1km	Acc@10km	Acc@100km
Δ Improvement(%)	-17,58%	-7,61%	+0.17%	-38.23%	-30,30%	+9.25%
Baseline	HLM top-ranked					
Approach	HLM@100km+GeoKB					
Accuracy Measure	Acc@1km	Acc@10km	Acc@100km	Acc@1km	Acc@10km	Acc@100km
Δ Improvement(%)	-16.19%	+6.83%	+15.68%	-26,11%	+2.00%	+11.47%
Baseline	HLM@100km					
Approach	HLM@100km+GeoKB					
Accuracy Measure	Acc@1km	Acc@10km	Acc@100km	Acc@1km	Acc@10km	Acc@100km
Δ Improvement(%)	+1.69%	+15.63%	+4.99%	+19.62%	+46.36%	+2.02%

Figure 6.13: Examples of experiments that show improvements in some evaluation measures (in grey background) with respect of the Data-Driven models using the combination of Geographical Knowledge and Data-Driven techniques applied to Textual Georeferencing.

Moreover, this Chapter presents a set of preliminary experiments with emergency scenarios derived from Flickr social network data. The results of these experiments show that Textual Georeferecing could be useful for geo-location prediction within the context of these events. The results also show that emergency-related images seem to carry text metadata that could be potentially useful for geo-location with or without automatic GPS annotation.

CHAPTER 7

Conclusions

This chapter outlines and describes the main contributions of this thesis, points out its limitations, and proposes further work.

7.1 Contributions

This PhD thesis contributes to the state-of-the-art of Geographical Information Access with these main contributions:

- **The presentation and description of several novel approaches for Geographical Information Access tasks.** These approaches deal with the following tasks related to Geographical Information Access: Geographical Information Retrieval, Geographical Question Answering and Textual Georeferencing.
- **The evaluation of these novel approaches for Geographical Information Access tasks.** These approaches have been evaluated in these contexts: 1) within official evaluation benchmarks, 2) after evaluation benchmarks with the benchmarks test collections, 3) with other specific datasets.
- **The effective use of Geographical Knowledge and Natural Language Processing for the Geographical Information Retrieval tasks evaluated**.
- ***Passage Retrieval Approaches for GIR***. Implementation and evaluation of two approaches that combine sucessfully Geographical Knowledge and Passage Retrieval presented at GeoCLEF 2005 and GeoCLEF 2006.
- ***TALPGeoIR***. An approach that combines Geographical Knowledge Re-Ranking, Natural Language Processing and Relevance Feedback for Geographical Information Retrieval that achieved state-of-the-art results in official GeoCLEF benchmarks (Ferrés and Rodríguez, 2008a; Mandl et al., 2008) and posterior experiments (Ferrés and Rodríguez, 2015a).

- ***GeoTALP-QA. A scope-based Geographical Question Answering Approach***. This thesis contributed to both GeoQA and Restricted-Domain QA state-of-the-art with the design and implementation of a Scope-based GeoQA system for Spanish and English and its evaluation with a set of questions of the Spanish geography (Ferrés and Rodríguez, 2006a).
- ***State-of-the-art Textual Georeferencing approaches***. This thesis presented four approaches to generic Textual Georeferencing for informal and formal documents that achieved state-of-the-art results in evaluation benchmarks (Ferrés and Rodríguez, 2014) and posterior experiments (Ferrés and Rodríguez, 2011a; Ferrés and Rodríguez, 2015b).
- ***A Geographical Query Parsing algorithm***. A Geographical Query Parsing algorithm that detects and extracts information from geographical queries that has been evaluated with search engine log queries in an evaluation benchmark (Ferrés and Rodríguez, 2008b).
- ***GikiTALP: a simple Data-Driven baseline for Geographical Question Answering over Wikipedia.***

7.1.1 Design of novel Geographical Information Access Approaches

This thesis presents 10 different approaches to the three GIR tasks investigated in this thesis: 3 approaches for GIR, 3 for GeoQA, and 4 for Textual Georeferencing (TG) (see Table 7.1). Regarding the novelty of these approaches: 9 of these approaches are novel and complex, and 1 of them (the GikiTALP 2009) is simple and not novel but its application in some experiments is a novelty.

Task	Approach	Type of Corpus in the Experiments	Language
GIR	GeoTALP-IR 2005 TALP-GeoIR 2006 TALP-GeoIR 2007	journalistic news	English
Geo QA	GeoTALP-QA	Corpus of Spanish Geography & Web snippets	Spanish
	GeoQuery 2007	MSN search logs	English & Spanish
	GikiTALP 2009	Wikipedia	English
TG	GeoKB IR-Approach IR-ReRanking GeoFusion	Flickr Metadata (informal docs) Wikipedia (formal docs)	multilingual English

Table 7.1: Details of the main characteristics of the approaches for Geographical Information Access tasks. Please consult Chapters 4,5,and 6 for more specific details about these systems.

7.1.2 Evaluation of Geographical Information Access Approaches

All the approaches have been evaluated for the tasks of Geographical Information Retrieval, Geographical Question Answering and Textual Georeferencing. Most of these algorithms have been presented in international benchmarking evaluations (see Table 7.2) and some of them achieved state-of-the-art results (including the top-performing results in GIR (GeoCLEF 2007) and TG (MediaEval 2014) benchmarks). The other experiments performed at official evaluation benchmarks presented in this thesis achieved average or low performance compared with the other participants in the benchmarks but these experiments have shown scientific relevance by: a) showing that Geographical Knowlegdge and Natural Language Processing combined with Data-Driven methods can improve effectiveness measures of GeoIA tasks, or b) establishing baselines for future improvements in the task. As an example, the GIR task results obtained at GeoCLEF 2005 and GeoCLEF 2006 were average and low but the experiments itself showed that the Geographical Knowledge used was helping to improve the Data-Driven baseline.

Task	Evaluation	Comparative Ranking	Experiment Type
GIR	GeoCLEF 2005	average (7st of 11)	Data-Driven+GeoKB
	GeoCLEF 2006	low (15st of 16)	Data-Driven+GeoKB
	GeoCLEF 2007	**top-ranked (1st of 11)**	Data-Driven+GeoKB
GeoQA (Wikipedia)	GikiCLEF 2009	low (6th of 8)[1]	Data-Driven (baseline)
GeoQA (Query Parsing)	GeoQuery 2007	average (3rd of 6)	GeoKB (baseline)
Text Georeferencing	MediaEval 2010	average	GeoKB (baseline)
	MediaEval 2011	average (3rd of 6)[2]	Data-Driven+GeoKB
	MediaEval 2014	**top (1st of 6)** [3]	Data-Driven+GeoKB

Table 7.2: Summary of participation in international evaluations of Geographical Information Access tasks

Moreover this thesis describes two other types of experiments: 1) experiments realized after the evaluation benchmarks in order to improve the approaches, and 2) experiments with closed collections (see a list of these two types of experiments in Table 7.3).

Task	Corpora	Experiment Type
GIR	**GeoCLEF 2006** (Post-eval)	Data-Driven+GeoKB
GIR	**GeoCLEF 2005-2008** (Ferrés and Rodríguez, 2015a)	Data-Driven+GeoKB
GeoQA (RDQA)	**Web snippets** (Ferrés and Rodríguez, 2006a)	Data-Driven+GeoKB
Textual Georeferencing	**Wikipedia** (Roller et al., 2012; Van Laere et al., 2010b)	Data-Driven+GeoKB
	MEPT2010 (Post-Eval) (Ferrés and Rodríguez, 2011a)	DataDriven+GeoKB
	MEPT2014 (Post-Eval) (Ferrés and Rodríguez, 2015b)	DataDriven+GeoKB
	Emergency data (MEPT2014) (Kordopatis-Zilos et al., 2015)	DataDriven+GeoKB

Table 7.3: Summary of experiments without participation in international evaluations of Geographical Information Access tasks. Note that in bold are marked the experiments that outperformed (partially or totally) the results of the participants in the evaluation or in the dataset.

[1] GikiCLEF 2009 Final Score results.

[2] Ranked 3st in the accuracy ranges from 10km to 10,000Km

[3] Ranked 1st in the accuracies distances of 1,000 and 5,000 kms, and 2st in accuracy distance of 100 kms

Most of the approaches that have been evaluated after the “post-evaluation” improvements showed that in several contexts can outperform the results of the author’s baseline algorithms presented in the benchmark and in some cases (GeoCLEF 2005, GeoCLEF 2006, GeoCLEF 2007, MediaEval 2010, and MediaEval 2014) even outperform the best official results obtained at the evaluation benchmark. On the other hand, some experiments in GIR, GeoQA and Textual Georeferencing have been performed in closed collections or the web without participation in international evaluations. These experiments also showed that in come contexts Geographical Knowledge in combination with Data-Driven methods can outperform the Data-Driven baselines.

This thesis also replicated partially the experiments of Kordopatis-Zilos et al. (2015) (see Chapter 6), which consisted in evaluate Textual Georeferencing over an ”emergency scenario” test set derived from the MEPT2014 test set (Choi et al., 2014). The results of these experiments show that emergency-related images seem to carry text metadata that is very useful for geo-location even without automatic GPS annotation.

7.1.3 Effective Use of Geographical Knowledge and Natural Language Processing in Geographical Information Access tasks

The approaches presented in this PhD thesis use the following sets of techniques alone or in a combination: 1) *Data-Driven:* techniques based on Information Retrival, 2) *Geographical Knowledge:* including Geographical Knowledge-Bases, Toponym Recognition and Disambiguation methods, and 3) *Natural Language Processing techniques*: stemming, lemmatization, stopwords filtering, NERC, parsing and Wordnet (see in Table 7.4 the resources employed for each approach).

		Data Driven	Geographical Knowledge		Natural Language Processing				
Task	Approach	IR Engine	GeoKB	Disambiguation	stem.	lemma.	stopwords	NERC	parsing& WordNet
GIR	GeoTALPIR 2005	✓	✓	-	✓	✓	✓	✓	-
	TALPGeoIR 2006	✓	✓	-	✓	✓	✓	✓	-
	TALPGeoIR 2007	✓	✓	✓	✓	✓	✓	✓	-
Geo QA	GeoTALP-QA	✓	✓	-	✓		✓	✓	✓
	GeoQuery 2007	-	✓	✓	-	✓	-	✓	✓
	GikiTALP 2009	✓	-	-	✓	-	✓	-	-
TG	GeoKB	-	✓	✓	-	-	✓	-	-
	IR-Approach	✓	-	-	-	-	✓	-	-
	IR-ReRanking	✓	✓	-	-	-	✓	-	-
	GeoFusion	✓	✓	✓	-	-	✓	-	-

Table 7.4: Details of the main characteristics of the approaches for Geographical Information Access tasks. Please consult Chapters 4,5,and 6 for more specific details about these systems.

The experiments reported in this PhD thesis show that the approaches can combine effectively geographical knowledge and natural language processing tools with data driven techniques (see in Table 7.5 by improving efectiveness measures in different cases of the three Geographical Information Access tasks investigated).

Task	Approach	Corpus	Effectiveness Measure	improvement Δ (%)
GIR	TF-IDF + GeoKR	GeoCLEF all	MAP (title)	+7.74%**
GeoQA (P. Retrieval)	GeoKB (QE)	Google (snippets)	accuracy	+7.84%
TG (informal)	HLM@100km+GeoKB	MEPT2014	acc@100km	+4.99%
TG (formal)	HLM+GeoKB	W&B/Wikipedia Spot	acc@10km	+1.98%

Table 7.5: Examples of some cases of improvements of effectiveness measures, with respect to the Data-Driven baselines, of the combination of Geographical Knowledge and Data-Driven techniques in the context of the different Geographical Information Access tasks presented in the thesis. The GIR approach has been evaluated with statistical significance and showed statistical significance for t-test and randomization tests with p-values <0.01.

The author employed in all the tasks toponym recognition methods (NLP and gazetteer lookup) to detect toponyms in text and other geographical information derived from Geographical Knowledge Bases or gazetteers.

For the Geographical Information Retrieval task the geographical information used is the population count, feature type, and geographical salience of the toponyms extracted. In GIR, three approaches (GeoTALP-IR, TALP-GeoIR2006 and TALPGeoIR 2007) evaluated at different GeoCLEF official benchmarks and posterior experiments used successfully Geographical Knowledge to improve the results of state-of-the-art IR algorithms by using a Geographical Re-Ranking process. The third approach, TALPGeoIR 2007, was evaluated with statistical significance testing with the full GeoCLEF corpus (100 topics) and showed that Geographical Knowledge Re-Ranking has statistical significance over the state-of-the-art IR algorithms TF-IDF, BM25 and InL2 in effectiveness measures such as MAP (in all cases with p-values < 0.01) and in R-Precision (in some cases).

In the QA task, the author used also feature types and an ontology of geographical classes that include geographical , political, numeric and magnitude elements. In GeoQA, the use of Query Expansion based on Geographical Knowledge improved the results in *answer accuracy* of the Passage Retrieval module that used the google search engine to get relevant snippets in some experiments.

Finally, the Textual Georeferencing algorithms use the following type of geographical knowledge: 1) geographical common-sense heuristics that use the population count, feature type, and 2) haversine distance between Geographical places (detected by their toponyms). In the task of Textual Georeferencing for informal documents, the GeoFusion approaches achieved the best results in the experiments at ranges from 10 km to 5,000 km with the ME2014PT Training dataset at the official ME2014PT benchmark, clearly outperforming our other approaches (Ferrés and Rodríguez, 2014; Ferrés and Rodríguez, 2015b) . The GeoFusion approaches achieved the best results at these evaluation ranges because these approaches combine high precision rules based on Toponym Disambiguation heuristics and predictions that come from an IR model when these rules are not activated. An evaluation of the Geographical Knowledge-Based heuristics showed that when these rules are activated (144,074 cases of 510,000), they achieve accuracy percentages of 87.37% (125,878 of

144,074 items) predicting up to 100 km. By contrast, the HLM IR model trained with the ME2014PT training set with Re-Ranking achieved a 78.34% of accuracy at 100 km when evaluated over this subset (144,074 cases).

For the task of Textual Georeferencing of formal documents, the GeoFusion approaches only achieved the best results predicting in the accuracies margin of 10km when trained with the WB Wikipedia training set and without using Re-Ranking. It has to be noted that in this task the geofusion results are not usually outperforming the IR and IR with Re-Ranking approaches. For this reason, to improve the GeoKB results for this kind of documents is needed a more sophisticated use of the Geograhical Toponym Focus Detection.

The proposal, implementation and evaluation of approaches that combine successfully Geographical Knowledge Heuristics based on human common-sense with Data-Driven methods in most of the Geographical Information Access Tasks is one of the most significant contributions of this thesis. The main idea of these approaches is that Knowledge-Based Heuristics based on human common-sense combined with Data-Driven approaches can improve the results of Data-Driven methods.

7.1.4 Passage Retrieval Approaches for GIR

The initial GIR approaches by this author's thesis were those presented at GeoCLEF 2005 and GeoCLEF2006. The approach presented by the author at GeoCLEF 2005 achieved its best MAP with 0.2231, which is clearly low respect with the best MAP presented (0.3936). The author presented 4 runs (experiments) at GeoCLEF 2005 and the two experiments that used geographical information outperformed the author's baselines, based on Passage Retrieval. Regarding the official results at GeoCLEF 2006, these ones shows that the system failed to use the Geographical Knowledge to improve the IR results. But this fact was due to that this approach used the non-geographical keywords for the textual IR baselines instead of using all the keywords as in the other two approaches. The experiments with the same system after GeoCLEF 2006 have proved this fact, and showed that when using all the keywords for the textual approach the Geographical Re-Ranking process outperforms the baselines;

In both evaluations the approaches used Passsage Retrieval systems for QA as baselines; being ones of the few attempts to try Passage Retrieval techniques for GIR. In GeoCLEF 2005 the baseline was a Passage Retrieval for QA applied in the GIR context, and for GeoCLEF 2006 the baseline was the JIRS passage retrieval system. In both cases the MAP measure obtained with these baselines was low compared with the MAP of the best systems presented (0.1923 and 0.1342 of MAP measure respectively in GeoCLEF2005 and GeoCLEF2006). The other Passage Retrieval system tested by other researches for GIR was IR-n (Pascual, 2002): Ferrández et al. (2005) and Toral et al. (2007) used it in GeoCLEF2005 and GeoCLEF2006 with MAP values of 32.53 and 0.2985 respectively and García-Vega et al. (2006b) used IR-n but combined with other IR engines in GeoCLEF2006.

7.1.5 TALPGeoIR, a State-of-the-art GIR Approach

This thesis proposed an approach that combines Geographical Knowledge, Natural Language Processing and Relevance Feedback for Geographical Information Retrieval that obtained state-of-the-art results in the GeoCLEF 2007 official benchmark (Ferrés and Rodríguez, 2008a) and posterior experiments with the GeoCLEF 2005, 2006, 2007, and 2008 English test collections in 2015 (Ferrés and Rodríguez, 2015a). This approach, the TALPGeoIR 2007 approach, achieved the top-ranked results at the official GeoCLEF 2007 evaluation (see Table 7.6). Four of the five runs were ranked as the first four runs in the GeoCLEF 2007 evaluation task (consult Mandl et al. (2007) for more details) both considering Mean Average Precision (ranging from 28.50% to 27.11%, next system was scored 26.42%) and R-Precision (ranging from 31.70% to 28.47%, next system was scored 27.23%).

The reason for these competitive results are due to the use of a combination of Geographical Knowledge Re-Ranking, Linguistics Processing adn Relevance Feedback. Evaluated separately and in combination each one of these methods has improved the MAP and R-Precision showing statistical significance with respect of the standard IR baselines TF-IDF, BM25 and InL2 in most of the experiments (Ferrés and Rodríguez, 2015a).

Approach		Best MAP
TALP-U.Politècnica Catalunya	(Ferrés and Rodríguez, 2007b)	0.2850
U.C. Berkeley	(R. R. Larson, 2007)	0.2642
U.Politècnica Valencia	(Buscaldi and Rosso, 2007)	0.2636
U. Groningen	(Andogah and Bouma, 2007)	0.2515
Cal State U.- San marcos	(Guillén, 2007)	0.2132
U.Lisbon	(Cardoso et al., 2007)	0.2180
ICL	(S. E. Overell et al., 2007)	0.1850
Moscow State Univ.	-	0.1761
linguit Ltd	(Leidner, n.d.)	0.1612
U.Hildesheim	(Kölle et al., 2007)	0.1535
Microsoft Asia	(Z. Li et al., 2007a)	0.1519

Table 7.6: GIR approaches in the context of the official GeoCLEF 2007 evaluation ordered by MAP. The results of the TALPGeoIR approach are colored with light grey.

In addition, some configurations with Geographical Knowledge Re-Ranking, Linguistic Processing and Query Expansion have improved the MAP of the best official results at GeoCLEF evaluations of 2005, 2006, and 2007 with improvements of 0.9%, 11.73%, and 3.05% of MAP respectively and without statistical significance with p-values < 0.05 detected (Ferrés and Rodríguez, 2015a) (See Table 7.7).

Table 7.7: MAP at 1,000 documents with the best configurations for the full collection applied to each GeoCLEF Monolingual English task. Includes the best official results (in MAP) at GeoCLEF evaluations.

Approach	MAP with GeoCLEF test collection (year)			
	2005	2006	2007	2008
median of official runs	0.2600	0.2700	0.2097	0.2370
best official results	0.3936[4]	0.3034[5]	0.2850[6]	**0.3037**[7]
InL2+S+L+GeoKR+Bo1(TDN)	0.3921	0.3303	**0.2937**	0.2208
InL2+S+L+GeoKR+KL(TDN)	**0.3974**	**0.3390**	**0.2924**	0.2178

Moreover this is the first system that combines NLP (lemmatization level), Geographical Knowledge Re-Ranking and Relevance Feedback and has been tested exhaustively with all the GeoCLEF topics (100 topics from the evaluation benchmarks of GeoCLEF2005, GeoCLEF2006, GeoCLEF2007,and GeoCLE2008). The testing was done with the TF-IDF, BM25 and InL2 baselines. All the improvements were tested separately and as a whole and with the three usual combinations of the metadata topic fields (Topic(T), Description(D) and Narrative(N): T, TD, and TDN.

7.1.6 GeoTALP-QA. An Scope-based Geographical Question Answering Approach

This thesis contributed to the Geographical Question Answering (GeoQA) state-of-the-art with the design and implementation of a scope-based textual GeoQA system for Spanish and English and its evaluation with a set of questions of the Spanish geography (Ferrés and Rodríguez, 2006a). The GeoTALP-QA approach has been adapted from an existing Open-Domain Question Answering system with geographically oriented resources and specific domain adaptation (special ontology and question types). This thesis describes all the geographical resources and domain adaptation needs to do this process. GeoTALP-QA (Ferrés and Rodríguez, 2006a) was one of the first approaches to propose a scope-based approach to deal with geographical questions regarding the geography of a country or a region and semi-automatically obtain resources (gazetteers, geographical feature types, groups of place names) related with a geographical region and a language.

7.1.7 New state-of-the-art Textual Georeferencing Approaches

An important contribution of this dissertation is the presentation and evaluation of four generic georeferencing approaches based on Geographical Knowledge Bases, Linguistic Processing, and Information Retrieval to deal with formal and informal documents. These approaches assume the "one geographical referent per discourse" hypothesis and some of them achieved competitive state-of-the-art results in official georeferencing evaluations and posterior experiments. These georeferencing approaches are the following ones:

[4] (R. Larson et al., 2006)

[5] (Martins et al., 2007b)

[6] (Ferrés and Rodríguez, 2008a)

[7] (R. Wang and Neumann, 2009)

1. **A Geographical Knowledge-Based Approach**. This novel approach uses only Geographical Knowledge for georeferencing. It uses Toponym Recognition and Geographical Focus Detection Heuristics to predict the georeferenced coordinates of an input text. The Toponym Recognition approach uses the Geonames gazetteer. Geonames has been used in GIR (Toral et al., 2006), Geographical Scope Resolution (Andogah et al., 2008) and georeferencing approaches (Serdyukov et al., 2009). The information contained in this gazetteer allows the approach to deal with the recognition of issues related with multilinguality, acronyms, lower and uppercase place names, and joined place names. Thus this gazetteer provides a robust Knowledge-Based method to solve these issues. The Geographical Focus Detection Heuristics use both Geographical Knowledge Based Heuristics and Population Heuristics described in the Toponym Disambiguation literature Hauptmann et al. (1999) and Leidner (2007). A main difference from the text and gazetteer based approaches systems of Kelm et al. (2010) and J. Perea-Ortega et al. (2010) with respect to our system is that the approach presented in this thesis does not use Named Entity Recognizers and NLP processors. The approach presented here only uses the Geonames Gazetteer and some limited NLP resources that are stopwords lists from several languages and an English dictionary to treat the geo/non-geo ambiguity of some toponyms.

   Regarding the results obtained by this approadch in evaluation benchmarks: in some experiments with the ME2010PT dataset (metadata of 5,091 videos) the Geographical Knowledge heuristics achieved high precision (when activated) in georeferencing (up to a 86.36%). In MEPT2014 when the GeoKB rules are activated (144,074 cases of 510,000), they achieved accuracy percentages of 87.37% (125,878 of 144,074 items) predicting up to 100 km. These results were very useful for establishing high confidence rules that could allow a high precision georeferencing detection in textual annotations and tags.

2. **An Information Retrieval Approach**. This approach treats the document to georeference as an IR query and uses existing state-of-the-art IR models to retrieve a set of weighted coordinates relevant to the query. The IR models used are the TF-IDF, BM25 and Hiemstra Language Model (Hiemstra, 2001).

   The novelty of this algorithm is the use of IR techniques for indexing and retrieval of relevant geographical coordinates as a way of georeferencing texts. The main novelty of the indexing approach is that instead of using grid-based indexing (or language modelling) approaches Serdyukov et al. (2009), Van Laere et al. (2013), Kordopatis-Zilos et al. (2014), and Popescu et al. (2014) each different coordinates pair in the original collection will generate an associated document with all the metadata related with this specific coordinate pair. This means that instead modelling regions this algorithm is modelling unique individual geographical coordinates. The IR weighting models used are the TF-IDF, BM25 and Hiemstra Language Model (HLM) (Hiemstra, 2001) implemented in the Terrier IR system[8] Ounis et al., 2006. To the best of the author's knowledge, these are the first approach and experiments with the HLM algorithm in Textual Georeferencing. Some experiments with informal documents also showed that the HLM algorithm outperforms TFIDF and BM25 in some cases (see Chapter 6).

[8] Terrier IR engine. `http://www.terrier.org`

3. **Information Retrieval with Re-Ranking Approach**. This approach is a variant of the IR approach that re-ranks the set of ranked predictions associated with a georeferenced document by means of a clustering process that uses a geographical distance function (Haversine distance). It achieved state of the art results (and some of the top results in some accuracy ranges) comparing with other systems at the MediaEval 2010, 2011 and 2014 international benchmarks. The novelty of the IR with Re-Ranking approach presented in this thesis is that it models individual coordinates (instead of grid cells or regions) in the first stage and then in a posterior the Re-Ranking phase a clustering process based on the Haversine distance re-ranks the top retrieved coordinates.

   In some experiments after the MEPT2010 benchmark this approach achieved our best results georeferencing at 1, 5 and 10 km of margin of errors distances using the HLM IR model. At the MEPT2014 the HLM approach with Re-Ranking obtained the best results (compared with the other approaches presented by the author) in distance ranges from 10m to 1 km because it captures non-geographical highly descriptive and unique keywords and place names appearing in the geographical coordinates' associated metadata that are not present in the gazetteer. In comparison with the results of the other participants at MEPT2014, the IR with Re-Ranking and GeoFusion approaches achieved competitive state-of-the-art results (Ferrés and Rodríguez, 2015b). The HLM with Re-Ranking approach obtained the best official results for accuracies at distances of 1,000 km and 5,000 km in the task where only the official training data can be used to predict (See Table 7.8).

4. **GeoFusion Approaches**. GeoFusion is a novel state-of-the-art approach that combines predictions from the Geographical Knowledge Heuristics and from the IR approaches (with and without Re-Ranking). This is a unique approach that combines heuristics based on common-sense (indicating a high confidence when one of this heuristic is activated) and uses Data-Driven (IR approaches) predictions (activated when the heuristics were not activated).

   In experiments after MEPT2010 the GeoFusion approach with Re-Ranking outperformed the best results in accuracy reported by the state-of-the art systems participating at MediaEval 2010 Placing task, achieving an 68.53% of accuracy georeferencing up to a distance of 100 Km (Ferrés and Rodríguez, 2010b). The approaches of Van Laere et al. (2010a) and Kelm et al. (2010) obtained a 67,23% and 60,46% of accuracy with the same test set at the official MediaEval 2010 placing task. In both official MEPT2014 evaluation and posterior experiments the GeoFusion approaches achieved the best results within the ranges of accuracies of 10km, 100km, 1,000km and 5,000km compared with the other approaches presented by the author.

   Regarding the official experiments with the ME2014PT dataset and posterior experiments with gazetteer use (see the results in Table 7.8): the official run1 at the benchmark was performed with the HLM model and a distance threshold of 100km for Re-Ranking and it achieved the best official results in accuracies at distances of 1,000km and 5,000km.

   For the task of Textual Georeferencing of formal documents, the GeoFusion approaches only achieved the best results predicting in the accuracies margin of 10km when trained with the W&B Wikipedia training set and without using Re-Ranking.

Table 7.8: Results of Run1 at ME2014PT (use provided training dataset only) and posterior experiments (with and without gazetteers used).[10]

		accuracy percentage						
	System	10m	100m	1km	10km	100km	1000km	5000km
Benchmark Results	CEALISTPopescu, et al, 2014	0.01	0.61	22.62	40.00	47.36	61.17	74.94
	RECOD(L. Li et al., 2014)	0.55	**6.06**	21.04	37.59	46.14	61.69	76.76
	SonSensCERTHKordopatis et al 2014	0.50	5.85	23.02	39.92	46.87	60.11	74.80
	UQ-DKE(Cao et al., 2014)	**1.07**	4.98	19.57	**41.71**	**52.46**	63.61	77.28
	USEMPPopescu, et al, 2014	0.78	1.61	**23.48**	40.77	48.11	61.79	75.30
	ICSI/TUDelft (Choi and X. Li, 2014)	0.24	3.15	16.65	34.70	45.58	60.67	75.03
	TALP-UPC[11](Ferrés, et al, 2014)	0.29	4.12	16.54	34.34	51.06	**64.67**	**78.63**
Post-Evaluation Experiments	GeoKB	0.07	0.89	11.31	34.44	42.26	48.45	58.32
	training with the MEPT2014 training dataset							
	HLM@10km	0.29	4.18	17.35	**41.99**	50.97	63.38	77.91
	HLM@1km	0.30	4.65	**24.03**	41.10	49.53	62.20	75.79
	HLM@0.1km	0.46	**7.20**	22.29	38.37	46.86	60.10	74.59
	TFIDF@100km	0.29	4.21	16.84	34.32	50.15	63.52	77.69
	BM25@100km	0.29	4.24	17.01	34.63	50.60	63.88	77.93
	HLM@100km+GeoKB	0.25	3.25	16.82	39.71	**53.61**	**66.78**	**80.06**
	HLM@10km+GeoKB	0.26	3.32	17.30	**43.48**	**53.47**	**65.67**	**79.47**
	HLM@1km+GeoKB	0.25	3.56	20.74	**42.80**	52.36	**64.76**	77.48
	HLM@0.1km+GeoKB	0.35	5.03	19.69	40.95	50.53	63.22	76.58
	TFIDF@100km+GeoKB	0.25	3.19	16.72	39.34	**53.07**	**66.10**	**79.39**
	BM25@100km+GeoKB	0.25	3.21	16.83	39.53	**53.31**	**66.30**	**79.52**
	HLM@100km+GeoKB+Wiki	0.25	3.25	16.82	39.72	**53.61**	**66.77**	**80.05**

7.1.8 A Geographical Query Parsing algorithm

A Geographical Query Parsing algorithm that detects and extract information from geographical queries that was evaluated with search engine log queries at the GeoQuery 2007 evaluation benchmark (Ferrés and Rodríguez, 2008b). This approach is based on a linguistic and geographical knowledge analysis of the queries. This system was ranked the third from 6 participants. The analysis of the results showed that the selection of a subset of the most important features to create a gazetteer of only the most important places implies a lost of coverage and thus missing geographical places and classifying queries as non-local.

7.1.9 GikiTALP: a Data-Driven Baseline for Geographical Question Answering over Wikipedia

This approach uses a Data-Driven algorithm with limited Natural Language Processing and without Geographical Knowledge. The use of the Sphinx seardch engine, with limited NLP and without Geographical Knowledge to the with GeoQA. This approach can be seen as a baseline for future improvements on this task.

[11] Note that the experiments with that include the "GeoKB" string only can be compared with the TALP-UPC approach (author's approach) because the other approaches in the benchmark only used the provided dataset for training.

[11] HLM@100km approach.

7.2 Limitations and Future work

This section discusses the limitations of the work presented in this thesis and describes some possible further research in the context of the three Geographical Information Access tasks investigated.

7.2.1 Geographical Information Retrieval

The GIR approaches presented in this thesis were evaluated with corpora from important newspapers in English. An evaluation of the algorithms with other kind of documents, other languages, and more locally oriented documents is necessary to know the bounds of the approach. It is expected that the approach will not be as good in local news due to the Toponym Disambiguation strategies based on salience and population counts. The Toponym Recognition and Disambiguation strategies performed with the TALPGeoIR approach were context independent and were using heuristics that gave more importance to some places (with more political salience or population count) and some geographical feature types. Although this method has proven to be useful to improve the effectivenes results of the GIR GeoCLEF English test collections in MAP and R-Precision, context aware methods that could disambiguate toponyms taking into account semantic clues in the text and inter-relationships with other toponyms in text will be expected to improve the results of both general and locally oriented documents. Regarding the evaluation with other languages, it is expected that the improvements shown in GIR still apply but it will be necessary to study the coverage of the name variants in different languages in the gazetteers employed to build the Geographical Knowledge Base and perhaps use a database such as Geonames to improve this coverage. For these reasons further work can include: 1) a change the NLP and NERC phases for a simple Geonames gazetteer lookup of tokens and evaluate the performance this method, 2) apply context aware Toponym Disambiguation heuristics that could disambiguate fully or partially (reducing the degree or geographical ambiguity) toponyms in the collection and the topics, 3) use Data-Driven methods for Toponym Disambiguation trained with data from Flickr or Wikipedia to help with difficult cases. 4) evaluate the approach with other existing test collections for other languages at GeoCLEF evaluations such as Spanish, Portuguese, and German. 5) perform more specific analysis of the heuristics,

Finally, it should be pointed out that the use of only the spatial operator and part-of-relationship ”in” as the default way of treat the geographical part of queries instead of a more detailed analysis (that could treat spatial relationships such as “near”, “close”, “north of”) means that there is also more room for improvement.

7.2.2 Geographical Question Answering

The Geographical Question Answering approach presented in this thesis, GeoTALP-QA, is a language and scope-based approach, meaning that it has to be adapted to a certain geographical scope region and a language. The Geographical Knowledge resources needed for this adaptation can be obtained from existing geographical Gazetteers with semi-automatic methods. The approach can solve geographical questions about this restricted region.

The evaluation of GeoTALP-QA was done in the scope of the Spanish Geography with Spanish as a source and target language and with google search as a passage retrieval method. A set of snippets extracted from google search were used for the evaluation of the Knowledge-Based and the Data-Driven Answer Extraction operational modes. But extracting the answer from snippets with a Knowledge-Based Answer Extraction module that applies Language Processing and some reasoning have some difficulties because the snippets are sometimes cut and sentences are broken. A new evaluation of good quality snippets or passages of full documents will be needed to test the Knowledge-Based Answer Extraction in a more fair environment for its requirements. As a future work the Knowledge-Based AE module could also be improved with the treatment of questions that need the extraction of the answer from two or more sentences (e.g. treating coreference).

7.2.3 Textual Georeferencing

The TG approaches presented in this thesis have the limitations of the “one georeference per text” that is assumed in this thesis. Future research can investigate to adapt the approaches for documents that can refer to several locations. Further work can also explore: 1) research about high precision Geographical Focus Detection heuristics for formal documents, 2) new experiments to combine other geo/non geo ambiguity approaches based on spatial probability distribution models (Hauff and Houben, 2011) with the presented approaches. 3) experiments with context-aware Toponym Disambiguation Heuristics that can be adapted to different kind of contexts; for these experiments the TAC-KBP Entity Linking[12] datasets that include newswire, web texts, and discussion fora could be used.

[12] `http://nlp.cs.rpi.edu/kbp/2014/`

APPENDIX A

Annex A: Test Collections

This Annex describes how to obtain the test collections employed in the experiments presented in this thesis.

A.1 Test Collections for Geographical Information Retrieval

The tests collections employed in this thesis for GIR are the GeoCLEF test collections. The GeoCLEF test topics, relevance assesments and the official experiments performed at GeoCLEF from 2005 to 2008 can be downloaded at `http://direct.dei.unipd.it/`. The collection of documents to perform the experiments are The Glasgow Herald (1995) (GH95) and Los Angeles Times (1994) (LAT94) collections (i.e. 169,477 documents). These two collections can be obtained through the ”The CLEF AdHoc-News Test Suites (2004-2008) Evaluation Package” or the ”CLEF Question Answering Test Suites (2003-2008) – Evaluation Package” corpora. These corpora are available at `http://catalog.elra.info`.

A.2 Test Collections for Geographical Question Answering

A.2.1 Geo-QA corpus: a set of 123 Geographical QA questions

A corpus of Geographical questions was obtained from Albayzin, a speech corpus (Diaz et al., 1998) that contains a geographical subcorpus with utterances of questions about the geography of Spain in Spanish. The 123 questions with answer in the SW were divided in two sets: 61 questions for development (setting thresholds and other parameters) and 62 for test. The test set questions are listed in the Annex D - Geo-QA Questions.

A.2.2 GeoQuery 2007 Dataset

The Geoquery 2007 Geographic Query Parsing Data Set (Microsoft) can be downloaded from the following site: `http://research.microsoft.com/en-us/people/xingx/geoquery.aspx`

A.2.3 GikiCLEF 2009 test collections

The GikiCLEF 2009 Test collections can be downloaded from the following site: `http://www.linguateca.pt/GikiCLEF/GIRA/`

A.3 Test Collections for Textual Georeferencing

A.3.1 MediaEval Placing Tasks of 2010 and 2011 Training and Test set.

The datasets of the MediaEval Placing Tasks of 2010 and 2011 can be downloaded from this site: `https://github.com/dferres/placingtask_2010-2011_textual_metadata`

A.3.2 MediaEval Placing Task 2014 Training and Test set.

The training data consists of 5 million geotagged photos and 25,000 geotagged videos (downloadable at `https://multimedia-commons.s3-us-west-2.amazonaws.com/subsets/YLI-GEO/docs/mediaeval2014_placing_train.bz2`) , whereas the test set consists of 500,000 photos and 10,000 videos (downloadable at `https://multimedia-commons.s3-us-west-2.amazonaws.com/subsets/YLI-GEO/docs/mediaeval2014_placing_test.bz2`). (see Choi et al. (2014) for more details about this evaluation and dataset). The training and the test set are mutually exclusive with respect to the users who contributed the media (i.e., the users in the training set will be different from the users in the test set). All photos and videos used in the benchmark have been taken from the YFCC100M dataset.

This data has been extracted from the YFCC100M[1] dataset (Yahoo Flickr Creative Commons 100M) (Thomee et al., 2015). This resource has 99.3 million images and 0.7 million videos.

The YFCC100M must be separately requested via the Yahoo! Webscope portal at `http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67`. Yahoo! Labs makes this data available for free, subject to the terms of the Webscope agreement on data use (which you will be asked to sign). Your use of the YFCC100M dataset is not tied in any way to your use of the YLI-GEO features, nor vice versa.

`https://multimedia-commons.s3-us-west-2.amazonaws.com/subsets/YLI-GEO/features/README_YLI-GEO_10-9-2015.txt`

`http://www.multimediacommons.org/`

A.3.3 W&B Training Set and Wikipedia UK spot test

In the georeferencing experiments for formal documents reported in this PhD thesis the Wikipedia UK Spot Test corpus was used as a test set, and the W&B corpus was used as a training set. The Wikipedia UK Spot Test Corpus is a set of 21,839 English Wikipedia documents that refer to a spot (railwaystation, landmarks, buildings, and schools) of United Kingdom and it was created by Van Laere et al. (2014). It can be downloaded at: `https://github.com/ovlaere/georeferencing_wikipedia` The W&B Wikipedia training set is a set of about 390,574 Wikipedia documents created by Roller et al. (2012). `http://web.corral.tacc.utexas.edu/utcompling/wing-baldridge-2011/enwiki-20100905/` This set

[1] `http://www.yli-corpus.org/`

has been processed to filter out the Wikipedia UK Spot Test Corpus documents, resulting a total of 376,110 documents (Van Laere et al., 2014).

A.3.4 Test set of Flickr Metadata for Emergency Scenarios

From the set of 510,000 items (photos/videos) of the MEPT2014 test set were extracted a total of 784 items that contain at least one of these keywords using the Keywords (user tags) metadata field. A total of 63 items contain the keyword "riot", 229 items contain the keyword "hurricane", 366 contain the keyword "flood", and 144 items contain the keyword "earthquake". The MEPT2014 test set can be downloaded at `https://multimedia-commons.s3-us-west-2.amazonaws.com/subsets/YLI-GEO/docs/mediaeval2014_placing_test.bz2`

APPENDIX B

Annex B: GeoCLEF Topics

This annex shows the GeoCLEF topics used in the GeoCLEF evaluation benchmarks from 2005 to 2008 (a total of 100 topics). This annex also includes several classification of these topics.

Figure B.1: List of GeoCLEF 2005 topics.

```
<GeoCLEF-2005-Topics-English>
<top>
<num> GC001 </num>
<orignum> C084 </orignum>
<EN-title> Shark Attacks off Australia and California </EN-title>
<EN-desc> Documents will report any information relating to shark attacks on humans. </EN-desc>
<EN-narr> Identify instances where a human was attacked by a shark, including where the attack
took place and the circumstances surrounding the attack. Only documents concerning specific
attacks are relevant; unconfirmed shark attacks or suspected bites are not relevant. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Shark Attacks </EN-concept>
<EN-spatialrelation> near </EN-spatialrelation>
<EN-location> Australia </EN-location>
<EN-location> California </EN-location>
</top>

<top>
<num> GC002 </num>
<orignum> C090 </orignum>
<EN-title> Vegetable Exporters of Europe </EN-title>
<EN-desc> What countries are exporters of fresh, dried or frozen vegetables? </EN-desc>
<EN-narr> Any report that identifies a country or territory that exports fresh, dried or frozen
 vegetables, or indicates the country of origin of imported vegetables is relevant. Reports
 regarding canned vegetables, vegetable juices or otherwise processed vegetables are not
 relevant. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Vegetable Exporters </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Europe </EN-location>
</top>

<top>
<num> GC003 </num>
<orignum> C091 </orignum>
<EN-title> AI in Latin America </EN-title>
<EN-desc> Amnesty International reports on human rights in Latin America. </EN-desc>
<EN-narr> Relevant documents should inform readers about Amnesty International reports
regarding human rights in Latin America, or on reactions to these reports. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Amnesty International Human Rights Reports </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Latin America </EN-location>
</top>

<top>
<num> GC004 </num>
<orignum> C126 </orignum>
<EN-title> Actions against the fur industry in Europe and the U.S.A. </EN-title>
<EN-desc> Find information on protests or violent acts against the fur industry. </EN-desc>
<EN-narr> Relevant documents describe measures taken by animal right activists against fur
farming and/or fur commerce, e.g. shops selling items in fur. Articles reporting actions
taken against people wearing furs are also of importance. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Animal Rights Actions against the fur industry </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Europe </EN-location>
<EN-location> United States </EN-location>
</top>
```

```
<top>
<num> GC005 </num>
<orignum> C145 </orignum>
<EN-title> Japanese Rice Imports </EN-title>
<EN-desc> Find documents discussing reasons for and consequences of the first imported
rice in Japan. </EN-desc>
<EN-narr> In 1994, Japan decided to open the national rice market for the first time to
other countries. Relevant documents will comment on this question. The discussion can
include the names of the countries from which the rice is imported, the types of rice,
and the controversy that this decision prompted in Japan. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Rice imports </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Japan </EN-location>
</top>

<top>
<num> GC006 </num>
<orignum> C147 </orignum>
<EN-title> Oil Accidents and Birds in Europe </EN-title>
<EN-desc> Find documents describing damage or injury to birds caused by accidental oil
spills or pollution. </EN-desc>
<EN-narr> All documents which mention birds suffering because of oil accidents are relevant.
Accounts of damage caused as a result of bilge discharges or oil dumping are
not relevant. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Oil Accidents </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Europe </EN-location>
</top>

<top>
<num> GC007 </num>
<orignum> C156 </orignum>
<EN-title> Trade Unions in Europe </EN-title>
<EN-desc> What are the differences in the role and importance of trade unions between
European countries? </EN-desc>
<EN-narr> Relevant documents must compare the role, status or importance of trade unions
between two or more European countries. Pertinent information will include level of
organisation, wage negotiation mechanisms, and the general climate of the labour market.
</EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Trade Unions </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Europe </EN-location>
</top>

<top>
<num> GC008 </num>
<orignum> C177 </orignum>
<EN-title> Milk Consumption in Europe </EN-title>
<EN-desc> Provide statistics or information concerning milk consumption
in European countries. </EN-desc>
<EN-narr> Relevant documents must provide statistics or other information about milk
consumption in Europe, or in single European nations. Reports on milk derivatives
are not relevant. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Milk Consumption </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Europe </EN-location>
</top>
```

```
<top>
<num> GC009 </num>
<orignum> C190 </orignum>
<EN-title> Child Labor in Asia </EN-title>
<EN-desc> Find documents that discuss child labor in Asia and proposals to eliminate it or to
improve working conditions for children. </EN-desc>
<EN-narr> Documents discussing child labor in particular countries in Asia, descriptions of
working conditions for children, and proposals of measures to eliminate child labor
are all relevant. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Child Labor </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Asia </EN-location>
</top>

<top>
<num> GC010 </num>
<orignum> C200 </orignum>
<EN-title> Flooding in Holland and Germany </EN-title>
<EN-desc> Find statistics on flood disasters in Holland and Germany in 1995. </EN-desc>
<EN-narr> Relevant documents will quantify the effects of the damage
caused by flooding that took place in Germany and the Netherlands in 1995 in terms of
numbers of people and animals evacuated and/or of economic losses. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Floods, Flooding </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Holland </EN-location>
<EN-location> Germany </EN-location>
</top>

<top>
<num> GC011 </num>
<EN-title> Roman cities in the UK and Germany </EN-title>
<EN-desc> Roman cities in the UK and Germany. </EN-desc>
<EN-narr> A relevant document will identify one or more cities in the United
Kingdom or Germany which were also cities in Roman times. </EN-narr>
<EN-concept> Roman Cities </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> UK </EN-location>
<EN-location> Germany </EN-location>
</top>

<top>
<num> GC012 </num>
<EN-title> Cathedrals in Europe </EN-title>
<EN-desc> Find stories about particular cathedrals in Europe, including the
United Kingdom and Russia. </EN-desc>
<EN-narr> In order to be relevant, a story must be about or describe a particular cathedral
 in a particular country or place within a country in Europe, the UK or Russia.  Not relevant
 are stories which are generally about tourist tours of cathedrals or about the funeral
 of a particular person in a cathedral. </EN-narr>
<EN-concept> Cathedrals </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Europe </EN-location>
</top>
```

```
<top>
<num> GC013 </num>
<EN-title> Visits of the American president to Germany </EN-title>
<EN-desc> Find articles about visits of President Clinton to Germany. </EN-desc>
<EN-narr> Relevant documents should describe the stay of President Clinton in Germany
not purely the status of American-German relations. </EN-narr>
<EN-concept> Visits of American President </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Germany </EN-location>
</top>

<top>
<num> GC014 </num>
<EN-title> Environmentally hazardous Incidents in the North Sea </EN-title>
<EN-desc> Find documents about environmental accidents and hazards in the North Sea region.
</EN-desc>
<EN-narr> Relevant documents will describe accidents and environmentally hazardous
actions in or around the North Sea. Documents about oil production
can be included if they describe environmental impacts. </EN-narr>
<EN-concept> Environmentally Hazardous Incidents </EN-concept>
<EN-spatialrelation> in or around </EN-spatialrelation>
<EN-location> North Sea </EN-location>
</top>

<top>
<num> GC015 </num>
<EN-title> Consequences of the genocide in Rwanda </EN-title>
<EN-desc> Find documents about genocide in Rwanda and its impacts. </EN-desc>
<EN-narr>
Relevant documents will describe the country's situation after the genocide and the
political, economic and other efforts involved in attempting to stabilize the country.
</EN-narr>
<EN-concept> Genocide </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Rwanda </EN-location>
</top>

<top>
<num> GC016 </num>
<EN-title> Oil prospecting and ecological problems in Siberia and the Caspian Sea </EN-title>
<EN-desc> Find documents about Oil or petroleum development and related
ecological problems in Siberia and the Caspian Sea regions. </EN-desc>
<EN-narr> Relevant documents will discuss the exploration for, and exploitation of
petroleum (oil) resources in the Russian region of Siberia and in or near
the Caspian Sea. Relevant documents will also discuss ecological issues or
problems, including disasters or accidents in these regions. </EN-narr>
<EN-concept> Oil Prospecting </EN-concept>
<EN-concept> Ecological Problems </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Siberia </EN-location>
<EN-spatialrelation> near </EN-spatialrelation>
<EN-location> Caspian Sea </EN-location>
</top>
```

```
<top>
<num> GC017 </num>
<EN-title> American Troops in Sarajevo, Bosnia-Herzegovina </EN-title>
<EN-desc> Find documents about American troop deployment in Bosnia-Herzegovina,
especially Sarajevo. </EN-desc>
<EN-narr> Relevant documents will discuss deployment of American (USA) troops as
part of the UN peacekeeping force in the former Yugoslavian regions of
Bosnia-Herzegovina, and in particular in the city of Sarajevo. </EN-narr>
<EN-concept> American Troops </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Sarajevo </EN-location>
<EN-location> Bosnia-Herzegovina </EN-location>
</top>

<top>
<num> GC018 </num>
<EN-title> Walking holidays in Scotland </EN-title>
<EN-desc> Find documents that describe locations for walking holidays in Scotland. </EN-desc>
<EN-narr> A relevant document will describe a place or places within Scotland where a
walking holiday could take place. </EN-narr>
<EN-concept> Walking holidays </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Scotland </EN-location>
</top>

<top>
<num> GC019 </num>
<EN-title> Golf tournaments in Europe </EN-title>
<EN-desc> Find information about golf tournaments held in European locations. </EN-desc>
<EN-narr> A relevant document will describe the planning, running and/or results of a golf
tournament held at a location in Europe. </EN-narr>
<EN-concept> Golf tournaments </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Europe </EN-location>
</top>

<top>
<num> GC020 </num>
<EN-title> Wind power in the Scottish Islands </EN-title>
<EN-desc> Find documents on electrical power generation using wind power
in the islands of Scotland. </EN-desc>
<EN-narr> A relevant document will describe wind power-based electricity generation schemes
 providing electricity for the islands of Scotland. </EN-narr>
<EN-concept> Wind power </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> the Scottish Islands </EN-location>
</top>

<top>
<num> GC021 </num>
<EN-title> Sea rescue in North Sea </EN-title>
<EN-desc> Find items about rescues in the North Sea. </EN-desc>
<EN-narr> A relevant document will report a sea rescue undertaken in North Sea. </EN-narr>
<EN-concept> Sea rescue </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> North Sea </EN-location>
</top>
```

```
<top>
<num> GC022 </num>
<EN-title> Restored buildings in Southern Scotland </EN-title>
<EN-desc> Find articles on the restoration of historic buildings in the southern
part of Scotland. </EN-desc>
<EN-narr> A relevant document will describe a restoration of historical buildings in
the southern Scotland. </EN-narr>
<EN-concept> Restored buildings </EN-concept>
<EN-spatialrelation> south of </EN-spatialrelation>
<EN-location> Scotland </EN-location>
</top>

<top>
<num> GC023 </num>
<EN-title> Murders and violence in South-West Scotland </EN-title>
<EN-desc> Find articles on violent acts including murders in the South West
part of Scotland. </EN-desc>
<EN-narr> A relevant document will give details of either specific acts of violence or death
 related to murder or information about the general state of violence in South West Scotland.
  This includes information about violence in places such as Ayr, Campeltown,
  Douglas and Glasgow. </EN-narr>
<EN-concept> Murders and violence </EN-concept>
<EN-spatialrelation> South-West of </EN-spatialrelation>
<EN-location> Scotland </EN-location>
</top>

<top>
<num> GC024 </num>
<EN-title> Factors influencing tourist industry in Scottish Highlands </EN-title>
<EN-desc> Find articles on the tourism industry in the Highlands of Scotland
and the factors affecting it. </EN-desc>
<EN-narr> A relevant document will provide information on factors which have affected or
influenced tourism in the Scottish Highlands. For example, the construction of roads or
railways, initiatives to increase tourism, the planning and construction of new attractions
 and influences from the environment (e.g. poor weather). </EN-narr>
<EN-concept> influences on tourist industry </EN-concept>
<EN-spatialrelation> in </EN-spatialrelation>
<EN-location> Scottish Highlands </EN-location>
</top>


<top>
<num> GC025 </num>
<EN-title> Environmental concerns in and around the Scottish Trossachs </EN-title>
<EN-desc> Find articles about environmental issues and concerns in
the Trossachs region of Scotland. </EN-desc>
<EN-narr> A relevant document will describe environmental concerns (e.g. pollution,
 damage to the environment from tourism) in and around the area in Scotland known as
  the Trossachs. Strictly speaking, the Trossachs is the narrow wooded glen between
   Loch Katrine and Loch Achray, but the name is now used to describe a much larger
    area between Argyll and Perthshire, stretching north from the Campsies and west
     from Callander to the eastern shore of Loch Lomond. </EN-narr>
<EN-concept> Environmental concerns </EN-concept>
<EN-spatialrelation> in and around </EN-spatialrelation>
<EN-location> the Scottish Trossachs </EN-location>
</top>

</GeoCLEF-2005-Topics-English>
```

Figure B.2: List of GeoCLEF 2006 topics.

```
<GeoCLEF-2006-Topics-English>
<top>
  <num>GC026</num>
  <EN-title>Wine regions around rivers in Europe</EN-title>
  <EN-desc>Documents about wine regions along the banks of European rivers</EN-desc>
  <EN-narr>Relevant documents describe a wine region along a major river in European countries.
To be relevant the document must name the region and the river.</EN-narr>
 </top>
 <top>
  <num>GC027</num>
  <EN-title>Cities within 100km of Frankfurt</EN-title>
  <EN-desc>Documents about cities within 100 kilometers of the city of Frankfurt in Western
  Germany</EN-desc>
  <EN-narr>Relevant documents discuss cities within 100 kilometers of Frankfurt am Main Germany,
  latitude 50.11222, longitude 8.68194.  To be relevant the document must describe the city or
  an event in that city. Stories about Frankfurt itself are not relevant</EN-narr>
 </top>
<top>
<num>GC028</num>
<EN-title>Snowstorms in North America</EN-title>
<EN-desc>Documents about snowstorms occurring in the north part of the American continent</EN-desc>
<EN-narr>Relevant documents state cases of snowstorms and their effects in North America.
  Countries are Canada, United States of America and Mexico.  Documents about other kinds
   of storms are not relevant (e.g. rainstorm, thunderstorm, electric storm, windstorm)</EN-narr>
</top>
<top>
<num>GC029</num>
<EN-title>Diamond trade in Angola and South Africa</EN-title>
<EN-desc>Documents regarding diamond trade in Angola and South Africa</EN-desc>
<EN-narr>Relevant documents are about diamond trading in these two countries and its
consequences (e.g. smuggling, economic and political instability)</EN-narr>
</top>
<top>
<num>GC030</num>
<EN-title>Car bombings near Madrid</EN-title>
<EN-desc>Documents about car bombings occurring near Madrid</EN-desc>
<EN-narr>Relevant documents treat cases of car bombings occurring in the capital of Spain
 and its outskirts</EN-narr>
</top>
<top>
<num>GC031</num>
<EN-title>Combats and embargo in the northern part of Iraq</EN-title>
<EN-desc>Documents telling about combats or embargo in the northern part of Iraq</EN-desc>
<EN-narr>Relevant documents are about combats and effects of the 90s embargo in the northern
part of Iraq. Documents about these facts happening in other parts of Iraq are not relevant</EN-narr>
</top>
<top>
<num>GC032</num>
<EN-title>Independence movement in Quebec</EN-title>
<EN-desc>Documents about actions in Quebec for the independence of this Canadian province</EN-desc>
<EN-narr>Relevant documents treat matters related to Quebec independence movement
(e.g. referendums) which take place in Quebec</EN-narr>
</top>
<top>
<num>GC033</num>
<EN-title> International sports competitions in the Ruhr area</EN-title>
<EN-desc> World Championships and international tournaments in
the Ruhr area</EN-desc>
<EN-narr> Relevant documents state the type or name of the competition,
the city and possibly results. Irrelevant are documents where only part of the competition
 takes place in the Ruhr area of Germany, e.g. Tour de France,
Champions League or UEFA-Cup games.</EN-narr>
</top>
```

```
<top>
<num> GC034 </num>
<EN-title> Malaria in the tropics </EN-title>
<EN-desc> Malaria outbreaks in tropical regions and preventive
vaccination </EN-desc>
<EN-narr> Relevant documents state cases of malaria in tropical regions
and possible preventive measures like chances to vaccinate against the
disease. Outbreaks must be of epidemic scope. Tropics are defined as the region between the
 Tropic of Capricorn, latitude 23.5 degrees South and the Tropic of Cancer, latitude 23.5
  degrees North.  Not relevant are documents about a single person's infection.  </EN-narr>
</top>
<top>
<num> GC035 </num>
<EN-title> Credits to the former Eastern Bloc </EN-title>
<EN-desc> Financial aid in form of credits by the International
Monetary Fund or the World Bank to countries formerly belonging to
the "Eastern Bloc" aka the Warsaw Pact, except the republics of the former USSR</EN-desc>
<EN-narr> Relevant documents cite agreements on credits, conditions or
consequences of these loans. The Eastern Bloc is defined as countries
under strong Soviet influence (so synonymous with Warsaw Pact) throughout
the whole Cold War.   Excluded are former USSR republics.  Thus the countries are Bulgaria,
 Hungary, Czech Republic, Slovakia, Poland and Romania.  Thus not all communist or
 socialist countries are considered relevant.</EN-narr>
</top>
<top>
<num> GC036 </num>
<EN-title> Automotive industry around the Sea of Japan </EN-title>
<EN-desc> Coastal cities on the Sea of Japan with automotive industry or factories </EN-desc>
<EN-narr> Relevant documents report on automotive industry or factories in
cities on the shore of the Sea of Japan (also named East Sea (of Korea))
including economic or social events happening there like planned joint-ventures
or strikes. In addition to Japan, the countries of North Korea, South Korea and Russia
are also on the Sea of Japan.</EN-narr>
</top>
<top>
<num> GC037 </num>
<EN-title> Archeology in the Middle East </EN-title>
<EN-desc> Excavations and archeological finds in the Middle East </EN-desc>
<EN-narr> Relevant documents report recent finds in some town, city, region or country
 of the Middle East, i.e. in Iran, Iraq, Turkey, Egypt, Lebanon, Saudi Arabia, Jordan,
 Yemen, Qatar, Kuwait, Bahrain, Israel, Oman, Syria, United Arab Emirates, Cyprus,
 West Bank, or the Gaza Strip</EN-narr>
</top>
<top>
<num> GC038 </num>
<EN-title> Solar or lunar eclipse in Southeast Asia </EN-title>
<EN-desc> Total or partial solar or lunar eclipses in Southeast Asia
</EN-desc>
<EN-narr> Relevant documents state the type of eclipse and the region or country
of occurrence, possibly also stories about people travelling to see it.  Countries of
Southeast Asia are Brunei, Cambodia, East Timor, Indonesia, Laos, Malaysia, Myanmar,
Philippines, Singapore, Thailand and Vietnam.
</EN-narr>
</top>
<top>
<num> GC039 </num>
<EN-title> Russian troops in the southern Caucasus </EN-title>
<EN-desc> Russian soldiers, armies or military bases in the Caucasus region
south of the Caucasus Mountains </EN-desc>
<EN-narr> Relevant documents report on Russian troops based at, moved to or
removed from the region. Also agreements on one of these actions or combats
are relevant. Relevant countries are: Azerbaijan, Armenia, Georgia, Ossetia,
Nagorno-Karabakh. Irrelevant are documents citing actions between troops of
nationality different from Russian (with Russian mediation between the two.)
</EN-narr>
</top>
```

```
<top>
<num> GC040 </num>
<EN-title> Cities near active volcanoes </EN-title>
<EN-desc> Cities, towns or villages threatened by the eruption of a volcano
</EN-desc>
<EN-narr> Relevant documents cite the name of the cities, towns, villages that
are near an active volcano which recently had an eruption or could erupt soon.
Irrelevant are reports which do not state the danger (i.e. for example necessary
preventive evacuations) or the consequences for specific cities , but just
tell that a particular volcano (in some country) is going to erupt, has erupted or that
a region has active volcanoes. </EN-narr>
</top>

<top>
<num>GC041</num>
<EN-title>Shipwrecks in the Atlantic Ocean</EN-title>
<EN-desc>Documents about shipwrecks in the Atlantic Ocean</EN-desc>
<EN-narr>Relevant documents should document shipwreckings in any part of the Atlantic Ocean
 or its coasts.</EN-narr>
</top>

<top>
<num>GC042</num>
<EN-title>Regional elections in Northern Germany</EN-title>
<EN-desc>Documents about regional elections in Northern Germany</EN-desc>
<EN-narr>Relevant documents are those reporting the campaign or results for the state
parliaments of any of the regions of Northern Germany. The states of northern Germany
 are commonly Bremen, Hamburg, Lower Saxony, Mecklenburg-Western Pomerania and
 Schleswig-Holstein. Only regional elections are relevant; municipal, national
  and European elections are not.</EN-narr>
</top>

<top>
<num>GC043</num>
<EN-title>Scientific research in New England Universities</EN-title>
<EN-desc>Documents about scientific research in New England universities</EN-desc>
<EN-narr>Valid documents should report specific scientific research or breakthroughs
 occurring in universities of New England. Both current and past research are relevant.
  Research regarded as bogus or fraudulent is also relevant.  New England states are:
   Connecticut, Rhode Island, Massachusetts, Vermont, New Hampshire, Maine. </EN-narr>
</top>

<top>
<num>GC044</num>
<EN-title>Arms sales in former Yugoslavia</EN-title>
<EN-desc>Documents about arms sales in former Yugoslavia</EN-desc>
<EN-narr>Relevant documents should report on arms sales that took place in the successor
 countries of the former Yugoslavia. These sales can be legal or not, and to any kind
  of entity in these states, not only the government itself. Relevant countries are:
   Slovenia, Macedonia, Croatia, Serbia and Montenegro, and Bosnia and Herzegovina.
</EN-narr>
</top>

<top>
<num>GC045</num>
<EN-title>Tourism in Northeast Brazil</EN-title>
<EN-desc>Documents about tourism in Northeastern Brazil</EN-desc>
<EN-narr>Of interest are documents reporting on tourism in Northeastern Brazil,
 including places of interest, the tourism industry and/or the reasons for taking
  or not a holiday there.  The states of northeast Brazil are Alagoas, Bahia, Ceara,
   Maranhao, Paraiba, Pernambuco, Piaui, Rio Grande do Norte and Sergipe.</EN-narr>
</top>
```

```
<top>
<num>GC046</num>
<EN-title>Forest fires in Northern Portugal</EN-title>
<EN-desc>Documents about forest fires in Northern Portugal</EN-desc>
<EN-narr>Documents should report the ocurrence, fight against, or aftermath of
 forest fires in Northern Portugal.  The regions covered are Minho, Douro Litoral,
  Tras-os-Montes and Alto Douro, corresponding to the districts of Viana do Castelo,
   Braga, Porto (or Oporto), Vila Real and Bragana.
</EN-narr>
</top>

<top>
<num>GC047</num>
<EN-title>Champions League games near the Mediterranean </EN-title>
<EN-desc>Documents about Champion League games played in European cities bordering the
 Mediterranean </EN-desc>
<EN-narr>Relevant documents should include at least a short description of a European
 Champions League game played in a European city bordering the Mediterranean Sea or
  any of its minor seas. European countries along the Mediterranean Sea are Spain,
   France, Monaco, Italy, the island state of Malta, Slovenia, Croatia, Bosnia and
    Herzegovina, Serbia and Montenegro, Albania, Greece, Turkey, and the
     island of Cyprus.</EN-narr>
</top>

<top>
<num>GC048</num>
<EN-title>Fishing in Newfoundland and Greenland</EN-title>
<EN-desc>Documents about fisheries around Newfoundland and Greenland</EN-desc>
<EN-narr>Relevant documents should document fisheries and economical, ecological or
 legal problems associated with it, around Greenland and the Canadian island
  of Newfoundland. </EN-narr>
</top>

<top>
<num>GC049</num>
<EN-title>ETA in France</EN-title>
<EN-desc>Documents about ETA activities in France</EN-desc>
<EN-narr>Relevant documents should document the activities of the Basque terrorist
 group ETA in France, of a paramilitary, financial, political nature or others. </EN-narr>
</top>

<top>
<num>GC050</num>
<EN-title>Cities along the Danube and the Rhine</EN-title>
<EN-desc>Documents describe cities in the shadow of the Danube or the Rhine</EN-desc>
<EN-narr>Relevant documents should contain at least a short description of
 cities through which the rivers Danube and Rhine pass, providing evidence for it.
  The Danube flows through nine countries (Germany, Austria, Slovakia, Hungary,
   Croatia, Serbia, Bulgaria, Romania, and Ukraine).  Countries along the Rhine
    are Liechtenstein, Austria, Germany, France, the Netherlands and Switzerland. </EN-narr>
</top>
</GeoCLEF-2006-Topics-English>
```

Figure B.3: List of GeoCLEF 2007 topics.

```
<topics>
<top lang="en">
<num>10.2452/51-GC</num>
<title>Oil and gas extraction found between the UK and the Continent</title>
<desc>To be relevant documents describing oil or gas production between the UK and the European
continent will be relevant</desc>
<narr>Oil and gas fields in the North Sea will be relevant.</narr>
</top>
<top lang="en">
<num>10.2452/52-GC</num>
<title>Crime near St Andrews</title>
<desc>To be relevant, documents must be about crimes occurring close to or
in St. Andrews.</desc>
<narr>Any event that refers to criminal dealings of some sort is relevant, from thefts to
corruption.</narr>
</top>
<top lang="en">
<num>10.2452/53-GC</num>
<title>Scientific research at east coast Scottish Universities</title>
<desc>For documents to be relevant, they must describe scientific research conducted by a
 Scottish University located on the east coast of Scotland</desc>
<narr>Universities in Aberdeen, Dundee, St Andrews and Edinburgh wil be considered
 relevant locations.</narr>
</top>
<top lang="en">
<num>10.2452/54-GC</num>
<title>Damage from acid rain in northern Europe</title>
<desc>Documents describing the damage caused by acid rain in the countries of
northern Europe</desc>
<narr>Relevant countries include Denmark, Estonia, Finland, Iceland, Republic of Ireland,
 Latvia, Lithuania, Norway, Sweden, United Kingdom and northeastern parts of Russia</narr>
</top>
<top lang="en">
<num>10.2452/55-GC</num>
<title>Deaths caused by avalanches occurring in Europe, but not in the Alps</title>
<desc>To be relevant a document must describe the death of a person caused by an avalanche
 that occurred away from the Alps but in Europe.</desc>
<narr>for example mountains in Scotland, Norway, Iceland</narr>
</top>
<top lang="en">
<num>10.2452/56-GC</num>
<title>Lakes with monsters</title>
<desc>To be relevant, the document must describe a lake where a monster is
supposed to exist.</desc>
<narr>The document must state the alledged existence of a monster in a particular lake
 and must name the lake. Activities which try to prove the existence of the monster and
  reports of witnesses who have seen the monster are relevant. Documents which mention
   only the name of a particular monster are not relevant.</narr>
</top>
<top lang="en">
<num>10.2452/57-GC</num>
<title>Whisky making in the Scottlsh Islands</title>
<desc>To be relevant, a document must describe a whisky made, or a whisky distillery
 located, on a Scottish island.</desc>
<narr>Relevant islands are Islay, Skye, Orkney, Arran, Jura, Mull.&#13;
Relevant whiskys are Arran Single Malt; Highland Park Single Malt; Scapa; Isle of Jura;
 Talisker; Tobermory;  Ledaig; Ardbeg; Bowmore; Bruichladdich; Bunnahabhain; Caol Ila;
  Kilchoman; Lagavulin; Laphroaig</narr>
</top>
```

```
<top lang="en">
<num>10.2452/58-GC</num>
<title>Travel problems at major airports near to London</title>
<desc>To be relevant, documents must describe travel problems at one of the major
airports close to London.</desc>
<narr>Major airports to be listed include Heathrow, Gatwick, Luton, Stanstead and
 London City airport.</narr>
</top>
<top lang="en">
<num>10.2452/59-GC</num>
<title>Meetings of the Andean Community of Nations (CAN)</title>
<desc>Find documents mentioning cities in on the meetings of the Andean Community of
Nations (CAN) took place</desc>
<narr>relevant documents mention cities in which meetings of the members of the Andean
 Community of Nations (CAN - member states Bolivia, Columbia, Ecuador, Peru).</narr>
</top>
<top lang="en">
<num>10.2452/60-GC</num>
<title>Casualties in fights in Nagorno-Karabakh</title>
<desc>Documents reporting on casualties in the war in Nagorno-Karabakh</desc>
<narr>Relevant documents report of casualties during the war or in fights in the Armenian
 enclave Nagorno-Karabakh</narr>
</top>
<top lang="en">
<num>10.2452/61-GC</num>
<title>Airplane crashes close to Russian cities</title>
<desc>Find documents mentioning airplane crashes close to Russian cities</desc>
<narr>Relevant documents report on airplane crashes in Russia. The location is to be
 specified by the name of a city mentioned in the document.</narr>
</top>
<top lang="en">
<num>10.2452/62-GC</num>
<title>OSCE meetings in Eastern Europe</title>
<desc>Find documents in which Eastern European conference venues of the  Organization
 for Security and Co-operation in Europe (OSCE) are mentioned</desc>
<narr>Relevant documents report on OSCE meetings in Eastern Europe. Eastern Europe
 includes Bulgaria, Poland, the Czech Republic, Slovakia, Hungary, Romania, Ukraine,
  Belarus, Lithuania, Estonia, Latvia and the European part of Russia.</narr>
</top>
<top lang="en">
<num>10.2452/63-GC</num>
<title>Water quality along coastlines of the Mediterranean Sea</title>
<desc>Find documents on the water quality at the coast of the Mediterranean Sea</desc>
<narr>Relevant documents report on the water quality along the coast and coastlines of
 the Mediterranean Sea. The coasts must be specified by their names.</narr>
</top>
<top lang="en">
<num>10.2452/64-GC</num>
<title>Sport events in the french speaking part of Switzerland</title>
<desc>Find documents on sport events in the french speaking part of Switzerland</desc>
<narr>Relevant documents report sport events in the french speaking part of Switzerland.
 Events in cities like Lausanne, Geneva, Neuch\^{a}tel and Fribourg are relevant.</narr>
</top>
```

```
<top lang="en">
<num>10.2452/65-GC</num>
<title>Free elections in Africa</title>
<desc>Documents mention free elections held in countries in Africa</desc>
<narr>Future elections or promises of free elections are not relevant</narr>
</top>
<top lang="en">
<num>10.2452/66-GC</num>
<title>Economy at the Bosphorus</title>
<desc>Documents on economic trends at the Bosphorus strait</desc>
<narr>Relevant documents report on economic trends and development in the Bosphorus
 region close to Istanbul</narr>
</top>
<top lang="en">
<num>10.2452/67-GC</num>
<title>F1 circuits where Ayrton Senna competed in 1994</title>
<desc>Find documents that mention circuits where  the Brazilian driver Ayrton Senna
 participated in 1994. The name and location of the circuit is required</desc>
<narr>Documents should indicate that Ayrton Senna participated in a race in a particular
 stadion, and the location of the race track.</narr>
</top>
<top lang="en">
<num>10.2452/68-GC</num>
<title>Rivers with floods</title>
<desc>Find documents that mention rivers that flooded. The name of the river
 is required.</desc>
<narr>Documents that mention floods but fail to name the rivers are not relevant.</narr>
</top>
<top lang="en">
<num>10.2452/69-GC</num>
<title>Death on the Himalaya</title>
<desc>Documents should mention deaths due to climbing mountains in the Himalaya range.</desc>
<narr>Only death casualties of mountaineering athletes in the Himalayan mountains,
 such as Mount Everest or Annapurna, are interesting. Other deaths, caused by e.g.
  political unrest in the region, are irrelevant.</narr>
</top>
<top lang="en">
<num>10.2452/70-GC</num>
<title>Tourist attractions in Northern Italy</title>
<desc>Find documents that identify tourist attractions in the North of Italy.</desc>
<narr>Documents should mention places of tourism in the North of Italy, either specifying
 particular tourist attractions (and where they are located) or mentioning that the place
  (town, beach, opera, etc.) attracts many tourists.</narr>
</top>
<top lang="en">
<num>10.2452/71-GC</num>
<title>Social problems in greater Lisbon</title>
<desc>Find information about social problems afllicting places in greater Lisbon.</desc>
<narr>Documents are relevant if they mention any social problem, such as drug consumption,
 crime, poverty, slums, unemployment or lack of integration of minorities, either for the
  region as a whole or in specific areas inside it. Greater Lisbon includes the Amadora,
   Cascais, Lisboa, Loures, Mafra, Odivelas, Oeiras , Sintra and Vila Franca de
    Xira districts.</narr>
</top>
```

```
<top lang="en">
<num>10.2452/72-GC</num>
<title>Beaches with sharks</title>
<desc>Relevant documents should name beaches or coastlines where there is danger of
 shark attacks. Both particular attacks and the mention of danger are relevant, provided
  the place is mentioned.</desc>
<narr>Provided that a geographical location is given, it is sufficient that fear or danger
 of sharks is mentioned. No actual accidents need to be reported.</narr>
</top>
<top lang="en">
<num>10.2452/73-GC</num>
<title>Events at St. Paul's Cathedral</title>
<desc>Any event that happened at St. Paul's cathedral is relevant, from concerts,
masses, ceremonies or even accidents or thefts.</desc>
<narr>Just the description of the church or its mention as a tourist attraction is
 not relevant. There are three relevant St. Paul's cathedrals for this topic: those
  of S\~{a}o Paulo, Rome and London.</narr>
</top>
<top lang="en">
<num>10.2452/74-GC</num>
<title>Ship traffic around the Portuguese islands</title>
<desc>Documents should mention ships or sea traffic connecting Madeira and the
 Azores to other places, and also connecting the several isles of each archipelago.
  All subjects, from wrecked ships, treasure finding, fishing, touristic tours to
   military actions, are relevant, except for historical narratives.</desc>
<narr>Documents have to mention that there is ship traffic connecting the isles to
 the continent (portuguese mainland), or between the several islands, or showing
  international traffic. Isles of Azores are: S\~{a}o Miguel, Santa Maria, Formigas,
   Terceira, Graciosa, S\~{a}o Jorge, Pico, Faial, Flores and Corvo. The Madeira
    islands are: Mardeira, Porto Santo,  Desertas islets and Selvagens islets.</narr>
</top>
<top lang="en">
<num>10.2452/75-GC</num>
<title>Violation of human rights in Burma</title>
<desc>Documents are relevant if they mention actual violation of human rights
 in Myanmar, previously named Burma.</desc>
<narr>This includes all reported violations of human rights in Burma, no matter
 when (not only by the present government). Declarations (accusations or denials)
  about the matter only, are not relevant.</narr>
</top>
</topics>
```

Figure B.4: List of GeoCLEF 2008 Topics

```
<topics>
<topic lang="en">
<identifier>10.2452/76-GC</identifier>
<title>Riots in South American prisons</title>
<description>Documents mentioning riots in prisons in South America</description>
<narrative>Relevant documents mention riots or uprising on the South American continent.
Countries in South America include Argentina, Bolivia, Brazil, Chile, Suriname, Ecuador,
Colombia, Guyana, Peru, Paraguay, Uruguay and Venezuela. French Guiana is a French
 province in South America.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/77-GC</identifier>
<title>Nobel prize winners from Northern European countries</title>
<description>Documents mentioning Noble prize winners born in a Northern European
country.</description>
<narrative>Relevant documents contain information about the field of research and the
country of origin of the prize winner. Northern European countries are: Denmark,
Finland, Iceland, Norway, Sweden, Estonia, Latvia, Belgium, the Netherlands, Luxembourg,
Ireland, Lithuania, and the UK. The north of Germany and Poland as well as the north-east
of Russia also belong to Northern Europe.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/78-GC</identifier>
<title>Sport events in the Sahara</title>
<description>Documents mentioning sport events occurring in (or passing through) the
Sahara.</description>
<narrative>Relevant documents must make reference to athletic events and to the place
where they take place. The Sahara covers huge parts of Algeria, Chad, Egypt, Libya,
Mali, Mauritania, Morocco, Niger, Western Sahara, Sudan, Senegal and Tunisia.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/79-GC</identifier>
<title>Invasion of Eastern Timor's capital by Indonesia</title>
<description>Documents mentioning the invasion of Dili by Indonesian troops</description>
<narrative>Relevant documents deal with the occupation of East Timor by Indonesia and
mention incidents between Indonesian soldiers and the inhabitants of Dili.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/80-GC</identifier>
<title>Politicians in exile in Germany</title>
<description>Documents mentioning exiled politicians in Germany</description>
<narrative>Relevant documents report about politicians who live in exile in Germany and
mention the nationality and political convictions of these politicians.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/81-GC</identifier>
<title>G7 summits in Mediterranean countries</title>
<description>Documents mentioning G7 summit meetings in Mediterranean countries</description>
<narrative>Relevant documents must mention summit meetings of the G7 in the mediterranean
countries: Spain, Gibraltar, France, Monaco, Italy, Malta, Slovenia, Croatia, Bosnia and
Herzegovina, Montenegro, Albania, Greece, Cyprus, Turkey, Syria, Lebanon, Israel,
Palestine, Egypt, Libya, Tunisia, Algeria and Morocco.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/82-GC</identifier>
<title>Agriculture in the Iberian Peninsula</title>
<description>Relevant documents relate to the state of agriculture in the Iberian
Peninsula</description>
<narrative>Relevant docments contain information about the state of agriculture
in the Iberian peninsula. Crops, protests and statistics are relevant. The countries
n the Iberian peninsula are Portugal, Spain and Andorra.</narrative>
</topic>
```

```
<topic lang="en">
<identifier>10.2452/83-GC</identifier>
<title>Demonstrations against terrorism in Northern Africa</title>
<description>Documents mentioning demonstrations against terrorism in Northern
Africa</description>
<narrative>Relevant documents must mention demonstrations against terrorism in the
North of Africa. The documents must mention the number of demonstrators and the reasons
for the demonstration. North Africa includes the Magreb region (countries: Algeria, Tunisia,
and Morocco, as well as the Western Sahara region) and Egypt, Sudan,
Libya and Mauritania.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/84-GC</identifier>
<title>Bombings in Northern Ireland</title>
<description>Documents mentioning bomb attacks in Northern Ireland</description>
<narrative>Relevant documents should contain information about bomb attacks in
Northern Ireland and should mention people responsible for and consequences
of the attacks.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/85-GC</identifier>
<title>Nuclear tests in the South Pacific</title>
<description>Documents mentioning the execution of nuclear tests in South
Pacific</description>
<narrative>Relevant documents should contain information about nuclear tests which were
carried out in the South Pacific. Intentions as well as plans for future nuclear tests
 in this region are not considered as relevant.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/86-GC</identifier>
<title>Most visited sights in the capital of France and its vicinity</title>
<description>Documents mentioning the most visited sights in Paris and
surroundings</description>
<narrative>Relevant documents should provide information about the most visited
sights of Paris and close to Paris and either give this information explicitly or
contain data which allows conclusions about which places were most visited.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/87-GC</identifier>
<title>Unemployment in the OECD countries</title>
<description>Documents mentioning issues related with the unemployment in the countries
of the Organisation for Economic Co-operation and Development (OECD)</description>
<narrative>Relevant documents should contain information about the unemployment (rate
of unemployment, important reasons and consequences) in the industrial states of the
OECD. The following states belong to the OECD: Australia, Belgium, Denmark, Germany,
Finland, France, Greece, Ireland, Iceland, Italy, Japan, Canada, Luxembourg, Mexico,
New Zealand, the Netherlands, Norway, Austria, Poland, Portugal, Sweden, Switzerland,
Slovakia, Spain, South Korea, Czech Republic, Turkey, Hungary, the United Kingdom and
the United States of America (USA).</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/88-GC</identifier>
<title>Portuguese immigrant communities in the world</title>
<description>Documents mentioning immigrant Portuguese communities in other countries
</description>
<narrative>Relevant documents contain information about Portguese communities who live
 as immigrants in other countries.</narrative>
</topic>
```

```
<topic lang="en">
<identifier>10.2452/89-GC</identifier>
<title>Trade fairs in Lower Saxony</title>
<description>Documents reporting about industrial or cultural fairs in Lower Saxony
</description>
<narrative>Relevant documents should contain information about trade or industrial
fairs which take place in the German federal state of Lower Saxony, i.e. name, type
and place of the fair. The capital of Lower Saxony is Hanover. Other cities include
Braunschweig, Osnabr{\"u}ck, Oldenburg and G{\"o}ttingen.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/90-GC</identifier>
<title>Environmental pollution in European waters</title>
<description>Documents mentioning environmental pollution in European rivers, lakes
and oceans.</description>
<narrative>Relevant documents should mention the kind and level of the pollution and
furthermore contain information about the type of the water and locate the affected
area and potential consequences.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/91-GC</identifier>
<title>Forest fires on Spanish islands</title>
<description>Documents mentioning forest fires on Spanish islands</description>
<narrative>Relevant documents should contain information about the location, causes
and consequences of the forest fires. Spanish Islands are: the Balearic Islands
(Majorca, Minorca, Ibiza, Formentera), the Canary Islands (Tenerife, Gran Canaria,
El Hierro, Lanzarote, La Palma, La Gomera, Fuerteventura) and some islands located
just off the Moroccan coast (Islas Chafarinas, Alhucemas, Albor\'{a}n, Perejil, Islas
Columbretes and Pen\'{o}n de V\'{e}lez de la Gomera).</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/92-GC</identifier>
<title>Islamic fundamentalists in Western Europe</title>
<description>Documents mentioning Islamic fundamentalists living in Western
Europe</description>
<narrative>Relevant Documents contain information about countries of origin and
current whereabouts and political and religious motives of the fundamentalists.
Western Europe consists of Western Europe consists of Belgium, Ireland,
Great Britain, Spain, Italy, Portugal, Andorra, Germany, France, Liechtenstein,
Luxembourg, Monaco, the Netherlands, Austria and Switzerland.</narrative> </topic>
<topic lang="en">
<identifier>10.2452/93-GC</identifier>
<title>Attacks in Japanese subways</title>
<description>Documents mentioning attacks in Japanese subways</description>
<narrative>Relevant documents contain information about attackers, reasons,
number of victims, places and consequences of the attacks in subways in
Japan.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/94-GC</identifier>
<title>Demonstrations in German cities</title>
<description>Documents mentioning demonstrations in German cities</description>
<narrative>Relevant documents contain information about participants, and number
of participants, reasons, type (peaceful or riots) and consequences of
demonstrations in German cities.</narrative>
</topic>
```

```
<topic lang="en">
<identifier>10.2452/95-GC</identifier>
<title>American troops in the Persian Gulf</title>
<description>Documents mentioning American troops in the Persian Gulf</description>
<narrative>Relevant documents contain information about functions/tasks of the
American troops and where exactly they are based. Countries with a coastline
with the Persian Gulf are: Iran, Iraq, Oman, United Arab Emirates, Saudi-Arabia,
Qatar, Bahrain and Kuwait.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/96-GC</identifier>
<title>Economic boom in Southeast Asia</title>
<description>Documents mentioning economic boom in countries in Southeast
Asia</description>
<narrative>Relevant documents contain information about (international) companies
in this region and the impact of the economic boom on the population. Countries of
Southeast Asia are: Brunei, Indonesia, Malaysia, Cambodia, Laos, Myanmar (Burma),
East Timor, the Phillipines, Singapore, Thailand and Vietnam.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/97-GC</identifier>
<title>Foreign aid in Sub-Saharan Africa</title>
<description>Documents mentioning foreign aid in Sub-Saharan Africa</description>
<narrative>Relevant documents contain information about the kind of foreign aid
and describe which countries or organizations help in which regions of Sub-Saharan
Africa. Countries of the Sub-Saharan Africa are: state of Central Africa (Burundi,
Rwanda, Democratic Republic of Congo, Republic of Congo, Central African Republic),
East Africa (Ethiopia, Eritrea, Kenya, Somalia, Sudan, Tanzania, Uganda, Djibouti),
Southern Africa (Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa,
Madagascar, Zambia, Zimbabwe, Swaziland), Western Africa (Benin, Burkina Faso, Chad,
C\^{o}te d'Ivoire, Gabon, Gambia, Ghana, Equatorial Guinea, Guinea-Bissau, Cameroon,
Liberia, Mali, Mauritania, Niger, Nigeria, Senegal, Sierra Leone, Togo) and the
African isles (Cape Verde, Comoros, Mauritius, Seychelles, S\~{o} Tom\'{e} and Pr\'{i}ncipe
and Madagascar).</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/98-GC</identifier>
<title>Tibetan people in the Indian subcontinent</title>
<description>Documents mentioning Tibetan people who live in countries of the Indian
subcontinent.</description>
<narrative>Relevant Documents contain information about Tibetan people living in
exile in countries of the Indian Subcontinent and mention reasons for the exile or
living conditions of the Tibetians. Countries of the Indian subcontinent are: India,
Pakistan, Bangladesh, Bhutan, Nepal and Sri Lanka.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/99-GC</identifier>
<title>Floods in European cities</title>
<description>Documents mentioning resons for and consequences of floods in
European cities</description>
<narrative>Relevant documents contain information about reasons and consequences
(damages, deaths, victims) of the floods and name the European city where the
flood occurred.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/100-GC</identifier>
<title>Natural disasters in the Western USA</title>
<description>Douments need to describe natural disasters in the Western
USA</description>
<narrative>Relevant documents report on natural disasters like earthquakes or
flooding which took place in Western states of the United States. To the Western
states belong California, Washington and Oregon.</narrative>
</topic>
</topics>
```

B.1 Classification of the GeoCLEF Topics

Table B.1: Classification of queries by query type (Overell PhD Thesis).

Counts	Query Type	Sample
80	Non-geo subject restricted to a place	Shark Attacks off Australia and California.
6	Geo subject with Non-Geo restriction	Cities near active volcanoes.
6	Geo subject restricted to a place	Cities along the Danube and the Rhine.
2	Non-geo subject related to a place	Independence movement in Quebec.
7	Non-geo subject related to regions	Water quality along coastlines of the Mediterranean Sea

Table B.2: Classification of queries by feature type.

Counts	Feature type	Sample
12	Geographic Feature	Cities near active volcanoes.
7	Body of water	Sea rescue in North Sea.
17	Continent	Trade Unions in Europe.
29	Country	Japanese Rice Imports.
9	State / County	Independence movement in Quebec.
6	City	Cities within 100km of Frankfurt.
0	Smaller than city.	
26	Imprecise region	Malaria in the tropics.

Table B.3: Classification of queries by location.

Counts	Location	Sample
9	Scotland	Walking holidays in Scotland.
1	California	Shark Attacks off Australia and California.
3	USA (excluding California)	Scientific research in New England Universities.
7	UK (excluding Scotland)	Roman cities in the UK and Germany.
46	Europe (exclud. the UK)	Trade Unions in Europe.
16	Asia	Solar or lunar eclipse in Southeast Asia.
7	Africa	Diamond trade in Angola and South Africa.
1	Australasia	Shark Attacks off Australia and California.
3	North America (excl. USA)	Fishing in Newfoundland and Greenland.
2	South America	Tourism in Northeast Brazil.
8	Other Specific Region	Shipwrecks in the Atlantic Ocean.
6	Other	Beaches with sharks.

APPENDIX C

Annex C: GeoCLEF PerQuery Results

This annex shows the GeoCLEF Per Query results of the experiments with the full GeoCLEF topics (100 topics) from GeoCLEF2005 to GeoCLEF2008 (consult Chapter 4 for more details about these experiments). The experiments were performed with the IR algorithms TF-IDF, BM25 and InL2 using the following configuration of topic's fields: Title (T), Title and Description (TD), and Title, Description and Narrative (TDN). The following table C indicates the legend of the different experiments.

T	topic number
base	baseline
S	stemming
L	lemmatization
LS	lemmatization+stemming
Bo1	Bose-Einstein1 Relevance Feedback
KL	Kullback-Leibler (KL) Relevance Feedback
G	GeoKB
SB1	Stemming+Bose-Einstein1
LSB1	lemmatization+stemming+Bose-Einstein1
LSB1G	lemmatization+stemming+Bose-Einstein1+GeoKB
LSG	lemmatization+stemming+GeoKB
SKL	stemming+Kullback-Leibler
LSKL	lemmatization+stemming+Kullback-Leibler
LSKLG	lemmatization+stemming+Kullback-Leibler+GeoKB

The next table C.1 contains the 100 GeoCLEF topics titles in order to help in the reading process of the perquery results.

Table C.1: GeoCLEF 2005-2008 topics' title.

01	Shark Attacks off Australia and California
02	Vegetable Exporters of Europe
03	AI in Latin America
04	Actions against the fur industry in Europe and the U.S.A.
05	Japanese Rice Imports
06	Oil Accidents and Birds in Europe
07	Trade Unions in Europe
08	Milk Consumption in Europe
09	Child Labor in Asia
10	Flooding in Holland and Germany
11	Roman cities in the UK and Germany
12	Cathedrals in Europe
13	Visits of the American president to Germany
14	Environmentally hazardous Incidents in the North Sea
15	Consequences of the genocide in Rwanda
16	Oil prospecting and ecological problems in Siberia and the Caspian Sea
17	American Troops in Sarajevo, Bosnia-Herzegovina
18	Walking holidays in Scotland
19	Golf tournaments in Europe
20	Wind power in the Scottish Islands
21	Sea rescue in North Sea
22	Restored buildings in Southern Scotland
23	Murders and violence in South-West Scotland
24	Factors influencing tourist industry in Scottish Highlands
25	Environmental concerns in and around the Scottish Trossachs
26	Wine regions around rivers in Europe
27	Cities within 100km of Frankfurt
28	Snowstorms in North America
29	Diamond trade in Angola and South Africa
30	Car bombings near Madrid
31	Combats and embargo in the northern part of Iraq
32	Independence movement in Quebec
33	International sports competitions in the Ruhr area
34	Malaria in the tropics
35	Credits to the former Eastern Bloc
36	Automotive industry around the Sea of Japan
37	Archeology in the Middle East
38	Solar or lunar eclipse in Southeast Asia
39	Russian troops in the southern Caucasus
40	Cities near active volcanoes
41	Shipwrecks in the Atlantic Ocean
42	Regional elections in Northern Germany
43	Scientific research in New England Universities
44	Arms sales in former Yugoslavia
45	Tourism in Northeast Brazil
46	Forest fires in Northern Portugal
47	Champions League games near the Mediterranean
48	Fishing in Newfoundland and Greenland
49	ETA in France
50	Cities along the Danube and the Rhine
51	Oil and gas extraction found between the UK and the Continent
52	Crime near St Andrews
53	Scientific research at east coast Scottish Universities
54	Damage from acid rain in northern Europe
55	Deaths caused by avalanches occurring in Europe, but not in the Alps
56	Lakes with monsters
57	Whisky making in the Scottlsh Islands
58	Travel problems at major airports near to London
59	Meetings of the Andean Community of Nations (CAN)
60	Casualties in fights in Nagorno-Karabakh
61	Airplane crashes close to Russian cities
62	OSCE meetings in Eastern Europe
63	Water quality along coastlines of the Mediterranean Sea
64	Sport events in the french speaking part of Switzerland
65	Free elections in Africa
66	Economy at the Bosphorus
67	F1 circuits where Ayrton Senna competed in 1994
68	Rivers with floods
69	Death on the Himalaya
70	Tourist attractions in Northern Italy
71	Social problems in greater Lisbon
72	Beaches with sharks
73	Events at St. Paul's Cathedral
74	Ship traffic around the Portuguese islands
75	Violation of human rights in Burma
76	Riots in South American prisons
77	Nobel prize winners from Northern European countries
78	Sport events in the Sahara
79	Invasion of Eastern Timor's capital by Indonesia
80	Politicians in exile in Germany
81	G7 summits in Mediterranean countries
82	Agriculture in the Iberian Peninsula
83	Demonstrations against terrorism in Northern Africa
84	Bombings in Northern Ireland
85	Nuclear tests in the South Pacific
86	Most visited sights in the capital of France and its vicinity
87	Unemployment in the OECD countries
88	Portuguese immigrant communities in the world
89	Trade fairs in Lower Saxony
90	Environmental pollution in European waters
91	Forest fires on Spanish islands
92	Islamic fundamentalists in Western Europe
93	Attacks in Japanese subways
94	Demonstrations in German cities
95	American troops in the Persian Gulf
96	Economic boom in Southeast Asia
97	Foreign aid in Sub-Saharan Africa
98	Tibetan people in the Indian subcontinent
99	Floods in European cities
100	Natural disasters in the Western USA

Table C.2: TFIDF per query results (T. only title tag used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6125	0.6753	**0.6845**	0.6746	0.6043	0.6048	0.5567	0.6626	0.6601	0.6596	0.6240	0.6623	0.6621	0.6581
02	0.0065	0.1528	0.0953	0.1531	0.0221	0.0226	0.0084	0.1360	0.1411	0.1466	**0.1536**	0.1250	0.1243	0.1287
03	0.0005	0.0007	0.0004	0.0006	0.0009	0.0008	0.0006	**0.0011**	**0.0011**	**0.0011**	0.0006	0.0010	0.0010	0.0010
04	0.1151	0.1433	0.1492	0.1443	**0.1848**	0.1762	0.0865	0.1648	0.1652	0.1246	0.1234	0.1679	0.1628	0.1236
05	0.5314	0.5102	0.5119	0.5063	0.6022	**0.6088**	0.5430	0.4968	0.4908	0.5245	0.5259	0.5206	0.5131	0.5511
06	0.2219	0.2528	0.1664	0.2494	0.2188	0.2283	0.2311	0.4048	0.3277	0.2842	0.2014	**0.4102**	0.3473	0.3061
07	0.3891	0.1730	0.1809	0.1727	0.4510	**0.4523**	0.3393	0.0872	0.0868	0.0721	0.1378	0.0876	0.0860	0.0723
08	0.0582	0.0568	0.0579	0.0572	0.0397	0.0379	**0.0803**	0.0409	0.0416	0.0633	0.0778	0.0407	0.0420	0.0668
09	0.2836	0.2915	0.2269	0.2432	0.2919	0.2991	0.3015	0.3101	0.2745	0.2812	0.2499	**0.3158**	0.2899	0.2992
10	0.4646	0.5759	0.4414	0.5681	0.6937	0.6102	0.4689	0.7465	0.7553	**0.7574**	0.5716	0.7287	0.7337	0.7359
11	0.0494	0.0495	0.0562	0.0561	0.1251	**0.1321**	0.0506	0.0708	0.0657	0.0659	0.0585	0.0690	0.0636	0.0638
12	0.0630	0.2536	0.2560	0.2546	0.1393	0.1327	0.0606	0.2912	0.2957	**0.3231**	0.2772	0.2924	0.2934	0.3220
13	0.1617	0.3333	0.3422	0.3333	0.5026	**0.5063**	0.3119	0.4307	0.4204	0.4209	0.3372	0.4395	0.4289	0.4297
14	0.1230	0.2570	0.1459	0.2580	0.4982	0.5053	0.1623	0.6337	0.6329	0.4326	0.2712	0.6330	**0.6338**	0.4279
15	0.6766	0.7239	0.6757	0.7219	0.7574	0.7539	0.6875	**0.8028**	0.7798	0.7808	0.7309	0.7997	0.7812	0.7819
16	0.8214	0.8385	0.8286	0.8306	0.8975	0.8862	0.7959	0.9110	0.9110	0.8042	0.7718	0.9159	**0.9165**	0.8096
17	0.4515	0.4462	0.4439	0.4474	0.4700	**0.4760**	0.4355	0.4487	0.4508	0.4352	0.4304	0.4577	0.4596	0.4435
18	0.1483	0.3105	0.2726	0.3106	0.1101	0.1088	0.1747	0.2927	0.2935	0.3293	0.3234	0.3172	0.3236	**0.3603**
19	0.1142	0.1365	0.1227	0.1384	0.1939	0.1823	0.1484	0.1925	0.2005	**0.2397**	0.1809	0.1793	0.1808	0.2199
20	0.2015	**0.3315**	0.2892	0.3198	0.1760	0.1842	0.0030	0.2197	0.1972	0.0071	0.0028	0.2252	0.2253	0.0076
21	0.4898	0.5468	0.5465	0.5455	0.5230	0.5119	0.4802	0.4396	0.4724	0.5899	0.5160	0.4554	0.4840	**0.5983**
22	0.2080	0.4381	0.2760	0.4287	0.3254	0.3194	0.2310	**0.5105**	0.4688	0.4894	0.4826	0.5060	0.4632	0.4854
23	0.0059	0.0218	0.0226	0.0219	0.0135	0.0138	0.0173	0.0535	0.0543	**0.1553**	0.0385	0.0502	0.0507	0.1549
24	0.4916	0.5162	0.4950	0.5150	0.5996	0.6029	0.4834	**0.6456**	0.6440	0.6270	0.5185	0.6412	0.6394	0.6223
25	0.3056	0.3687	0.3687	0.3778	0.8095	0.6984	0.3333	0.7917	0.8095	**0.8333**	0.4028	0.7917	0.7917	0.8095
26	0.0088	0.2413	0.2532	0.2418	0.0176	0.0167	0.0119	0.2719	0.2723	**0.2920**	0.2512	0.2495	0.2496	0.2875
27	0.0190	0.0200	0.0201	0.0200	0.0375	**0.0403**	0.0190	0.0107	0.0133	0.0133	0.0200	0.0140	0.0176	0.0176
28	0.0275	0.0526	0.0523	0.0523	0.0909	0.0913	0.0004	**0.0964**	0.0914	0.0037	0.0018	0.0957	0.0916	0.0043
29	0.1884	**0.2410**	0.2395	0.2381	0.1098	0.1055	0.1788	0.1114	0.1124	0.1063	0.2245	0.1352	0.1358	0.1310
30	0.3501	**1.0000**	0.4723	**1.0000**	0.1585	0.1363	0.5225	0.9444	0.9444	0.6393	0.6728	0.9306	0.9583	0.6394
31	0.3825	0.3623	0.3517	0.3602	0.3966	0.3899	0.4069	0.4083	0.4097	**0.4165**	0.3859	0.3986	0.3976	0.4043
32	0.6707	0.7047	0.6668	0.7058	0.9019	0.8897	0.6560	0.9167	0.9078	0.8852	0.6843	**0.9168**	0.9160	0.8934
33	0.0014	0.0013	0.0016	0.0013	0.0027	**0.0029**	0.0012	0.0019	0.0019	0.0017	0.0012	0.0020	0.0020	0.0019
34	0.3056	0.5667	0.3056	0.5667	0.4444	0.5139	**0.7500**	0.6556	0.6465	0.7381	0.7436	0.6556	0.6556	0.7436
35	0.0052	0.0040	0.0041	0.0040	0.0083	0.0081	0.0052	0.0089	0.0089	0.0089	0.0040	0.0090	**0.0092**	**0.0092**
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0121	0.0247	0.0116	0.0231	**0.0702**	0.0643	0.0022	0.0698	0.0625	0.0127	0.0024	0.0688	0.0605	0.0115
38	0.0667	0.0435	0.0435	0.0435	**0.0769**	**0.0769**	0.0015	0.0476	0.0500	0.0000	0.0011	0.0526	0.0588	0.0000
39	0.0691	0.0695	0.0697	0.0694	0.0409	0.0407	0.0721	0.0404	0.0405	0.0409	**0.0724**	0.0399	0.0400	0.0404
40	0.0000	0.2106	0.1495	0.2109	0.1166	0.1148	0.0000	**0.2394**	0.2363	0.2363	0.2109	0.2389	0.2382	0.2382
41	0.0003	0.0003	0.0003	0.0003	0.0003	0.0000	0.0009	**0.0053**	0.0006	0.0008	0.0008	0.0036	0.0005	0.0008
42	0.0486	0.0672	0.1288	0.0672	0.1769	**0.1909**	0.0943	0.0123	0.0121	0.0536	0.1357	0.0167	0.0166	0.0629
43	0.0020	0.0156	0.0101	0.0152	0.0032	0.0033	0.0029	0.0108	0.0105	0.0233	0.0157	0.0108	0.0105	**0.0259**
44	0.2115	0.2169	0.1975	0.2190	**0.2267**	0.2217	0.2170	0.1824	0.1847	0.1730	0.2173	0.1872	0.1853	0.1741
45	0.0084	0.0084	0.0085	0.0085	0.0781	0.0726	0.0140	0.0863	**0.0879**	**0.0879**	0.0140	0.0822	0.0834	0.0834
46	0.8095	0.7167	0.7193	0.7167	0.5222	0.5193	**1.0000**	0.7292	0.7292	0.7917	0.7667	0.7255	0.7255	0.8333
47	0.0527	0.0325	0.0696	0.0305	0.0725	0.0662	0.0527	**0.0963**	0.0601	0.0601	0.0305	0.0911	0.0485	0.0485
48	0.7331	0.7254	0.7332	0.7258	0.8970	0.9038	0.5586	0.9181	0.9200	0.5978	0.5605	0.9196	**0.9204**	0.5976
49	**0.6111**	**0.6111**	**0.6111**	**0.6111**	0.2679	0.2679	0.5000	0.2679	0.2679	0.5000	0.5000	0.2679	0.2679	0.5000
50	0.1850	0.2511	0.2782	0.2495	0.2430	0.2442	0.3507	0.1703	0.1719	**0.4697**	0.4271	0.1742	0.1767	0.4525
51	0.4088	0.3786	0.3941	0.3725	0.4133	0.4225	0.4781	**0.6126**	0.4944	0.5508	0.4409	0.6008	0.4875	0.5497
52	0.0265	0.0335	0.0254	0.0314	0.0280	0.0235	0.0284	**0.0827**	0.0596	0.0604	0.0389	0.0676	0.0596	0.0603
53	0.0948	0.1433	0.1538	0.1448	0.1178	0.1199	0.1201	0.1217	0.1222	0.1873	0.1894	0.1365	0.1373	**0.1951**
54	0.1018	0.1169	0.1006	0.1172	0.0996	0.1036	0.1278	0.1240	0.1255	0.1455	0.1341	0.1202	0.1300	**0.1532**
55	0.0111	0.1054	0.1374	0.1066	0.1519	0.1487	0.0136	0.2091	**0.2092**	0.2069	0.1299	0.2044	0.2058	0.2041
56	0.0003	0.2316	0.0258	0.2323	0.0006	0.0006	0.0003	0.3392	**0.3523**	**0.3523**	0.2323	0.3014	0.2956	0.2956
57	0.1023	0.1520	0.0733	0.1520	0.1232	0.1173	0.1023	0.2261	**0.2286**	**0.2286**	0.1520	0.2233	0.2253	0.2253
58	0.0254	0.0193	0.0205	0.0194	0.0367	0.0373	0.0410	0.0348	0.0349	**0.0554**	0.0337	0.0311	0.0316	0.0531
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7520	0.7704	0.7624	0.7721	0.7630	0.7676	0.7527	0.7837	0.7814	0.7814	0.7721	0.7864	**0.7897**	**0.7897**
61	0.0975	0.1391	0.1467	0.1445	0.0295	0.0255	0.1675	0.0709	0.0730	0.5167	0.3576	0.0699	0.0728	**0.5299**
62	0.2812	0.3352	0.2718	0.3350	0.2813	0.2831	0.2839	0.3305	0.3117	0.3118	**0.3377**	0.3320	0.3222	0.3225
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0046	0.0135	0.0099	0.0134	0.0047	0.0048	0.0071	0.0115	0.0076	0.0108	**0.0181**	0.0120	0.0074	0.0109
65	0.3298	**0.4242**	0.4224	0.4191	0.3274	0.3237	0.3394	0.3847	0.3884	0.3884	0.4235	0.3832	0.3857	0.3857
66	**0.3436**	0.3425	0.3430	0.3430	0.1825	0.1291	**0.3436**	0.1863	0.1307	0.1307	0.3430	0.1042	0.0698	0.0698
67	0.2922	0.3331	0.3302	0.3335	0.3769	0.3796	0.2916	**0.4284**	0.4241	0.4241	0.3332	0.4225	0.4206	0.4206
68	0.3014	0.7594	0.7385	0.7589	0.4282	0.4230	0.3014	0.7676	0.7683	0.7683	0.7589	0.7681	**0.7691**	**0.7691**
69	0.0001	0.0572	0.0003	0.0569	0.0001	0.0001	0.0001	0.2836	**0.2890**	**0.2890**	0.0569	0.2806	0.2792	0.2792
70	0.0280	0.0313	0.0209	0.0312	0.0112	0.0105	**0.0859**	0.0143	0.0144	0.0731	0.0715	0.0110	0.0111	0.0705
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.3310	0.5670	**0.5772**	0.5677	0.4446	0.4435	0.3310	0.5145	0.5172	0.5172	0.5677	0.5162	0.5200	0.5200
73	**0.0223**	0.0205	0.0207	0.0207	0.0202	0.0202	0.0100	0.0160	0.0162	0.0053	0.0097	0.0162	0.0163	0.0053
74	0.0196	0.4167	0.3426	0.3939	0.0093	0.0076	0.0600	0.3889	0.2407	0.2679	0.4340	**0.5556**	0.2500	0.2834
75	0.2299	0.2744	0.2743	0.2749	0.5540	0.5103	0.4748	**0.7089**	0.7086	0.6782	0.4289	0.6815	0.6805	0.6607
76	0.0000	**0.4316**	0.2319	**0.4316**	0.0000	0.0000	0.0000	0.1370	0.1370	0.1370	**0.4316**	0.1837	0.1837	0.1837
77	0.3768	0.4016	**0.4023**	0.4003	0.3832	0.3828	0.3768	0.3771	0.3757	0.3757	0.4003	0.3869	0.3863	0.3863
78	0.0205	0.1269	0.1111	0.1269	0.0324	0.0272	0.0622	**0.2306**	0.2286	0.2071	0.1031	0.2253	0.2255	0.2027
79	0.8028	0.8028	0.8028	0.8028	**0.9306**	0.9107	0.8028	0.9107	0.9107	0.9107	0.8028	0.9107	0.9107	0.9107
80	0.0347	0.0859	0.0860	0.0859	0.1169	0.1042	0.0618	0.2500	0.2500	**0.2917**	0.1303	0.2500	0.2679	0.2679
81	0.0366	0.0586	0.0578	0.0587	0.0719	0.0729	0.0366	0.0788	0.0789	0.0789	0.0587	0.0797	**0.0799**	**0.0799**
82	0.0018	0.0004	0.0018	0.0004	0.0032	0.0028	0.0015	**0.0034**	**0.0034**	0.0029	0.0003	0.0026	0.0026	0.0022
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.0577	**0.2488**	0.1006	0.2477	0.0352	0.0357	0.0582	0.0610	0.0645	0.0671	0.2477	0.0631	0.0649	0.0689
85	0.3865	0.3400	0.3612	0.3379	**0.4782**	0.4763	0.3420	0.4481	0.4448	0.3744	0.2990	0.4234	0.4195	0.3574
86	0.0020	0.0028	0.0020	0.0028	0.0004	0.0004	0.0048	0.0017	0.0017	0.0029	**0.0050**	0.0018	0.0019	0.0029
87	0.2329	0.2685	0.2704	0.2677	0.2331	0.2286	0.2329	0.2939	0.2933	0.2933	0.2677	**0.2962**	0.2958	0.2958
88	0.0799	0.2963	0.2374	0.2889	0.1664	0.1724	0.1292	0.3641	0.3520	0.2581	0.2514	**0.3675**	0.3533	0.2557
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90	0.0781	0.1439	0.0742	**0.1461**	0.0460	0.0470	0.0781	0.1172	0.1177	0.1177	**0.1461**	0.1150	0.1171	0.1171
91	0.1250	0.5051	0.1306	0.5052	0.0235	0.0228	0.2500	0.2575	0.2576	0.5012	0.5010	0.5083	**0.5085**	0.5012
92	0.0891	0.0707	0.0869	0.0712	0.1323	0.1378	0.1839	0.1198	0.1179	0.2401	0.1868	0.1164	0.1166	**0.2483**
93	0.3525	0.8053	0.8052	0.8046	0.8467	0.8559	0.3929	0.8893	0.8888	0.8710	0.8231	**0.8957**	0.8940	0.8769
94	0.1157	0.4136	0.2799	0.4135	0.0498	0.0514	0.1220	0.0516	0.0503	0.0506	**0.4191**	0.0469	0.0465	0.0468
95	0.4469	0.4518	0.4628	0.4523	**0.6677**	0.6558	0.3305	0.6662	0.6650	0.3826	0.3208	0.6559	0.6512	0.3782
96	0.2104	0.2718	0.2170	0.2736	0.2471	0.2428	0.2256	0.2890	0.2860	0.2614	**0.2925**	0.2755	0.2750	0.2531
97	0.0359	0.0356	0.0360	0.0356	**0.0409**	0.0399	0.0328	0.0407	0.0408	0.0372	0.0326	0.0390	0.0392	0.0355
98	0.2083	0.2485	0.2101	0.2481	0.3482	0.3342	0.2771	0.7351	0.7361	**0.7848**	0.3839	0.7495	0.7495	0.7804
99	0.1020	0.1350	0.1005	0.1365	0.2140	**0.2179**	0.1020	0.1645	0.1666	0.1666	0.1365	0.1699	0.1730	0.1730
100	0.0110	0.0162	0.0178	0.0161	0.0343	0.0323	0.0037	0.0414	**0.0426**	0.0210	0.0093	0.0375	0.0381	0.0194
all	0.1938	0.2642	0.2333	0.2631	0.2372	0.2339	0.2088	0.2926	0.2869	0.2899	0.2647	**0.2954**	0.2893	0.2898

Table C.3: BM25 per query results (T. only title tag used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6158	0.6753	**0.6853**	0.6753	0.6020	0.6031	0.5616	0.6635	0.6616	0.6618	0.6246	0.6667	0.6674	0.6646
02	0.0063	0.1533	0.0946	0.1537	0.0229	0.0220	0.0087	0.1399	0.1465	0.1530	**0.1542**	0.1295	0.1316	0.1371
03	0.0005	0.0006	0.0004	0.0005	0.0010	0.0008	0.0006	**0.0013**	0.0012	**0.0013**	0.0006	0.0010	0.0010	0.0010
04	0.1262	0.1574	0.1608	0.1575	**0.1868**	0.1755	0.0960	0.1645	0.1628	0.1229	0.1320	0.1662	0.1640	0.1255
05	0.5328	0.5048	0.5121	0.5009	0.6007	**0.6067**	0.5446	0.4877	0.4769	0.5089	0.5230	0.5124	0.4955	0.5312
06	0.2203	0.2534	0.1798	0.2499	0.1950	0.2111	0.2311	0.3956	0.3405	0.2973	0.2023	**0.4121**	0.3391	0.2952
07	0.3898	0.1667	0.1766	0.1663	**0.4671**	0.4661	0.3408	0.0844	0.0784	0.0654	0.1331	0.0830	0.0799	0.0674
08	0.0578	0.0563	0.0574	0.0569	0.0395	0.0376	**0.0801**	0.0404	0.0410	0.0626	0.0793	0.0399	0.0408	0.0627
09	0.2792	0.2840	0.2060	0.2037	0.2857	0.2973	0.2956	0.3047	0.2704	0.2747	0.2087	**0.3156**	0.2787	0.2836
10	0.4651	0.5755	0.4421	0.5677	0.7057	0.6290	0.4696	0.7465	0.7539	**0.7561**	0.5713	0.7280	0.7329	0.7351
11	0.0511	0.0410	0.0418	0.0403	0.1255	**0.1322**	0.0523	0.0670	0.0247	0.0247	0.0424	0.0625	0.0312	0.0312
12	0.0635	0.2569	0.2590	0.2585	0.1408	0.1355	0.0610	0.2939	0.2960	**0.3230**	0.2810	0.2925	0.2934	0.3213
13	0.1264	0.3258	0.3406	0.3260	0.4501	0.4579	0.2176	0.4396	0.4395	0.4401	0.3287	**0.4636**	0.4603	0.4608
14	0.1178	0.2463	0.1408	0.2476	0.4970	0.5094	0.1628	0.6313	0.6345	0.4333	0.2721	**0.6364**	0.6349	0.4329
15	0.6767	0.7234	0.6757	0.7221	0.7582	0.7568	0.6877	**0.7836**	0.7811	0.7821	0.7310	0.7820	0.7797	0.7802
16	0.8215	0.8386	0.8287	0.8386	0.8992	0.8889	0.7959	0.9128	0.9128	0.8042	0.7801	0.9145	**0.9151**	0.8042
17	0.4351	0.4266	0.4275	0.4276	0.4642	**0.4666**	0.4199	0.4402	0.4180	0.4030	0.4115	0.4447	0.4204	0.4055
18	0.1401	0.2916	0.2590	0.2930	0.1089	0.1064	0.1715	0.2768	0.2803	0.3151	0.3141	0.3062	0.3120	**0.3485**
19	0.1134	0.1355	0.1223	0.1374	0.1999	0.1878	0.1477	0.1953	0.2042	**0.2439**	0.1805	0.1817	0.1822	0.2214
20	0.2017	0.3290	0.3006	**0.3293**	0.1734	0.2018	0.0031	0.1979	0.1969	0.0073	0.0033	0.2222	0.2179	0.0080
21	0.4657	0.5419	0.5455	0.5440	0.5187	0.5039	0.4859	0.4207	0.4639	0.5858	0.5216	0.4464	0.4744	**0.5891**
22	0.2045	0.4263	0.2697	0.4083	0.2981	0.3016	0.2353	0.4967	0.4405	0.4611	0.4712	**0.5031**	0.4373	0.4639
23	0.0065	0.0246	0.0244	0.0266	0.0129	0.0129	0.0211	0.0566	0.0562	**0.1714**	0.0604	0.0520	0.0524	0.1672
24	0.4850	0.5088	0.4884	0.5064	0.5973	0.5997	0.4778	**0.6403**	0.6398	0.6234	0.5109	0.6396	0.6397	0.6234
25	0.3056	0.3687	0.3687	0.3778	0.8095	0.8095	0.3333	0.7778	0.8095	**0.8333**	0.4028	0.7917	0.8095	**0.8333**
26	0.0089	0.2428	0.2540	0.2430	0.0180	0.0170	0.0121	0.2841	0.2845	**0.3099**	0.2529	0.2688	0.2505	0.2886
27	0.0197	0.0199	0.0200	0.0198	0.0375	**0.0438**	0.0197	0.0101	0.0101	0.0101	0.0198	0.0125	0.0122	0.0122
28	0.0302	0.0550	0.0538	0.0537	**0.1053**	0.1040	0.0004	0.0964	0.0934	0.0044	0.0019	0.0972	0.0960	0.0044
29	0.1891	0.2435	**0.2456**	0.2449	0.1119	0.1075	0.1789	0.1125	0.1133	0.1065	0.2304	0.1369	0.1379	0.1321
30	0.3446	**1.0000**	0.4854	**1.0000**	0.1447	0.1289	0.5058	0.9167	0.9167	0.6393	0.6727	0.9306	0.9444	0.6393
31	0.3742	0.3563	0.3494	0.3554	0.3984	0.3998	0.3981	0.4115	0.4076	**0.4147**	0.3810	0.4018	0.4016	0.4084
32	0.6711	0.7078	0.6672	0.7066	0.9027	0.8902	0.6560	0.9095	**0.9175**	0.8950	0.6846	0.9150	0.9160	0.8934
33	0.0013	0.0013	0.0016	0.0013	**0.0024**	**0.0024**	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0013
34	0.3056	0.5667	0.3056	0.5667	0.4333	0.4583	**0.7500**	0.4798	0.4798	0.7381	0.7436	0.6556	0.6465	0.7381
35	0.0051	0.0041	0.0041	0.0041	0.0086	0.0082	0.0051	0.0091	0.0092	0.0092	0.0041	0.0092	**0.0093**	**0.0093**
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0143	0.0268	0.0136	0.0245	**0.0760**	0.0683	0.0029	0.0732	0.0637	0.0127	0.0025	0.0695	0.0644	0.0127
38	0.0667	0.0435	0.0435	0.0435	0.0714	**0.0769**	0.0015	0.0435	0.0476	0.0000	0.0011	0.0526	0.0526	0.0000
39	0.0697	0.0708	0.0712	0.0705	0.0408	0.0405	0.0725	0.0400	0.0400	0.0404	**0.0734**	0.0395	0.0395	0.0399
40	0.0000	0.2370	0.1653	0.2366	0.1270	0.1251	0.0000	0.2413	0.2374	0.2374	0.2366	**0.2418**	0.2401	0.2401
41	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0009	**0.0066**	0.0007	0.0009	0.0008	0.0046	0.0006	0.0008
42	0.0486	0.0682	0.1455	0.0672	0.1769	**0.2159**	0.0911	0.0105	0.0105	0.0524	0.1345	0.0147	0.0149	0.0577
43	0.0025	0.0133	0.0076	0.0128	0.0033	0.0033	0.0033	0.0106	0.0103	0.0224	0.0143	0.0107	0.0102	**0.0229**
44	0.2136	0.2171	0.1988	0.2138	**0.2201**	0.2181	0.2174	0.1772	0.1764	0.1644	0.2115	0.1788	0.1784	0.1669
45	0.0084	0.0084	0.0085	0.0084	0.0821	0.0760	0.0140	0.0897	**0.0903**	**0.0903**	0.0140	0.0841	0.0865	0.0865
46	0.8095	0.7167	0.7222	0.7222	0.3069	0.5222	**1.0000**	0.7333	0.5769	0.6111	0.7667	0.7333	0.7500	0.8333
47	0.0546	0.0335	0.0642	0.0323	0.0720	0.0657	0.0546	**0.0970**	0.0943	0.0943	0.0323	0.0951	0.0939	0.0939
48	0.7331	0.7233	0.7332	0.7235	0.8990	0.9064	0.5586	0.9203	0.9219	0.5980	0.5605	0.9225	**0.9228**	0.5977
49	**0.6111**	**0.6111**	**0.6111**	**0.6111**	0.2679	0.2679	0.5000	0.2917	0.2917	0.5000	0.5000	0.2679	0.2679	0.5000
50	0.1873	0.2402	0.2723	0.2402	0.2471	0.2490	0.3539	0.1936	0.1905	0.5042	0.4168	0.1944	0.1938	**0.5051**
51	0.4097	0.3813	0.4075	0.3816	0.4515	0.4534	0.4787	0.5585	0.5028	**0.5608**	0.4502	0.5447	0.4967	0.5595
52	0.0265	0.0335	0.0216	0.0315	0.0262	0.0232	0.0284	**0.0919**	0.0596	0.0604	0.0389	0.0767	0.0596	0.0603
53	0.0920	0.1356	0.1483	0.1369	0.1150	0.1169	0.1208	0.1151	0.1155	0.1844	0.1835	0.1291	0.1297	**0.1926**
54	0.1151	0.1182	0.1060	0.1186	0.0972	0.1041	0.1374	0.1234	0.1250	0.1447	0.1365	0.1222	0.1323	**0.1545**
55	0.0120	0.1252	0.1421	0.1270	0.1652	0.1689	0.0145	0.2080	**0.2097**	0.2068	0.1454	0.2042	0.2053	0.2027
56	0.0003	0.2321	0.0254	0.2323	0.0006	0.0006	0.0003	**0.3569**	0.3553	0.3553	0.2323	0.3052	0.3121	0.3121
57	0.1075	0.1542	0.0733	0.1521	0.1383	0.1267	0.1075	0.2294	**0.2309**	**0.2309**	0.1521	0.2272	0.2295	0.2295
58	0.0259	0.0183	0.0216	0.0179	0.0522	0.0481	0.0419	0.0454	0.0453	**0.0674**	0.0322	0.0391	0.0406	0.0630
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7520	0.7708	0.7628	0.7737	0.7596	0.7629	0.7527	0.7812	0.7820	0.7820	0.7737	**0.7894**	0.7875	0.7875
61	0.0982	0.1456	0.1527	0.1491	0.0289	0.0265	0.1703	0.0798	0.0818	**0.5167**	0.3866	0.0806	0.0821	0.5152
62	0.2812	0.3350	0.2735	0.3348	0.2826	0.2834	0.2838	0.3305	0.3168	0.3168	**0.3372**	0.3318	0.3179	0.3179
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0046	0.0146	0.0101	0.0144	0.0054	0.0053	0.0076	0.0079	0.0082	0.0106	**0.0185**	0.0087	0.0089	0.0119
65	0.3289	**0.4244**	**0.4244**	0.4184	0.3236	0.3214	0.3378	0.3781	0.3802	0.3803	0.4217	0.3783	0.3804	0.3805
66	**0.3436**	0.3425	0.3430	0.3430	0.0992	0.1287	**0.3436**	0.1869	0.1307	0.1307	0.3430	0.1035	0.0579	0.0579
67	0.2919	0.3327	0.3304	0.3333	0.3765	0.3765	0.2913	**0.4271**	0.4213	0.4213	0.3330	0.4257	0.4247	0.4247
68	0.3015	0.7586	0.7394	0.7587	0.4384	0.4362	0.3015	0.7678	**0.7688**	**0.7688**	0.7587	0.7683	0.7684	0.7684
69	0.0001	0.0581	0.0003	0.0582	0.0001	0.0001	0.0001	**0.2920**	0.2911	0.2911	0.0582	0.2877	0.2869	0.2869
70	0.0279	0.0326	0.0207	0.0329	0.0112	0.0106	**0.0878**	0.0169	0.0169	0.0756	0.0718	0.0150	0.0151	0.0737
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.3311	0.5694	**0.5811**	0.5733	0.4631	0.4462	0.3311	0.5151	0.4934	0.4934	0.5733	0.5162	0.5201	0.5201
73	**0.0222**	0.0207	0.0208	0.0208	0.0203	0.0202	0.0100	0.0157	0.0160	0.0053	0.0097	0.0159	0.0160	0.0053
74	0.0196	0.4000	0.3431	0.4000	0.0090	0.0085	0.0600	0.3889	0.3889	0.3981	0.4499	0.5556	0.5556	**0.5627**
75	0.2389	0.3186	0.3020	0.3200	0.5750	0.5266	0.4791	0.7127	**0.7159**	0.6796	0.4362	0.7030	0.6984	0.6693
76	0.0000	**0.4429**	0.2412	0.4400	0.0000	0.0000	0.0000	0.1459	0.1418	0.1418	0.4400	0.1955	0.1821	0.1821
77	0.3815	0.4046	**0.4075**	0.4063	0.3742	0.3804	0.3815	0.3760	0.3760	0.3760	0.4063	0.3852	0.3863	0.3863
78	0.0241	0.1268	0.1055	0.1268	0.0453	0.0431	0.0622	0.2306	**0.2308**	0.2087	0.1031	0.2305	0.2306	0.2094
79	0.8028	0.8028	0.8028	0.7974	**0.9306**	0.9107	0.8028	**0.9306**	**0.9306**	**0.9306**	0.7974	0.9107	0.9107	0.9107
80	0.0366	0.0859	0.0860	0.0859	0.1167	0.1083	0.0688	0.2500	0.2500	**0.2917**	0.1303	0.2500	0.2500	0.2679
81	0.0368	0.0571	0.0572	0.0580	0.0745	0.0744	0.0368	0.0807	0.0806	0.0806	0.0580	0.0815	**0.0816**	**0.0816**
82	0.0018	0.0004	0.0018	0.0004	0.0034	0.0031	0.0015	**0.0041**	**0.0041**	0.0035	0.0003	0.0030	0.0030	0.0025
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.0576	**0.2499**	0.1010	0.2482	0.0363	0.0357	0.0582	0.0606	0.0604	0.0633	0.2482	0.0609	0.0603	0.0640
85	0.3881	0.3400	0.3633	0.3391	**0.4858**	0.4802	0.3441	0.4588	0.4564	0.3812	0.3000	0.4336	0.4293	0.3646
86	0.0020	0.0027	0.0019	0.0028	0.0004	0.0003	0.0048	0.0015	0.0015	0.0027	**0.0050**	0.0016	0.0016	0.0027
87	0.2339	0.2811	0.2828	0.2804	0.2359	0.2331	0.2339	0.3046	0.3040	0.3040	0.2805	**0.3075**	**0.3075**	**0.3075**
88	0.0824	**0.3284**	0.2551	0.3280	0.1654	0.1700	0.1370	0.3217	0.3214	0.2384	0.2655	0.3276	0.3264	0.2409
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90	0.0780	0.1434	0.0719	**0.1447**	0.0462	0.0473	0.0780	0.1173	0.1184	0.1184	**0.1447**	0.1151	0.1158	0.1158
91	0.1250	0.5049	0.1721	**0.5050**	0.0246	0.0258	0.2500	0.1744	0.1746	0.5011	0.5000	0.2579	0.1750	0.5012
92	0.0901	0.0718	0.1111	0.0720	0.1466	0.1419	0.1867	0.1232	0.1223	0.2434	0.1912	0.1199	0.1206	**0.2535**
93	0.3542	0.8004	0.8005	0.7993	0.8505	0.8591	0.3951	0.8859	0.8861	0.8692	0.8210	**0.8929**	0.8917	0.8750
94	0.1157	0.4136	0.2688	0.4136	0.0491	0.0504	0.1222	0.0774	0.0615	0.0617	**0.4185**	0.1085	0.0577	0.0579
95	0.4432	0.4538	0.4587	0.4527	0.6743	0.6612	0.3294	**0.6776**	0.6772	0.3908	0.3268	0.6745	0.6751	0.3891
96	0.2117	0.2744	0.2204	0.2757	0.2492	0.2443	0.2269	0.2914	0.2869	0.2602	**0.2937**	0.2757	0.2756	0.2527
97	0.0358	0.0356	0.0359	0.0356	0.0409	0.0401	0.0327	0.0407	**0.0412**	0.0381	0.0326	0.0397	0.0398	0.0366
98	0.2195	0.2767	0.2207	0.2791	0.6896	0.7180	0.3047	0.7570	0.8002	**0.8418**	0.4205	0.7530	0.7908	0.8389
99	0.1084	0.1740	0.1403	0.1753	0.2138	**0.2193**	0.1085	0.1939	0.1942	0.1942	0.1753	0.1907	0.1952	0.1952
100	0.0111	0.0160	0.0176	0.0156	0.0365	0.0342	0.0037	0.0258	**0.0427**	0.0220	0.0093	0.0226	0.0389	0.0213
all	0.1935	0.2653	0.2353	0.2643	0.2384	0.2399	0.2086	0.2898	0.2854	0.2906	0.2661	**0.2940**	0.2899	0.2939

Table C.4: InL2 per query results (T. only title tag used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6185	0.6633	**0.6737**	0.6645	0.5955	0.5950	0.5408	0.6469	0.6432	0.6329	0.6154	0.6460	0.6487	0.6467
02	0.0080	0.1694	0.1034	0.1724	0.0240	0.0251	0.0094	0.1635	0.1703	**0.1739**	0.1727	0.1420	0.1428	0.1458
03	0.0006	0.0007	0.0005	0.0006	0.0010	0.0009	0.0006	**0.0013**	0.0012	**0.0013**	0.0006	0.0010	0.0010	0.0010
04	0.1140	0.1425	0.1543	0.1426	0.1779	0.1720	0.0885	0.2082	0.2067	0.1660	0.1242	**0.2136**	0.2036	0.1640
05	0.5337	0.5112	0.5175	0.5095	0.6031	**0.6120**	0.5459	0.4861	0.4806	0.5136	0.5300	0.5062	0.5043	0.5383
06	0.1792	0.2189	0.1611	0.2249	0.1930	0.2244	0.1887	0.3373	**0.3552**	0.2856	0.1850	0.3476	0.3518	0.2894
07	0.3819	0.1566	0.1660	0.1562	0.4404	**0.4447**	0.3295	0.0802	0.0740	0.0650	0.1260	0.0782	0.0667	0.0624
08	0.0612	0.0593	0.0606	0.0598	0.0447	0.0438	**0.0849**	0.0455	0.0462	0.0704	0.0828	0.0443	0.0451	0.0692
09	0.2781	0.2846	0.2248	0.2253	0.2825	0.2996	0.2935	0.2990	0.2708	0.2752	0.2296	**0.3090**	0.2803	0.2866
10	0.4682	0.5776	0.4889	0.5803	0.6997	0.6311	0.4712	0.7635	0.7821	**0.7837**	0.5854	0.7438	0.7526	0.7544
11	0.0841	0.0617	0.0615	0.0619	0.1333	0.1327	0.0853	0.1560	0.1566	**0.1591**	0.0639	0.1548	0.1548	0.1571
12	0.0697	0.2741	0.2757	0.2755	0.1446	0.1401	0.0665	0.2949	0.2989	0.3262	0.2963	0.2979	0.2993	**0.3279**
13	0.1245	0.3283	0.3402	0.3300	0.4586	**0.4744**	0.2152	0.4160	0.3408	0.3437	0.3329	0.4708	0.4705	0.4712
14	0.1187	0.2525	0.1419	0.2569	0.4954	0.5025	0.1602	0.6232	0.6238	0.4255	0.2716	**0.6317**	0.6313	0.4286
15	0.6735	0.7214	0.6728	0.7185	0.7743	0.7713	0.6855	**0.8057**	0.8030	0.8023	0.7282	0.8033	0.7991	0.7979
16	0.8168	0.8316	0.8240	0.8315	0.8818	0.8771	0.7903	0.9094	0.9102	0.8041	0.7717	**0.9126**	**0.9126**	0.8040
17	0.4546	0.4508	0.4490	0.4526	0.4847	**0.4929**	0.4399	0.4624	0.4636	0.4484	0.4368	0.4701	0.4713	0.4555
18	0.1470	0.3110	0.2841	0.3130	0.1131	0.1150	0.1733	0.2922	0.2894	0.3206	0.3300	0.3201	0.3260	**0.3614**
19	0.1214	0.1516	0.1385	0.1536	0.2209	0.2046	0.1544	0.2053	0.2137	**0.2499**	0.1917	0.1937	0.1952	0.2320
20	0.1987	**0.3097**	0.2810	0.3091	0.1735	0.2011	0.0028	0.1985	0.1978	0.0071	0.0028	0.2770	0.2742	0.0072
21	0.4751	0.5517	0.5563	0.5556	0.5012	0.4972	0.4885	0.4789	0.4671	0.5875	0.5366	0.4739	0.4656	**0.5909**
22	0.2188	0.4351	0.2886	0.4136	0.3185	0.3067	0.2485	**0.5556**	0.4773	0.5264	0.4673	0.5439	0.4808	0.5215
23	0.0074	0.0267	0.0269	0.0270	0.0135	0.0150	0.0231	0.0556	0.0562	**0.1658**	0.0588	0.0513	0.0512	0.1644
24	0.4740	0.4938	0.4728	0.4939	0.5950	0.6007	0.4681	0.6303	0.6290	0.6145	0.4956	**0.6330**	0.6328	0.6167
25	0.2833	0.3619	0.3619	0.3619	0.7917	0.7917	0.3063	0.7917	0.7917	**0.8095**	0.3869	0.7917	0.7917	**0.8095**
26	0.0116	0.2582	0.2666	0.2644	0.0220	0.0211	0.0145	0.3263	0.3268	**0.3540**	0.2788	0.3214	0.2847	0.3112
27	0.0190	0.0200	0.0202	0.0201	0.0375	**0.0403**	0.0190	0.0127	0.0127	0.0127	0.0201	0.0174	0.0175	0.0175
28	0.0319	0.0586	0.0587	0.0587	0.1065	0.1072	0.0004	0.1462	0.1468	0.0052	0.0019	0.1493	**0.1586**	0.0052
29	0.1848	0.2353	**0.2377**	0.2332	0.1577	0.1555	0.1753	0.1254	0.1261	0.1204	0.2199	0.1302	0.1322	0.1276
30	0.3105	**1.0000**	0.4764	**1.0000**	0.5725	0.6139	0.5058	0.9583	0.9583	0.6393	0.6727	0.9583	0.9583	0.6393
31	0.3898	0.3588	0.3508	0.3597	0.4124	0.4067	0.4123	0.5558	0.5549	**0.5604**	0.3842	0.5385	0.5364	0.5423
32	0.6619	0.6995	0.6640	0.6984	0.9108	0.9019	0.6496	0.9160	0.9235	0.9002	0.6773	0.9227	**0.9280**	0.9047
33	0.0016	0.0012	0.0018	0.0013	0.0031	**0.0032**	0.0014	0.0011	0.0012	0.0012	0.0012	0.0012	0.0012	0.0013
34	0.3417	0.5667	0.3417	0.5667	0.4444	0.4583	**0.7576**	0.4798	0.4798	0.7381	0.7436	0.6556	0.6556	0.7436
35	0.0047	0.0036	0.0037	0.0036	0.0075	0.0072	0.0047	0.0078	0.0078	0.0078	0.0036	0.0078	**0.0079**	**0.0079**
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0155	0.0270	0.0151	0.0262	0.0761	0.0681	0.0028	**0.1458**	0.1359	0.0191	0.0025	0.1419	0.1321	0.0188
38	0.0667	0.0417	0.0435	0.0435	0.0714	**0.0769**	0.0015	0.0455	0.0455	0.0000	0.0015	0.0500	0.0526	0.0000
39	0.0692	0.0699	0.0698	0.0704	0.0417	0.0413	0.0722	0.0403	0.0412	0.0416	**0.0734**	0.0422	0.0405	0.0408
40	0.0000	0.2132	0.1483	0.2139	0.1133	0.1103	0.0000	**0.2398**	0.2368	0.2368	0.2139	0.2396	0.2371	0.2371
41	0.0003	0.0003	0.0003	0.0003	0.0003	0.0000	0.0008	**0.0050**	0.0007	0.0008	0.0007	0.0033	0.0005	0.0007
42	0.0520	0.0627	**0.1385**	0.0621	0.1190	0.1010	0.0972	0.0122	0.0123	0.0498	0.1250	0.0159	0.0155	0.0551
43	0.0024	0.0156	0.0091	0.0153	0.0032	0.0032	0.0040	0.0104	0.0102	**0.0218**	0.0145	0.0105	0.0102	0.0203
44	0.2219	0.2312	0.2062	0.2329	**0.2334**	0.2230	0.2251	0.1815	0.1801	0.1691	0.2314	0.1838	0.1838	0.1735
45	0.0097	0.0097	0.0098	0.0098	0.0859	0.0821	0.0157	0.0969	**0.0982**	**0.0982**	0.0158	0.0896	0.0908	0.0908
46	0.8095	0.7121	0.7121	0.7121	0.2879	0.3178	**1.0000**	0.7381	0.7381	0.7778	0.7500	0.7292	0.7292	0.7917
47	0.0507	0.0354	0.0937	0.0333	0.0745	0.0680	0.0507	0.0866	0.0880	0.0880	0.0333	0.0848	**0.0941**	**0.0941**
48	0.7332	0.7265	0.7330	0.7268	0.8971	0.9038	0.5626	0.9168	0.9178	0.5977	0.5603	0.9187	**0.9200**	0.5977
49	**0.6111**	**0.6111**	**0.6111**	**0.6111**	0.2679	0.2679	0.5000	0.2917	0.2917	0.5000	0.5000	0.2679	0.2679	0.5000
50	0.1862	0.2436	0.2810	0.2457	0.2445	0.2468	0.3235	0.1825	0.1865	**0.4927**	0.4038	0.1832	0.1822	0.4655
51	0.3921	0.3595	0.3789	0.3539	0.4003	0.4068	0.4502	**0.5849**	0.5198	0.5631	0.4234	0.5747	0.5086	0.5611
52	0.0298	0.0335	0.0272	0.0315	0.0280	0.0250	0.0344	**0.0919**	0.0672	0.0679	0.0389	0.0768	0.0596	0.0603
53	0.0924	0.1394	0.1520	0.1394	0.1129	0.1153	0.1198	0.1150	0.1150	0.1849	0.1869	0.1308	0.1305	**0.1925**
54	0.1006	0.1078	0.0971	0.1094	0.1053	0.1167	0.1303	0.1259	0.1290	0.1498	0.1283	0.1341	0.1363	**0.1601**
55	0.0110	0.0979	0.1287	0.0982	0.1573	0.1598	0.0134	**0.2072**	0.2066	0.2042	0.1168	0.2027	0.2016	0.1993
56	0.0003	0.2417	0.0236	0.2418	0.0006	0.0010	0.0003	0.3730	**0.3812**	**0.3812**	0.2418	0.3284	0.3208	0.3208
57	0.1050	0.1516	0.0764	0.1566	0.1624	0.1546	0.1050	0.2309	**0.2326**	**0.2326**	0.1566	0.2266	0.2267	0.2267
58	0.0244	0.0188	0.0203	0.0190	0.0334	0.0346	0.0397	0.0328	0.0338	**0.0555**	0.0331	0.0311	0.0307	0.0522
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7503	0.7716	0.7629	0.7732	0.7539	0.7585	0.7511	0.7769	0.7766	0.7766	0.7732	0.7834	**0.7876**	**0.7876**
61	0.0997	0.1477	0.1499	0.1597	0.0279	0.0253	0.1770	0.0524	0.0661	0.4243	0.3653	0.0539	0.0648	**0.4324**
62	0.2763	0.3349	0.2678	0.3347	0.0768	0.1464	0.2788	0.2624	0.2926	0.2926	**0.3372**	0.2727	0.3025	0.3025
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0037	0.0132	0.0095	0.0126	0.0042	0.0029	0.0064	0.0101	0.0086	0.0105	**0.0167**	0.0093	0.0085	0.0105
65	0.3313	**0.4320**	0.4294	0.4255	0.3248	0.3258	0.3396	0.3822	0.3826	0.3827	0.4272	0.3835	0.3859	0.3860
66	0.3433	0.3423	0.3430	0.3430	0.3433	0.3450	0.3433	0.3529	0.3524	0.3524	0.3430	**0.3535**	0.1869	0.1869
67	0.2888	0.3283	0.3211	0.3238	0.3653	0.3190	0.2882	**0.4220**	0.4138	0.4138	0.3235	0.4128	0.4138	0.4138
68	0.3096	0.7579	0.7376	0.7576	0.4298	0.4265	0.3096	0.7691	**0.7704**	**0.7704**	0.7577	0.7695	0.7687	0.7687
69	0.0001	0.0579	0.0003	0.0568	0.0001	0.0001	0.0001	0.2780	**0.2802**	**0.2802**	0.0568	0.2766	0.2762	0.2762
70	0.0288	0.0346	0.0247	0.0351	0.0130	0.0122	**0.0896**	0.0155	0.0154	0.0742	0.0756	0.0133	0.0135	0.0719
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.3145	0.5604	**0.5671**	0.5608	0.4600	0.4604	0.3145	0.4931	0.4946	0.4946	0.5608	0.4924	0.4926	0.4926
73	**0.0269**	0.0248	0.0249	0.0248	0.0227	0.0227	0.0105	0.0179	0.0181	0.0053	0.0102	0.0182	0.0184	0.0053
74	0.0208	0.4286	0.3448	0.4167	0.0123	0.0119	0.0679	0.3902	0.3902	0.4009	0.5055	0.5556	0.5556	**0.5634**
75	0.2572	0.2931	0.2804	0.2943	0.5553	0.5062	0.4535	0.7042	**0.7043**	0.6696	0.4229	0.6901	0.6904	0.6622
76	0.0000	**0.4316**	0.2289	**0.4316**	0.0000	0.0000	0.0000	0.1301	0.1301	0.1301	**0.4316**	0.1515	0.1515	0.1515
77	0.3832	0.3934	0.3953	0.3946	0.3874	0.3855	0.3832	**0.4185**	0.4151	0.4151	0.3946	0.4177	0.4171	0.4171
78	0.0278	0.1708	0.1138	0.1708	0.0463	0.0428	0.0622	**0.2456**	0.2423	0.2095	0.1447	0.2359	0.2361	0.2073
79	0.8258	0.8139	0.8193	0.8139	**0.9583**	0.9107	0.8258	0.9306	0.9306	0.9306	0.8139	0.9107	0.9107	0.9107
80	0.0386	0.1022	0.1273	0.1272	0.0945	0.0913	0.0770	0.2500	0.2500	**0.2917**	0.1713	0.2361	0.2500	**0.2917**
81	0.0311	0.0581	0.0555	0.0564	0.0681	0.0680	0.0311	0.0795	0.0793	0.0793	0.0564	0.0800	**0.0801**	**0.0801**
82	0.0017	0.0004	0.0017	0.0004	0.0032	0.0028	0.0014	0.0035	**0.0036**	**0.0036**	0.0003	0.0025	0.0025	0.0021
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.0996	0.2975	0.0997	**0.2981**	0.0350	0.0356	0.1004	0.0615	0.0603	0.0630	**0.2981**	0.0623	0.0600	0.0634
85	0.3851	0.3332	0.3616	0.3330	**0.5144**	0.4956	0.3402	0.4672	0.4650	0.3877	0.2933	0.4463	0.4446	0.3742
86	0.0035	0.0040	0.0028	0.0042	0.0004	0.0004	0.0061	0.0023	0.0023	0.0033	**0.0067**	0.0021	0.0021	0.0033
87	0.2193	0.2558	0.2575	0.2555	0.1848	0.1851	0.2193	0.2815	0.2830	0.2830	0.2555	0.2870	**0.2874**	**0.2874**
88	0.0823	0.3281	0.2713	0.3255	0.1687	0.1797	0.1388	0.3377	0.3707	0.2677	0.2707	0.3375	**0.3742**	0.2663
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90	0.0806	0.1478	0.0809	**0.1511**	0.0556	0.0561	0.0806	0.1165	0.1173	0.1173	**0.1511**	0.1147	0.1157	0.1157
91	0.1250	0.5057	0.2562	**0.5058**	0.0238	0.0437	0.2500	0.1751	0.1754	0.5011	0.5000	0.1758	0.2593	0.5012
92	0.0892	0.0725	0.0882	0.0728	0.1420	0.1407	0.1867	0.1022	0.1013	0.2357	0.1904	0.1009	0.1010	**0.2387**
93	0.3576	0.8107	0.8098	0.8100	0.8521	0.8570	0.3952	0.8926	0.8910	0.8709	0.8239	**0.8985**	0.8948	0.8760
94	0.1163	0.3687	0.2798	0.3686	0.0500	0.0533	0.1237	0.0627	0.0611	0.0613	**0.3738**	0.0573	0.0567	0.0570
95	0.4382	0.4484	0.4554	0.4472	0.6542	0.6484	0.3225	**0.6853**	0.6614	0.3824	0.3173	0.6798	0.6530	0.3804
96	0.2195	0.2781	0.2302	0.2786	0.2514	0.2433	0.2344	0.2877	0.2878	0.2615	**0.3001**	0.2777	0.2778	0.2555
97	0.0369	0.0363	0.0369	0.0362	0.0394	0.0394	0.0333	**0.0404**	**0.0404**	0.0378	0.0328	0.0384	0.0386	0.0361
98	0.2087	0.2502	0.2094	0.2536	0.3583	0.3440	0.2832	0.8149	0.8149	**0.8451**	0.3961	0.8064	0.8055	0.8418
99	0.1021	0.1384	0.1115	0.1397	0.2086	**0.2101**	0.1022	0.1659	0.1682	0.1682	0.1398	0.1653	0.1682	0.1682
100	0.0125	0.0180	0.0197	0.0178	**0.0376**	0.0350	0.0044	0.0205	0.0204	0.0140	0.0116	0.0185	0.0188	0.0137
all	0.1939	0.2649	0.2370	0.2646	0.2388	0.2384	0.2078	0.2969	0.2949	0.2974	0.2663	**0.3001**	0.2978	0.2976

Table C.5: TFIDF per query results (TD. title and description tags used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6480	0.6398	**0.6590**	0.6446	0.6053	0.5991	0.5971	0.6405	0.6420	0.6580	0.6217	0.6379	0.6400	0.6560
02	0.0257	0.1666	0.0729	0.1339	0.0099	0.0116	0.0381	0.0562	0.0124	0.0246	**0.1706**	0.0849	0.0149	0.0288
03	0.0934	0.0777	0.0604	0.0780	0.0490	0.0463	**0.0936**	0.0278	0.0230	0.0231	0.0780	0.0206	0.0173	0.0174
04	0.1546	0.1873	0.1843	0.1877	0.1811	0.1822	0.1205	0.1843	**0.1889**	0.1426	0.1516	0.1815	0.1835	0.1367
05	0.5817	0.6002	0.6080	0.5968	0.6427	0.6431	0.5895	0.6318	0.6512	0.6602	0.6024	0.6349	0.6541	**0.6659**
06	0.3807	**0.3891**	0.3872	0.3760	0.3579	0.3814	0.3463	0.1992	0.2392	0.2098	0.3620	0.2142	0.2533	0.2226
07	0.3452	0.1012	0.1143	0.1008	0.4423	**0.4545**	0.2865	0.0285	0.0298	0.0290	0.0839	0.0246	0.0247	0.0263
08	0.0422	0.0450	0.0403	0.0396	0.0354	0.0365	**0.0577**	0.0358	0.0351	0.0524	0.0540	0.0362	0.0346	0.0522
09	0.3675	0.4042	0.2957	0.3133	0.4193	0.3912	0.4023	**0.4312**	0.3657	0.4074	0.3458	0.4262	0.3707	0.4049
10	0.4743	0.6656	0.5386	0.6499	0.6850	0.5607	0.4791	0.8541	0.8487	0.8552	0.6546	0.8589	0.8534	**0.8645**
11	0.0494	0.0495	0.0562	0.0561	0.1251	**0.1321**	0.0506	0.0708	0.0657	0.0659	0.0585	0.0690	0.0636	0.0638
12	0.0147	0.1344	0.1249	0.1332	0.0033	0.0031	0.0142	0.2043	0.2060	0.2348	0.1315	0.2096	0.2133	**0.2438**
13	0.2729	0.3988	0.4258	0.4199	0.4589	**0.4783**	0.3882	0.3207	0.3180	0.3536	0.4433	0.3356	0.3377	0.3758
14	0.2492	0.2213	0.2703	0.2281	0.5792	0.5740	0.2236	0.6453	**0.6491**	0.4475	0.2744	0.6444	0.6484	0.4464
15	0.6757	0.7111	0.6859	0.7121	0.7698	0.7683	0.6868	0.7767	0.7862	**0.7885**	0.7206	0.7766	0.7812	0.7832
16	0.8292	0.8667	0.8416	0.8592	0.8736	0.8764	0.7995	**0.9312**	0.9034	0.8043	0.7941	0.9240	0.9111	0.8102
17	0.3812	**0.4443**	0.4411	**0.4443**	0.4098	0.4102	0.3643	0.4107	0.4080	0.3911	0.4225	0.4157	0.4162	0.3993
18	0.1709	0.3074	0.2910	0.3218	0.1205	0.1186	0.1977	0.2754	0.2789	0.3133	0.3561	0.3138	0.3192	**0.3602**
19	0.1328	0.1457	0.1457	0.1487	0.1581	0.1509	0.1737	0.1760	0.1980	0.2331	0.1979	0.1707	0.1991	**0.2353**
20	0.2526	0.1320	**0.3592**	0.1570	0.1224	0.1217	0.0059	0.0435	0.0226	0.0019	0.0028	0.0460	0.0312	0.0019
21	0.3518	0.5069	0.5433	0.5466	0.4197	0.4183	0.3763	0.4563	0.4283	0.6226	0.5423	0.4621	0.4408	**0.6242**
22	0.3269	0.3526	0.3389	0.3606	0.4168	0.4082	0.3018	0.4278	0.4308	0.4484	0.3998	0.4256	0.4323	**0.4548**
23	0.0104	0.0160	0.0251	0.0239	0.0071	0.0070	0.0288	0.0322	0.0342	0.1177	0.0488	0.0316	0.0314	**0.1210**
24	0.5399	0.5800	0.5898	0.5841	0.5963	0.5963	0.5327	**0.6398**	0.6387	0.6285	0.5788	0.6371	0.6360	0.6251
25	0.4444	0.4111	0.4000	0.3909	0.5889	0.5889	0.4778	**0.8095**	0.7667	0.7778	0.4111	**0.8095**	0.7778	0.7917
26	0.0118	0.2797	0.2523	0.2824	0.0186	0.0177	0.0157	0.3298	0.3304	**0.3587**	0.2981	0.3261	0.3260	0.3541
27	0.0448	0.0211	0.0215	0.0211	0.0387	**0.0739**	0.0470	0.0644	0.0645	0.0648	0.0236	0.0661	0.0662	0.0664
28	0.0318	0.0742	0.0725	0.0730	0.0922	0.0922	0.0005	0.2534	**0.2582**	0.0069	0.0020	0.2458	0.2461	0.0069
29	0.1873	**0.2402**	0.2377	0.2363	0.1097	0.1053	0.1780	0.1112	0.1112	0.1056	0.2245	0.1350	0.1337	0.1293
30	0.3445	**1.0000**	0.5501	**1.0000**	0.2318	0.2233	0.5225	0.9444	0.9444	0.6393	0.6729	0.9306	0.9583	0.6394
31	0.3713	0.3558	0.3483	0.3599	0.3916	0.3833	0.3965	**0.5191**	0.4890	0.4960	0.3870	0.4933	0.4519	0.4606
32	0.8589	0.9218	0.8747	0.9226	0.9238	0.9297	0.8273	0.9620	0.9620	0.9325	0.8898	0.9619	**0.9650**	0.9355
33	0.0047	0.0031	0.0042	0.0031	0.0042	0.0039	0.0027	0.0033	0.0033	**0.0061**	0.0036	0.0030	0.0027	0.0042
34	0.5889	0.4583	0.5000	0.4583	0.5000	0.5000	**0.8095**	0.4444	0.4444	0.7500	0.7576	0.4444	0.4444	0.7500
35	**0.1234**	0.1067	0.1078	0.1067	0.0425	0.0395	0.0554	0.0750	0.0763	0.0508	0.0540	0.0793	0.0792	0.0548
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0631	0.0476	0.0995	0.0509	**0.1002**	0.0981	0.0044	0.0739	0.0790	0.0198	0.0036	0.0707	0.0757	0.0199
38	0.0098	0.0417	0.0417	0.0417	0.0333	0.0400	0.0010	0.0588	0.0588	0.0011	0.0012	**0.0769**	**0.0769**	0.0012
39	0.0619	0.0731	0.0690	0.0712	0.0480	0.0466	0.0670	0.0484	0.0485	0.0487	**0.0774**	0.0476	0.0478	0.0480
40	0.1487	**0.2701**	0.2655	0.2681	0.2472	0.2507	0.1487	0.2475	0.2467	0.2467	0.2681	0.2488	0.2487	0.2487
41	0.0003	0.0003	0.0003	0.0003	0.0003	0.0000	0.0009	**0.0135**	0.0052	0.0027	0.0007	0.0121	0.0040	0.0027
42	0.0479	0.0492	0.0878	0.0490	0.1769	**0.1909**	0.0943	0.0065	0.0064	0.0719	0.1037	0.0071	0.0070	0.0777
43	0.0019	0.0156	0.0093	0.0151	0.0032	0.0032	0.0029	0.0146	0.0137	0.0312	0.0311	0.0153	0.0145	**0.0330**
44	0.2018	0.2107	0.1903	0.2123	0.2376	**0.2383**	0.2155	0.1814	0.1837	0.1721	0.2142	0.1864	0.1847	0.1736
45	0.0497	0.0473	0.0475	0.0472	0.1563	0.1473	0.0601	0.1730	**0.1788**	**0.1788**	0.0573	0.1746	0.1746	0.1746
46	0.8095	0.7167	0.7193	0.7167	0.5222	0.5193	**1.0000**	0.7292	0.7292	0.7917	0.7667	0.7255	0.7255	0.8333
47	0.0324	0.0530	0.0704	0.0528	0.0710	0.0565	0.0324	0.0717	**0.0733**	**0.0733**	0.0528	0.0685	0.0682	0.0682
48	0.7522	0.7557	0.7432	0.7557	0.8919	0.8994	0.5470	0.9072	0.9084	0.5998	0.5491	0.9082	**0.9087**	0.5996
49	**0.6000**	**0.6000**	**0.6000**	**0.6000**	0.3929	0.4167	0.5000	0.4167	0.4167	0.5000	0.5000	0.4167	0.4167	0.5000
50	0.1836	0.2534	0.2835	0.2522	0.2426	0.2432	0.3502	0.1739	0.1756	**0.4702**	0.4400	0.1792	0.1816	0.4590
51	0.5642	0.5220	0.5516	0.5186	0.6556	0.6633	0.5778	0.6916	0.6844	0.6667	0.5308	**0.6954**	0.6886	0.6718
52	0.0077	0.0077	0.0071	0.0074	0.0055	0.0054	0.0097	0.0121	0.0144	**0.0239**	0.0144	0.0117	0.0126	0.0213
53	0.1273	0.1415	0.1415	0.1425	0.1379	0.1392	0.1375	0.1525	0.1539	0.1980	0.1581	0.1562	0.1564	**0.1988**
54	0.1009	0.1090	0.0815	0.1092	0.0981	0.1000	0.1197	0.1176	0.1189	0.1431	0.1237	0.1211	0.1223	**0.1508**
55	0.0464	0.0885	0.1437	0.0892	0.1999	0.1999	0.0533	**0.2053**	**0.2053**	0.2031	0.1111	0.2000	0.1997	0.1989
56	0.0174	0.1954	0.1178	0.1958	0.0117	0.0134	0.0174	**0.4836**	0.4777	0.4777	0.1958	0.3974	0.3832	0.3832
57	0.2108	0.1894	0.1786	0.1837	0.1937	0.1851	**0.2226**	0.2145	0.2148	0.2169	0.1963	0.2103	0.2120	0.2151
58	0.0319	0.0227	0.0259	0.0223	0.0520	0.0481	0.0486	0.0556	0.0549	**0.0664**	0.0364	0.0470	0.0464	0.0603
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7641	0.7766	0.7781	0.7778	0.7647	0.7723	0.7613	0.7788	0.7816	0.7757	0.7742	0.7840	**0.7853**	0.7781
61	0.0919	0.1477	0.1590	0.1541	0.0218	0.0201	0.1592	0.0598	0.0501	0.4512	0.3334	0.0606	0.0518	**0.4731**
62	0.3350	0.3952	0.3356	0.3950	0.3175	0.3333	0.3377	0.3353	0.2296	0.2296	**0.3976**	0.1760	0.1803	0.1803
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0034	0.0109	0.0080	0.0104	0.0080	0.0042	0.0049	0.0105	0.0116	**0.0149**	0.0145	0.0112	0.0114	0.0147
65	0.3082	0.4419	0.4352	0.4297	0.3230	0.3239	0.3219	0.4495	0.4427	0.4439	0.4373	0.4471	0.4499	**0.4500**
66	0.1673	0.3361	**0.3365**	0.3362	0.0208	0.0145	0.1673	0.1689	0.1689	0.1689	0.3362	0.1683	0.1684	0.1684
67	0.3634	0.4012	0.3883	0.4016	0.3551	0.3578	0.1931	**0.4183**	0.4141	0.1715	0.1821	0.4088	0.4065	0.1745
68	0.3858	0.7506	0.7208	0.7465	0.5943	0.5846	0.3858	0.7476	0.7517	0.7517	0.7465	0.7514	**0.7541**	**0.7541**
69	0.0724	0.1345	0.1060	0.1339	0.3671	**0.3729**	0.0724	0.2871	0.2958	0.2958	0.1339	0.2737	0.2818	0.2818
70	0.0238	0.0254	0.0158	0.0251	0.0068	0.0067	**0.0794**	0.0114	0.0113	0.0660	0.0639	0.0097	0.0098	0.0643
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.5094	0.7034	0.6996	0.7043	0.4574	0.4878	0.5094	**0.7182**	0.7072	0.7072	0.7043	0.7069	0.7001	0.7001
73	0.0181	0.0169	0.0176	0.0170	0.0166	0.0175	0.0080	0.0261	**0.0263**	0.0088	0.0082	0.0252	0.0254	0.0088
74	0.3349	0.3505	0.3348	0.3490	0.3343	0.3342	0.3769	0.3415	0.3405	0.4551	**0.5000**	0.3420	0.3411	0.4524
75	0.4045	0.4214	0.3992	0.4193	**0.6421**	0.6074	0.5317	0.6231	0.6233	0.6039	0.4943	0.5970	0.5971	0.5834
76	0.0000	0.4462	0.2505	0.4462	0.0000	0.0000	0.0000	0.4780	0.4887	0.3000	0.2800	0.4886	**0.5029**	0.3000
77	0.2883	0.2921	0.2783	0.2758	0.3653	0.3751	0.2883	**0.3976**	0.3868	0.3868	0.2758	0.3911	0.3771	0.3771
78	0.0165	0.1016	0.0723	0.1010	0.0297	0.0227	0.0502	0.2285	**0.2287**	0.2054	0.0938	0.2281	0.2264	0.2044
79	**0.9762**	0.9583	0.9583	0.9583	**0.9762**	**0.9762**	**0.9762**	**0.9762**	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**
80	0.0120	0.0734	0.0854	0.0734	0.0617	0.0513	0.0149	**0.2250**	0.2000	0.2111	0.1041	0.1667	0.1714	0.1769
81	0.0729	0.0733	0.0699	0.0684	0.0829	0.0820	0.0729	**0.0932**	0.0931	0.0931	0.0684	0.0919	0.0920	0.0920
82	0.0014	0.0000	0.0013	0.0000	**0.0025**	0.0022	0.0011	0.0023	0.0023	0.0019	0.0000	0.0017	0.0017	0.0015
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.3794	0.3001	0.2015	0.2971	0.2803	0.2847	**0.3812**	0.0840	0.0839	0.0866	0.2981	0.0850	0.0845	0.0882
85	0.3834	0.3343	0.3608	0.3329	**0.4776**	0.4757	0.3418	0.4579	0.4590	0.3883	0.2960	0.4409	0.4377	0.3725
86	0.0275	0.0199	0.0157	0.0209	0.0542	0.0545	0.0462	0.1550	0.1588	**0.1639**	0.0328	0.1508	0.1504	0.1548
87	0.1769	0.2050	0.2033	0.2034	0.2015	0.1993	0.1769	**0.2263**	0.2243	0.2243	0.2034	0.2228	0.2213	0.2213
88	0.0816	0.2724	0.2491	0.2657	0.1614	0.1762	0.1497	0.3634	0.3632	0.2653	0.2529	**0.3717**	0.3710	0.2664
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90	0.1164	0.1588	0.0853	**0.1590**	0.0926	0.0948	0.1164	0.0994	0.1004	0.1004	**0.1590**	0.1061	0.1070	0.1070
91	0.1250	0.5047	0.1302	0.5048	0.0235	0.0228	0.2500	0.2573	0.2574	0.5012	0.5000	0.5080	**0.5081**	0.5013
92	0.0935	0.0782	0.0990	0.0808	0.1365	0.1425	0.1952	0.1196	0.1211	**0.2600**	0.2003	0.1154	0.1175	0.2558
93	0.3396	0.7896	0.7903	0.7896	0.8348	0.8394	0.3824	0.8897	0.8889	0.8736	0.8121	**0.8952**	0.8951	0.8797
94	0.1151	0.2812	0.1851	0.2811	0.0495	0.0507	0.1202	0.0765	0.0763	0.0773	**0.2857**	0.0763	0.0737	0.0740
95	0.4307	0.4398	0.4530	0.4407	0.5599	0.5431	0.3275	**0.6573**	0.6567	0.3805	0.3165	0.6475	0.6432	0.3787
96	0.1858	0.2374	0.2014	0.2367	0.2403	0.2245	0.1995	0.2791	**0.2800**	0.2586	0.2513	0.2679	0.2682	0.2494
97	0.0344	0.0364	0.0371	0.0365	0.0405	0.0396	0.0311	**0.0425**	**0.0425**	0.0388	0.0332	0.0410	0.0411	0.0376
98	0.1770	0.3133	0.2025	0.2680	0.3158	0.3006	0.2349	0.7196	0.7196	**0.8500**	0.4081	0.7132	0.7132	0.8389
99	0.1001	0.1281	0.1008	0.1280	**0.2362**	0.2228	0.1002	0.1700	0.1716	0.1716	0.1280	0.1751	0.1783	0.1783
100	0.0139	0.0216	0.0208	0.0215	0.0530	**0.0549**	0.0052	0.0330	0.0330	0.0147	0.0114	0.0310	0.0309	0.0142
all	0.2238	0.2740	0.2573	0.2726	0.2541	0.2531	0.2307	**0.3007**	0.2977	0.2988	0.2735	0.3001	0.2987	0.2978

Table C.6: BM25 per query results (TD. title and description tags used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6467	0.6480	0.6668	0.6535	0.6041	0.5989	0.5967	0.6650	0.6649	0.6655	0.6374	0.6649	0.6652	**0.6690**
02	0.0250	0.1549	0.0728	0.1276	0.0097	0.0112	0.0388	0.0492	0.0171	0.0319	**0.1703**	0.0801	0.0217	0.0404
03	0.1097	0.0781	0.0626	0.0783	0.1886	**0.1899**	0.1099	0.0310	0.0274	0.0275	0.0783	0.0220	0.0215	0.0216
04	0.1567	0.2036	0.1942	**0.2040**	0.1866	0.1867	0.1212	0.1847	0.1899	0.1432	0.1661	0.1821	0.1879	0.1408
05	0.5668	0.5983	0.5983	0.5963	0.6402	**0.6505**	0.5738	0.6322	0.6305	0.6384	0.6001	0.6390	0.6313	0.6417
06	0.3654	0.3902	**0.3927**	0.3899	0.3375	0.3799	0.3403	0.1973	0.2369	0.2084	0.3681	0.2111	0.2519	0.2380
07	0.3272	0.0904	0.1053	0.0904	0.4282	**0.4455**	0.2684	0.0325	0.0337	0.0332	0.0782	0.0283	0.0310	0.0300
08	0.0425	0.0441	0.0397	0.0392	0.0353	0.0362	**0.0568**	0.0352	0.0343	0.0498	0.0529	0.0357	0.0343	0.0518
09	0.3665	0.3827	0.2608	0.2721	0.4239	0.3922	0.4015	**0.4316**	0.3903	0.4082	0.2932	0.4232	0.3787	0.4018
10	0.5162	0.6706	0.5827	0.6740	0.7520	0.7211	0.5240	0.8481	0.8512	0.8607	0.6789	0.8592	0.8642	**0.8784**
11	0.0511	0.0410	0.0418	0.0403	0.1255	**0.1322**	0.0523	0.0670	0.0247	0.0247	0.0424	0.0625	0.0312	0.0312
12	0.0140	0.1384	0.1260	0.1370	0.0031	0.0029	0.0135	0.1999	0.2017	0.2292	0.1368	0.2068	0.2098	**0.2382**
13	0.2622	0.4171	0.4289	0.4422	0.4580	**0.5271**	0.3482	0.3146	0.4676	0.4731	0.4599	0.3412	0.5001	0.5073
14	0.2342	0.2087	0.2601	0.2113	0.5896	0.5844	0.2303	0.6396	0.6369	0.4402	0.2729	**0.6444**	0.6409	0.4432
15	0.6748	0.7113	0.6836	0.7142	0.7698	0.7714	0.6860	0.7770	0.7890	**0.7912**	0.7224	0.7759	0.7825	0.7847
16	0.8206	0.8512	0.8459	0.8452	0.8764	0.8734	0.7994	0.9050	0.9034	0.8043	0.7940	0.9092	**0.9129**	0.8100
17	0.3692	**0.4210**	0.4193	0.4196	0.3944	0.3918	0.3527	0.3993	0.3983	0.3820	0.3984	0.4033	0.4035	0.3872
18	0.1635	0.2958	0.2812	0.3109	0.1186	0.1143	0.1944	0.2603	0.2660	0.3032	0.3396	0.2975	0.3064	**0.3470**
19	0.1350	0.1487	0.1551	0.1562	0.1608	0.1533	0.1761	0.1943	0.2032	**0.2369**	0.2055	0.1968	0.2003	0.2358
20	0.2563	0.1386	**0.3626**	0.1670	0.0819	0.1564	0.0081	0.0451	0.0214	0.0018	0.0035	0.0457	0.0314	0.0020
21	0.3558	0.5074	0.5482	0.5464	0.4001	0.3991	0.3991	0.4477	0.4223	0.6145	0.5444	0.4533	0.4270	**0.6175**
22	0.3295	0.3337	0.3218	0.3411	0.4163	0.4073	0.3118	0.3784	0.3795	**0.4508**	0.4112	0.3765	0.3773	0.4478
23	0.0117	0.0200	0.0245	0.0270	0.0068	0.0081	0.0338	0.0337	0.0350	0.1435	0.0598	0.0316	0.0332	**0.1441**
24	0.5505	0.5896	0.6021	0.5941	0.5945	0.5989	0.5452	0.6339	**0.6382**	0.6289	0.5869	0.6331	0.6330	0.6238
25	0.4444	0.4583	0.4444	0.4583	**0.8667**	0.8333	0.4778	0.8095	0.8095	0.8333	0.4583	0.8095	0.8095	0.8333
26	0.0116	0.2846	0.2586	0.2851	0.0204	0.0196	0.0153	0.3314	0.3315	**0.3602**	0.3027	0.3277	0.3278	0.3556
27	0.0451	0.0221	0.0220	0.0218	0.0456	0.0458	0.0471	0.0633	0.0630	0.0634	0.0236	0.0652	0.0650	**0.0653**
28	0.0332	0.0714	0.0696	0.0711	0.0951	0.0925	0.0005	0.2781	**0.2812**	0.0111	0.0020	0.2646	0.2501	0.0098
29	0.1878	0.2426	**0.2437**	0.2430	0.1118	0.1072	0.1780	0.1123	0.1121	0.1058	0.2304	0.1368	0.1356	0.1303
30	0.3239	**1.0000**	0.5464	**1.0000**	0.6566	0.6677	0.5058	0.9167	0.9167	0.6393	0.6728	0.9306	0.9444	0.6394
31	0.3619	0.3513	0.3449	0.3553	0.3935	0.3938	0.3871	**0.5313**	0.3735	0.3830	0.3811	0.5031	0.3639	0.3729
32	0.8649	0.9234	0.8742	0.9223	0.9341	0.9337	0.8306	0.9620	0.9630	0.9335	0.8894	**0.9647**	0.9639	0.9334
33	**0.0044**	0.0031	0.0037	0.0031	0.0040	0.0036	0.0028	0.0027	0.0029	0.0027	0.0037	0.0027	0.0030	0.0025
34	0.5556	0.4583	0.4762	0.4583	0.4444	0.5000	**0.8095**	0.4444	0.4333	0.7436	0.7576	0.4444	0.4444	0.7500
35	**0.1238**	0.1072	0.1088	0.1077	0.0389	0.0364	0.0552	0.0592	0.0611	0.0397	0.0535	0.0636	0.0645	0.0408
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0679	0.0505	**0.1033**	0.0522	0.1016	0.0997	0.0064	0.0757	0.0812	0.0282	0.0039	0.0726	0.0769	0.0199
38	0.0115	0.0417	0.0417	0.0417	0.0323	0.0345	0.0012	0.0588	0.0588	0.0011	0.0012	**0.0714**	**0.0714**	0.0012
39	0.0620	0.0732	0.0691	0.0764	0.0497	0.0492	0.0675	0.0479	0.0481	0.0483	**0.0830**	0.0469	0.0471	0.0473
40	0.1592	**0.2992**	0.2704	0.2973	0.2489	0.2506	0.1592	0.2478	0.2454	0.2454	0.2973	0.2470	0.2478	0.2478
41	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0009	**0.0163**	0.0069	0.0029	0.0007	0.0152	0.0052	0.0028
42	0.0478	0.0487	0.0928	0.0483	0.1769	**0.2159**	0.0911	0.0062	0.0061	0.0709	0.1027	0.0062	0.0062	0.0857
43	0.0024	0.0136	0.0084	0.0131	0.0033	0.0032	0.0033	0.0156	0.0152	0.0388	0.0305	0.0153	0.0147	**0.0408**
44	0.2042	0.2109	0.1906	0.2072	**0.2377**	0.2344	0.2158	0.1764	0.1759	0.1640	0.2086	0.1783	0.1775	0.1663
45	0.0570	0.0534	0.0539	0.0539	0.1656	0.1538	0.0662	**0.2039**	0.1875	0.1875	0.0635	0.2016	0.1841	0.1841
46	0.8095	0.7167	0.7222	0.7222	0.3069	0.5222	**1.0000**	0.7333	0.5769	0.6111	0.7667	0.7333	0.7500	0.8333
47	0.0377	0.0615	**0.0830**	0.0614	0.0747	0.0613	0.0377	0.0712	0.0732	0.0732	0.0614	0.0683	0.0682	0.0682
48	0.7570	0.7607	0.7497	0.7609	0.8933	0.9001	0.5479	0.9095	0.9107	0.5999	0.5488	0.9103	**0.9112**	0.5996
49	**0.6000**	**0.6000**	**0.6000**	**0.6000**	0.3929	0.3929	0.5000	0.3929	0.3929	0.5000	0.5000	0.4167	0.4167	0.5000
50	0.1852	0.2466	0.2825	0.2467	0.2465	0.2487	0.3532	0.2027	0.1996	0.5313	0.4270	0.1992	0.1999	**0.5323**
51	0.5585	0.5138	0.5481	0.5150	0.6504	0.6636	0.5737	0.6960	0.6903	0.6695	0.5271	**0.6984**	0.6935	0.6744
52	0.0075	0.0078	0.0070	0.0070	0.0047	0.0047	0.0094	0.0119	0.0131	**0.0238**	0.0142	0.0117	0.0118	0.0192
53	0.1319	0.1326	0.1314	0.1313	0.1352	0.1361	0.1471	0.1396	0.1405	0.1909	0.1547	0.1424	0.1432	**0.1939**
54	0.1013	0.1112	0.0888	0.1114	0.0980	0.1000	0.1207	0.1133	0.1149	0.1381	0.1263	0.1224	0.1237	**0.1516**
55	0.0559	0.1016	0.1507	0.1022	0.1976	0.2020	0.0652	0.1920	**0.2103**	0.2070	0.1214	0.1931	0.2016	0.2001
56	0.0175	0.1889	0.1056	0.1891	0.0116	0.0134	0.0175	**0.4915**	0.4873	0.4873	0.1891	0.4108	0.4108	0.4108
57	0.1764	0.1921	0.1782	0.1846	0.2124	0.2006	0.1894	0.2085	0.2147	0.2167	0.1974	0.2145	0.2152	**0.2181**
58	0.0324	0.0219	0.0271	0.0212	0.0529	0.0490	0.0493	0.0552	0.0548	**0.0668**	0.0356	0.0482	0.0480	0.0619
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7641	0.7775	0.7785	0.7781	0.7633	0.7706	0.7613	0.7785	0.7783	0.7727	0.7744	**0.7875**	**0.7875**	0.7787
61	0.0924	0.1507	0.1617	0.1580	0.0213	0.0197	0.1620	0.0481	0.0420	0.4842	0.3671	0.0496	0.0482	**0.4946**
62	0.3357	0.3954	0.3302	0.3951	0.3183	0.3287	0.3384	0.3190	0.2144	0.2144	**0.3977**	0.1951	0.1688	0.1688
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0031	0.0122	0.0085	0.0119	0.0082	0.0043	0.0050	0.0088	0.0081	0.0103	**0.0154**	0.0083	0.0102	0.0122
65	0.3024	**0.4431**	0.4324	0.4228	0.3232	0.3226	0.3148	0.4418	0.4083	0.4084	0.4294	0.4374	0.4254	0.4255
66	0.1667	0.3362	**0.3365**	0.3363	0.0093	0.0067	0.1667	0.1688	0.1134	0.1134	0.3363	0.1684	0.1129	0.1129
67	0.3586	0.3982	0.3864	0.3990	0.3532	0.3511	0.1871	**0.4195**	0.4194	0.1716	0.1817	0.3975	0.3995	0.1732
68	0.3972	0.7490	0.7204	0.7452	0.6066	0.5976	0.3972	0.7467	**0.7536**	0.7535	0.7454	0.7523	0.7529	0.7529
69	0.0722	0.1349	0.1078	0.1338	0.2077	0.2306	0.0722	0.2900	**0.2921**	**0.2921**	0.1338	0.2762	0.2756	0.2756
70	0.0236	0.0266	0.0160	0.0284	0.0068	0.0058	**0.0801**	0.0127	0.0144	0.0693	0.0655	0.0105	0.0122	0.0660
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.5183	0.7086	0.6999	0.7065	0.4574	0.4880	0.5183	**0.7202**	0.7084	0.7084	0.7065	0.7088	0.7094	0.7094
73	0.0175	0.0161	0.0166	0.0161	0.0164	0.0175	0.0080	0.0265	0.0269	0.0062	0.0081	0.0269	**0.0270**	0.0072
74	0.3351	0.3543	0.3355	0.3533	0.3344	0.3343	0.3787	0.3411	0.3414	0.4537	**0.5033**	0.3423	0.3425	0.4500
75	0.4122	0.4395	0.4386	0.4385	0.6332	0.6025	0.5241	**0.6408**	0.6389	0.6174	0.4982	0.6011	0.6014	0.5837
76	0.0000	0.4500	0.2622	0.4500	0.0000	0.0000	0.0000	0.4783	0.4890	0.3000	0.2800	0.4889	**0.5032**	0.3000
77	0.3001	0.3086	0.2923	0.2905	0.3666	0.3697	0.3001	**0.4020**	0.3887	0.3887	0.2905	0.3985	0.3837	0.3837
78	0.0188	0.1071	0.0793	0.1071	0.0456	0.0479	0.0503	0.2301	**0.2307**	0.2072	0.0939	0.2295	0.2241	0.2030
79	**0.9762**	0.9583	0.9583	0.9583	**0.9762**	**0.9762**	**0.9762**	**0.9762**	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**
80	0.0122	0.0733	0.0734	0.0733	0.0763	0.0530	0.0155	0.2250	0.2250	**0.2429**	0.1040	0.1833	0.1769	0.1833
81	0.0705	0.0686	0.0697	0.0673	0.0854	0.0840	0.0705	**0.0941**	0.0925	0.0925	0.0673	0.0935	0.0902	0.0902
82	0.0014	0.0000	0.0014	0.0000	0.0028	0.0026	0.0011	0.0029	**0.0033**	0.0027	0.0000	0.0021	0.0025	0.0021
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.3872	0.3007	0.2075	0.2983	0.2986	0.3170	**0.3932**	0.0709	0.0715	0.0721	0.2991	0.0792	0.0789	0.0815
85	0.3840	0.3359	0.3554	0.3357	**0.4852**	0.4796	0.3439	0.4691	0.4684	0.3942	0.2981	0.4493	0.4453	0.3787
86	0.0316	0.0239	0.0171	0.0241	0.0732	0.0740	0.0506	0.1746	0.1762	0.1835	0.0339	0.1595	0.1752	**0.1855**
87	0.1753	0.2159	0.2109	0.2148	0.2052	0.2046	0.1753	0.2574	0.2396	0.2396	0.2148	**0.2585**	0.2358	0.2358
88	0.0811	0.2893	0.2592	0.2866	0.1498	0.1631	0.1436	**0.3649**	0.3608	0.2629	0.2593	0.3629	0.3580	0.2608
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90	0.1166	0.1572	0.0791	**0.1579**	0.0934	0.0955	0.1166	0.0976	0.0984	0.0984	**0.1579**	0.1043	0.1051	0.1051
91	0.1250	0.5044	0.1716	**0.5045**	0.0245	0.0258	0.2500	0.1742	0.1742	0.5011	0.5000	0.2576	0.1749	0.5012
92	0.0969	0.0792	0.0995	0.0818	0.1390	0.1461	0.1997	0.1228	0.1238	**0.2623**	0.2041	0.1183	0.1205	0.2602
93	0.3411	0.7823	0.7819	0.7808	0.8388	0.8467	0.3848	0.8867	0.8868	0.8718	0.8097	**0.8936**	0.8932	0.8778
94	0.1151	0.2729	0.1735	0.2729	0.0488	0.0498	0.1204	0.0817	0.0815	0.0829	**0.2848**	0.0941	0.0913	0.0919
95	0.4244	0.4391	0.4463	0.4381	**0.6611**	0.6469	0.3260	0.6580	0.6579	0.3815	0.3179	0.6519	0.6485	0.3798
96	0.1859	0.2365	0.2024	0.2384	0.2399	0.2258	0.1992	0.2821	**0.2826**	0.2599	0.2525	0.2697	0.2694	0.2498
97	0.0342	0.0362	0.0368	0.0362	0.0405	0.0397	0.0309	**0.0426**	0.0425	0.0392	0.0329	0.0412	0.0413	0.0381
98	0.1803	0.3252	0.2109	0.3253	0.6400	0.6428	0.2387	0.7020	0.7257	**0.8600**	0.3997	0.7178	0.7178	**0.8600**
99	0.1043	0.1838	0.1385	0.1847	**0.2381**	0.2309	0.1049	0.2056	0.2072	0.2072	0.1847	0.2257	0.2283	0.2283
100	0.0135	0.0206	0.0209	0.0201	0.0560	**0.0576**	0.0048	0.0328	0.0322	0.0158	0.0114	0.0296	0.0296	0.0147
all	0.2237	0.2756	0.2589	0.2752	0.2635	0.2676	0.2312	0.2997	0.2951	0.2983	0.2756	0.2991	0.2962	**0.3002**

Table C.7: InL2 per query results (TD. title and description tags used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6326	0.6412	**0.6585**	0.6456	0.6018	0.6126	0.5918	0.6336	0.6337	0.6446	0.6225	0.6466	0.6362	0.6502
02	0.0270	0.1773	0.0734	0.1558	0.0099	0.0115	0.0398	0.0577	0.0290	0.0513	**0.1880**	0.0906	0.0392	0.0629
03	0.0946	0.0733	0.0623	0.0735	0.1724	**0.1788**	0.0948	0.0273	0.0272	0.0272	0.0736	0.0229	0.0229	0.0229
04	0.1526	0.1755	0.1556	0.1764	0.1733	0.1770	0.1205	**0.1869**	0.1843	0.1391	0.1439	0.1837	0.1822	0.1381
05	0.5709	0.5935	0.5962	0.5965	0.6340	**0.6422**	0.5787	0.6315	0.6281	0.6384	0.6011	0.6253	0.6220	0.6347
06	0.3729	0.3655	**0.3764**	0.3677	0.2019	0.2295	0.3376	0.1881	0.2246	0.1985	0.3537	0.2017	0.2391	0.2146
07	0.3349	0.0874	0.1041	0.0882	0.4112	**0.4226**	0.2767	0.0189	0.0190	0.0188	0.0769	0.0182	0.0183	0.0180
08	0.0431	0.0448	0.0414	0.0399	0.0379	0.0382	**0.0597**	0.0368	0.0365	0.0551	0.0548	0.0372	0.0369	0.0550
09	0.3707	0.3857	0.2621	0.2877	0.4300	0.3955	0.4025	**0.4324**	0.3772	0.3986	0.3104	0.4291	0.3687	0.3947
10	0.5273	0.6689	0.6049	0.6798	0.7245	0.6757	0.5311	0.7334	0.8525	0.8586	0.6841	0.7516	0.8577	**0.8653**
11	0.0841	0.0617	0.0615	0.0619	0.1333	0.1327	0.0853	0.1560	0.1566	**0.1591**	0.0639	0.1548	0.1548	0.1571
12	0.0237	0.1480	0.1367	0.1484	0.0057	0.0060	0.0229	0.2161	0.2158	0.2442	0.1481	0.2204	0.2214	**0.2516**
13	0.2520	0.3913	0.4170	0.4156	0.4568	0.4312	0.3274	0.2669	0.4737	0.4795	0.4286	0.3009	0.4867	**0.4943**
14	0.2393	0.2162	0.2640	0.2228	0.5893	0.5861	0.2267	0.6400	0.6437	0.4455	0.2740	0.6443	**0.6492**	0.4485
15	0.6716	0.7088	0.6822	0.7138	0.7784	0.7801	0.6839	0.7818	0.7940	**0.7956**	0.7230	0.7794	0.7879	0.7897
16	0.8331	0.8688	0.8485	0.8695	0.8760	0.8738	0.8038	**0.9312**	**0.9312**	0.8099	0.7942	0.9284	0.9270	0.8149
17	0.3841	0.4417	0.4411	**0.4426**	0.4088	0.4087	0.3681	0.4125	0.4285	0.4132	0.4220	0.4121	0.4339	0.4182
18	0.1688	0.3118	0.2972	0.3302	0.1238	0.1244	0.1969	0.2753	0.2805	0.3185	0.3620	0.3125	0.3238	**0.3650**
19	0.1415	0.1593	0.1653	0.1674	0.1630	0.1571	0.1831	0.1961	0.2306	**0.2605**	0.2134	0.2004	0.2221	0.2550
20	0.2510	0.1268	**0.3296**	0.1684	0.1199	0.1162	0.0060	0.0339	0.0358	0.0025	0.0030	0.0502	0.0370	0.0023
21	0.3690	0.5225	0.5671	0.5667	0.4055	0.4039	0.4179	0.4329	0.4304	**0.6392**	0.5548	0.4399	0.4369	0.6352
22	0.3461	0.3600	0.3443	0.3566	0.4315	0.4240	0.3332	0.4623	0.4296	0.4780	0.4194	0.4588	0.4191	**0.4782**
23	0.0118	0.0194	0.0281	0.0268	0.0069	0.0080	0.0325	0.0189	0.0358	0.1295	0.0570	0.0179	0.0342	**0.1329**
24	0.5271	0.5587	0.5791	0.5620	0.5983	0.5959	0.5220	0.6286	**0.6313**	0.6197	0.5556	0.6278	0.6267	0.6148
25	0.4444	0.4583	0.4444	0.4444	**0.9167**	0.8667	0.4778	0.8095	0.8095	0.8333	0.4583	0.8095	0.8095	0.8333
26	0.0140	0.2763	0.2775	0.2768	0.0289	0.0282	0.0183	0.4118	0.4128	**0.4508**	0.3241	0.4077	0.4078	0.4443
27	0.0398	0.0228	0.0219	0.0219	0.0384	0.0473	0.0415	0.0634	0.0635	0.0640	0.0238	0.0655	0.0656	**0.0658**
28	0.0340	0.0828	0.0832	0.0829	0.0839	0.0842	0.0005	0.3256	**0.3399**	0.0092	0.0020	0.3207	0.3280	0.0081
29	0.1837	0.2344	**0.2355**	0.2310	0.1573	0.1552	0.1744	0.1252	0.1252	0.1200	0.2199	0.1300	0.1305	0.1264
30	0.2979	**1.0000**	0.5520	**1.0000**	0.5696	0.6100	0.5058	0.9583	0.9583	0.6393	0.6727	0.9583	0.9583	0.6393
31	0.3786	0.3595	0.3542	0.3667	0.4284	0.4236	0.4018	0.6476	0.6525	**0.6575**	0.3931	0.6332	0.6354	0.6405
32	0.8646	0.9281	0.8854	0.9250	0.9371	0.9381	0.8302	0.9641	0.9641	0.9336	0.8908	**0.9669**	**0.9669**	0.9364
33	**0.0048**	0.0030	0.0041	0.0030	0.0040	0.0040	0.0028	0.0017	0.0021	0.0025	0.0035	0.0019	0.0022	0.0024
34	0.5889	0.4583	0.5556	0.4583	0.4583	0.5139	**0.8095**	0.5000	0.4889	0.7436	0.7576	0.5000	0.5000	0.7500
35	**0.1230**	0.1064	0.1081	0.1064	0.0468	0.0430	0.0551	0.0604	0.0686	0.0520	0.0546	0.0661	0.0696	0.0520
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0800	0.0533	**0.1021**	0.0505	**0.1021**	0.1008	0.0055	0.0869	0.0664	0.0386	0.0038	0.0803	0.0665	0.0282
38	0.0102	0.0385	0.0417	0.0400	0.0256	0.0294	0.0012	0.0588	0.0588	0.0011	0.0012	**0.0714**	**0.0714**	0.0011
39	0.0616	0.0729	0.0710	0.0730	0.0476	0.0472	0.0668	0.0487	0.0488	0.0490	**0.0793**	0.0479	0.0484	0.0486
40	0.1472	0.2747	0.2641	**0.2762**	0.2443	0.2479	0.1472	0.2473	0.2485	0.2485	**0.2762**	0.2491	0.2491	0.2491
41	0.0003	0.0003	0.0003	0.0003	0.0003	0.0000	0.0008	**0.0115**	0.0052	0.0027	0.0007	0.0109	0.0041	0.0026
42	0.0513	0.0478	0.0947	0.0478	**0.1190**	0.1010	0.0972	0.0075	0.0076	0.0552	0.0942	0.0075	0.0075	0.0558
43	0.0023	0.0157	0.0091	0.0154	0.0031	0.0031	0.0039	0.0165	0.0163	0.0378	0.0257	0.0165	0.0160	**0.0380**
44	0.2112	0.2255	0.1994	0.2270	**0.2438**	0.2380	0.2223	0.1805	0.1793	0.1684	0.2270	0.1830	0.1826	0.1723
45	0.0726	0.0682	0.0682	0.0692	0.1816	0.1773	0.0784	**0.2068**	0.2061	0.2061	0.0764	0.1993	0.1949	0.1949
46	0.8095	0.7121	0.7121	0.7121	0.2879	0.3178	**1.0000**	0.7381	0.7381	0.7778	0.7500	0.7292	0.7292	0.7917
47	0.0324	0.0564	**0.0791**	0.0557	0.0484	0.0474	0.0324	0.0719	0.0739	0.0739	0.0557	0.0688	0.0686	0.0686
48	0.7512	0.7546	0.7436	0.7552	0.8904	0.8978	0.5466	0.9071	0.9082	0.5998	0.5487	0.9077	**0.9084**	0.5996
49	0.6000	0.6000	0.6000	0.6000	0.3929	0.3929	0.5000	**0.6429**	**0.6429**	0.5000	0.5000	**0.6429**	0.6250	0.5000
50	0.1835	0.2520	0.2894	0.2530	0.2434	0.2463	0.3225	0.2140	0.2121	**0.5503**	0.4107	0.2075	0.2088	0.5382
51	0.5516	0.5001	0.5385	0.4994	0.6759	0.6809	0.5620	0.6832	0.6098	0.6235	0.5106	**0.6905**	0.6086	0.6239
52	0.0104	0.0084	0.0093	0.0075	0.0060	0.0061	0.0136	0.0108	0.0101	**0.0181**	0.0165	0.0130	0.0113	0.0180
53	0.1280	0.1351	0.1356	0.1365	0.1311	0.1380	0.1411	0.1464	0.1461	0.1945	0.1583	0.1499	0.1507	**0.1980**
54	0.0922	0.1089	0.0869	0.1102	0.0969	0.0962	0.1139	0.1184	0.1206	0.1460	0.1285	0.1267	0.1286	**0.1575**
55	0.0441	0.0808	0.1303	0.0818	0.1988	0.1933	0.0501	0.1879	**0.2019**	0.1990	0.1001	0.1922	0.1989	0.1974
56	0.0170	0.2008	0.1178	0.2056	0.0142	0.0145	0.0170	**0.4861**	**0.4861**	**0.4861**	0.2056	0.4174	0.4259	0.4259
57	0.1791	0.1999	0.1717	0.1888	0.1903	0.1861	0.1928	0.1933	0.2104	0.2129	0.2009	0.1962	0.2099	**0.2174**
58	0.0320	0.0235	0.0281	0.0230	0.0397	0.0378	0.0481	0.0716	0.0703	**0.0791**	0.0368	0.0590	0.0676	0.0785
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7647	0.7764	0.7780	0.7775	0.7636	0.7690	0.7638	0.7754	0.7768	0.7724	0.7738	0.7795	**0.7806**	0.7763
61	0.0943	0.1568	0.1673	0.1630	0.0215	0.0199	0.1679	0.0556	0.0406	0.3995	0.3364	0.0570	0.0487	**0.4069**
62	0.3346	0.3532	0.3154	0.3537	0.3177	0.3250	0.3370	0.3186	0.1995	0.1995	**0.3558**	0.1947	0.1512	0.1512
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0025	0.0105	0.0077	0.0101	0.0083	0.0041	0.0043	0.0089	0.0064	0.0086	**0.0136**	0.0084	0.0062	0.0086
65	0.3129	**0.4539**	0.4380	0.4342	0.3326	0.3341	0.3243	0.3797	0.3793	0.3795	0.4413	0.3947	0.4005	0.4010
66	0.1667	0.3362	**0.3366**	0.3363	0.0152	0.0104	0.1667	0.3357	0.3355	0.3355	0.3363	0.3351	0.3351	0.3351
67	0.3555	0.3969	0.3858	0.3972	0.3526	0.3494	0.1819	**0.4132**	0.4085	0.1716	0.1768	0.3997	0.3995	0.1727
68	0.3912	0.7530	0.7222	0.7486	0.6121	0.5870	0.3912	0.7500	0.7551	0.7551	0.7487	0.7526	**0.7578**	0.7577
69	0.0733	0.1307	0.1074	0.1308	0.3648	**0.3678**	0.0733	0.2913	0.2907	0.2907	0.1308	0.2769	0.2762	0.2762
70	0.0253	0.0295	0.0175	0.0311	0.0073	0.0070	**0.0808**	0.0142	0.0140	0.0717	0.0693	0.0107	0.0109	0.0671
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.5085	0.6886	0.6728	0.6898	0.4930	0.4891	0.5085	0.7116	0.7124	0.7124	0.6898	0.7111	**0.7127**	**0.7127**
73	0.0215	0.0203	0.0210	0.0205	0.0186	0.0197	0.0084	**0.0328**	0.0318	0.0101	0.0098	0.0314	0.0319	0.0102
74	0.3348	0.3522	0.3359	0.3512	0.3340	0.3333	0.3709	0.3426	0.3420	0.4601	**0.4952**	0.3439	0.3429	0.4621
75	0.4004	0.4160	0.4135	0.4134	**0.6304**	0.6015	0.5140	0.6084	0.6030	0.5828	0.4774	0.5884	0.5880	0.5693
76	0.0000	0.4400	0.2488	0.4400	0.0000	0.0000	0.0000	0.4621	**0.4688**	0.2800	0.2800	**0.4688**	**0.4688**	0.2800
77	0.2876	0.2929	0.2774	0.2750	0.3682	0.3743	0.2876	**0.3994**	0.3924	0.3924	0.2750	0.3960	0.3819	0.3819
78	0.0208	0.1087	0.0822	0.1087	0.0549	0.0567	0.0503	0.2266	**0.2267**	0.2057	0.0939	0.2249	0.2250	0.2026
79	0.9762	0.9583	0.9762	0.9583	**1.0000**	**1.0000**	0.9762	0.9762	0.9762	0.9762	0.9583	0.9762	0.9762	0.9762
80	0.0113	0.1016	0.1267	0.1266	**0.1964**	**0.1964**	0.0141	0.1526	0.1714	0.1769	0.1703	0.1455	0.1556	0.1588
81	0.0681	0.0737	0.0699	0.0731	0.0833	0.0865	0.0681	**0.0938**	0.0935	0.0935	0.0731	0.0926	0.0923	0.0923
82	0.0013	0.0000	0.0013	0.0000	**0.0025**	0.0022	0.0011	0.0023	0.0023	0.0019	0.0000	0.0017	0.0017	0.0016
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.3948	0.3209	0.2083	0.3206	0.1871	0.1890	**0.4008**	0.0722	0.0722	0.0729	0.3211	0.0786	0.0766	0.0785
85	0.3820	0.3293	0.3532	0.3290	**0.5138**	0.4950	0.3399	0.4700	0.4701	0.3948	0.2914	0.4575	0.4568	0.3895
86	0.0275	0.0279	0.0163	0.0281	0.0795	0.0795	0.0436	0.1739	0.1746	**0.1823**	0.0391	0.1628	0.1724	0.1805
87	0.1624	0.1944	0.1930	0.1931	0.1866	0.1877	0.1624	**0.2171**	0.2156	0.2156	0.1931	0.2133	0.2125	0.2125
88	0.0812	0.3113	0.2889	0.3103	0.1625	0.1931	0.1520	0.3767	0.3746	0.2715	0.2727	**0.3795**	0.3788	0.2724
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90	0.1170	**0.1639**	0.0902	0.1638	0.0942	0.0980	0.1170	0.0895	0.0904	0.0904	0.1638	0.0992	0.1002	0.1002
91	0.1250	0.5053	0.2557	**0.5054**	0.0237	0.0437	0.2500	0.1750	0.1754	0.5011	0.5000	0.1757	0.2592	0.5012
92	0.0968	0.0795	0.0997	0.0821	0.1425	0.1427	0.1981	0.1228	0.1269	**0.2663**	0.2024	0.1178	0.1212	0.2604
93	0.3451	0.7967	0.7962	0.7972	0.8410	0.8473	0.3848	0.8933	0.8904	0.8733	0.8170	**0.8973**	0.8972	0.8797
94	0.1158	0.2449	0.1846	0.2528	0.0499	0.0525	0.1221	0.0735	0.0727	0.0730	**0.2572**	0.0700	0.0691	0.0694
95	0.4195	0.4319	0.4431	0.4299	0.5619	0.5510	0.3193	**0.6549**	0.6536	0.3815	0.3116	0.6436	0.6423	0.3786
96	0.1863	0.2531	0.2029	0.2530	0.2394	0.2324	0.1991	**0.2702**	0.2698	0.2469	0.2665	0.2603	0.2596	0.2390
97	0.0358	0.0364	0.0371	0.0364	0.0390	0.0390	0.0323	**0.0413**	**0.0413**	0.0388	0.0330	0.0402	0.0402	0.0377
98	0.1768	0.2655	0.2095	0.2662	0.2430	0.2365	0.2320	0.7249	0.7249	**0.8500**	0.3836	0.7221	0.7213	**0.8500**
99	0.0979	0.1462	0.1040	0.1479	**0.2233**	0.2183	0.0979	0.1686	0.1702	0.1702	0.1479	0.1819	0.1822	0.1822
100	0.0156	0.0222	0.0241	0.0222	0.0352	0.0354	0.0060	0.0379	**0.0383**	0.0200	0.0119	0.0348	0.0347	0.0181
all	0.2240	0.2745	0.2612	0.2748	0.2595	0.2592	0.2307	0.3052	**0.3067**	0.3052	0.2745	0.3041	0.3061	0.3047

Table C.8: TFIDF per query results (TDN. title,description and narrative tags used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.5912	0.6351	0.6247	0.5987	0.6064	0.6102	0.6096	0.6389	0.6448	**0.6696**	0.6011	0.6579	0.6445	0.6656
02	0.0200	0.0447	0.0151	0.0299	0.0064	0.0065	0.0376	0.0052	0.0066	0.0146	**0.0547**	0.0061	0.0074	0.0162
03	0.2733	0.2378	0.2474	0.2309	**0.4122**	0.4000	0.2744	0.2929	0.3311	0.3314	0.2311	0.3100	0.3312	0.3316
04	0.3217	0.3136	**0.3243**	0.2969	0.3137	0.3230	0.2581	0.2875	0.2621	0.2094	0.2443	0.2861	0.2761	0.2161
05	0.6445	0.6759	0.6792	0.6726	0.7125	0.7099	0.6559	0.7049	0.7010	0.7094	0.6801	0.7048	0.7039	**0.7169**
06	0.3622	0.4491	0.3837	**0.4513**	0.2277	0.2702	0.3213	0.3138	0.3163	0.2841	0.4431	0.2947	0.2961	0.2994
07	**0.3725**	0.1378	0.1610	0.1381	0.3056	0.3319	0.3104	0.0495	0.0501	0.0457	0.1204	0.0439	0.0441	0.0403
08	0.0556	0.0423	0.0469	0.0416	0.0367	0.0352	**0.0862**	0.0363	0.0290	0.0446	0.0532	0.0374	0.0299	0.0463
09	0.3876	0.3925	0.2960	0.3017	0.4106	0.3944	**0.4280**	0.4176	0.3523	0.3904	0.3518	0.4091	0.3582	0.4003
10	0.7431	0.8467	0.7008	0.8522	0.7661	0.7771	0.8533	0.7956	0.8091	**0.9213**	0.9067	0.8000	0.8050	0.9210
11	0.0679	0.0704	0.0734	0.0709	0.1069	0.1056	0.0703	0.1617	0.1616	0.1639	0.0796	0.1605	0.1615	**0.1640**
12	0.0558	0.2267	0.2263	0.2291	0.0359	0.0435	0.0545	0.2490	0.2495	0.2684	0.2409	0.2471	0.2479	**0.2690**
13	0.4540	0.4635	0.4774	0.4764	0.4834	0.4871	0.4631	0.4303	0.4303	0.4353	**0.4943**	0.4481	0.4659	0.4727
14	0.2865	0.2580	0.3055	0.2703	0.6042	0.5951	0.2599	0.6234	**0.6293**	0.4360	0.3314	0.6227	0.6279	0.4360
15	0.6267	0.6726	0.6238	0.6713	0.7794	0.7810	0.6940	0.8081	0.8057	**0.8091**	0.7229	0.8047	0.8032	0.8060
16	0.8480	0.8949	0.8650	0.8949	0.8882	0.8797	0.8719	0.9240	0.9252	**0.9284**	0.9073	0.9193	0.9171	0.9272
17	0.4232	0.3955	**0.4447**	0.3969	0.4219	0.4224	0.4187	0.3574	0.3643	0.3597	0.3929	0.3635	0.3653	0.3606
18	0.2680	0.3181	0.2899	0.3114	0.1548	0.1960	0.2958	0.3614	0.3366	0.3542	0.3353	**0.3769**	0.3500	0.3706
19	0.1384	0.1259	0.1241	0.1334	0.1882	0.1744	0.1775	0.1745	0.1800	**0.2223**	0.1848	0.1650	0.1707	0.2130
20	0.1926	0.1921	**0.2565**	0.2055	0.0304	0.0293	0.0119	0.0301	0.0283	0.0021	0.0042	0.0375	0.0328	0.0020
21	0.3913	0.4945	0.5402	0.5314	0.3481	0.3421	0.3886	0.4354	0.3977	0.6320	0.5273	0.4487	0.4073	**0.6338**
22	0.3231	0.3126	0.2806	0.3250	0.3204	0.3386	0.3120	0.3559	0.3838	0.4396	0.3960	0.3553	0.3875	**0.4433**
23	0.0350	0.0456	0.0546	0.0516	0.0218	0.0205	0.0543	0.0457	0.0508	**0.1801**	0.1065	0.0457	0.0508	0.1753
24	0.5023	0.5028	0.4931	0.4985	0.5309	0.5320	0.5065	0.5604	0.5586	0.5618	0.5027	0.5599	0.5594	**0.5627**
25	0.3278	0.3167	0.3444	0.3167	**0.7917**	0.7778	0.3278	0.5909	0.5909	0.5909	0.3167	0.5909	0.4242	0.4242
26	0.0665	0.2090	0.2092	0.2116	0.0652	0.0709	0.0825	0.2676	0.2692	0.2926	0.2541	0.2699	0.2707	**0.2955**
27	0.0329	0.0232	0.0245	0.0237	0.0158	0.0205	**0.0347**	0.0087	0.0085	0.0090	0.0268	0.0100	0.0095	0.0100
28	0.0083	**0.0474**	0.0247	0.0443	0.0007	0.0007	0.0084	0.0073	0.0072	0.0072	0.0447	0.0055	0.0054	0.0054
29	0.2002	0.2560	**0.2700**	0.2511	0.1795	0.1724	0.1854	0.1937	0.1941	0.1880	0.2315	0.1922	0.1902	0.1852
30	0.2181	0.8965	0.3520	0.8965	0.4101	0.4136	0.4932	**0.9444**	**0.9444**	0.8346	0.8068	**0.9444**	**0.9444**	0.8347
31	0.3294	0.3325	0.3276	0.3391	0.2561	0.2585	0.3453	0.4393	0.4588	**0.4668**	0.3567	0.4140	0.4353	0.4432
32	0.8563	0.9355	0.8717	0.9365	0.9439	0.9472	0.8266	0.9652	0.9659	0.9365	0.9058	0.9655	**0.9680**	0.9385
33	0.0080	0.0063	0.0027	0.0060	0.0012	0.0024	**0.0208**	0.0023	0.0022	0.0104	0.0166	0.0024	0.0019	0.0105
34	0.4111	0.3417	0.2708	0.3418	0.4333	0.4333	**0.6692**	0.2118	0.2118	0.6678	0.6677	0.2108	0.2266	0.6678
35	0.0414	0.0366	0.0405	0.0372	0.0250	0.0260	**0.0445**	0.0258	0.0258	0.0276	0.0394	0.0246	0.0248	0.0267
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0025	**0.0032**	0.0025	0.0029	0.0007	0.0007	0.0025	0.0012	0.0012	0.0012	**0.0032**	0.0012	0.0012	0.0012
38	0.0097	0.0128	0.0154	0.0145	0.0038	0.0039	0.0105	0.0050	0.0059	0.0060	**0.0161**	0.0049	0.0057	0.0058
39	0.2721	0.3045	0.3023	0.3045	0.2418	0.2365	0.2724	0.4248	0.4450	**0.4451**	0.3053	0.4319	0.4323	0.4323
40	0.2335	**0.2665**	0.2538	0.2664	0.2642	0.2649	0.2335	0.2433	0.2397	0.2397	0.2664	0.2452	0.2438	0.2438
41	0.0003	0.0007	0.0008	0.0007	0.0003	0.0003	0.0008	**0.0287**	0.0223	0.0136	0.0007	0.0236	0.0207	0.0131
42	0.3269	0.6111	0.7000	0.6111	0.7000	0.6667	0.3500	**0.8333**	**0.8333**	**0.8333**	0.6111	**0.8333**	**0.8333**	**0.8333**
43	0.0052	0.0220	0.0107	0.0210	0.0074	0.0070	0.0010	0.0232	0.0215	0.0077	0.0083	**0.0233**	0.0209	0.0079
44	0.1815	0.1629	0.1679	0.1639	0.1779	0.1849	**0.1941**	0.1278	0.1465	0.1479	0.1707	0.1497	0.1516	0.1520
45	0.2525	0.2577	0.2428	0.2586	**0.7592**	0.7487	0.2576	0.7140	0.7157	0.7157	0.2670	0.6991	0.7001	0.7001
46	0.8095	0.7101	0.7222	0.7121	0.8095	0.8095	**0.9167**	0.5833	0.7292	0.7576	0.7500	0.6000	0.7167	0.7667
47	0.0329	0.0232	0.0309	0.0231	0.0499	**0.0570**	0.0335	0.0421	0.0422	0.0430	0.0263	0.0504	0.0502	0.0509
48	0.7857	0.7934	0.7659	0.7927	0.8867	0.8937	0.7794	0.9003	0.9008	0.8720	0.7932	0.9011	**0.9016**	0.8704
49	0.5625	0.6111	0.6111	0.6111	0.5625	0.5625	0.5000	**0.6429**	**0.6429**	0.5000	0.5000	**0.6429**	**0.6429**	0.5000
50	0.1344	0.3160	0.3165	0.3217	0.1783	0.1773	0.1425	0.3003	0.3138	0.3302	**0.3338**	0.3012	0.3113	0.3280
51	0.7447	0.7172	0.7301	0.7145	0.7478	0.7478	0.7755	0.7683	0.7653	**0.7775**	0.7525	0.7618	0.7597	0.7763
52	0.0061	0.0018	0.0037	0.0018	0.0047	0.0047	0.0075	0.0008	0.0021	**0.0104**	0.0070	0.0007	0.0019	0.0094
53	0.1365	0.2100	0.2094	0.2109	0.1266	0.1297	0.1331	0.2563	0.2563	0.2370	0.1967	0.2580	**0.2581**	0.2406
54	0.0054	0.0207	0.0124	0.0208	0.0014	0.0015	0.0055	0.0025	0.0025	0.0026	**0.0228**	0.0033	0.0026	0.0026
55	0.0831	0.1554	0.2137	0.1561	0.2079	0.2113	0.0827	**0.2267**	0.2239	0.2163	0.1637	0.2209	0.2185	0.2100
56	0.1931	0.2932	0.2487	**0.2957**	0.0878	0.0815	0.1931	0.2163	0.2296	0.2296	**0.2957**	0.1973	0.2043	0.2043
57	0.4506	0.4440	0.4353	0.4468	**0.4783**	0.4731	0.3672	0.4238	0.4244	0.3805	0.3728	0.4174	0.4210	0.3784
58	0.1172	0.1078	0.1223	0.1038	0.1273	0.1304	0.1312	**0.1474**	0.1221	0.1316	0.1108	0.1458	0.1069	0.1159
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7805	0.7808	0.7795	0.7841	0.7709	0.7758	0.7946	0.7811	0.7818	0.7429	**0.8017**	0.7909	0.7932	0.7549
61	0.0821	0.1017	0.1025	0.1003	0.0165	0.0158	0.1467	0.0385	0.0367	0.4878	0.2246	0.0389	0.0411	**0.4895**
62	0.0502	**0.0521**	0.0517	0.0506	0.0374	0.0392	0.0504	0.0436	0.0431	0.0432	0.0506	0.0437	0.0434	0.0435
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0185	0.1178	0.0644	0.1102	0.0222	0.0225	0.0288	0.0647	0.0662	0.0699	**0.1242**	0.0586	0.0584	0.0637
65	0.2514	0.4236	0.4120	0.4231	0.3748	0.3747	0.3122	0.4141	0.4549	0.4580	0.4570	0.4314	0.4745	**0.4775**
66	0.0476	**0.1759**	0.0882	0.1208	0.0062	0.0045	0.0530	0.0439	0.0439	0.1495	0.1435	0.0356	0.0393	0.1522
67	0.4007	0.4005	0.3882	0.3992	0.4209	**0.4282**	0.1894	0.4082	0.3994	0.1798	0.1889	0.4031	0.3989	0.1878
68	0.3605	0.7371	0.7064	0.7368	0.5573	0.5659	0.3605	0.7556	**0.7577**	**0.7577**	0.7369	0.7553	**0.7577**	0.7576
69	0.3260	0.2079	0.3352	0.2057	**0.3589**	0.3548	0.2737	0.2980	0.2939	0.2237	0.1976	0.2960	0.2870	0.2226
70	0.0200	0.0114	0.0112	0.0114	0.0052	0.0054	**0.0978**	0.0041	0.0043	0.0633	0.0743	0.0036	0.0035	0.0624
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.5217	0.6128	**0.6469**	0.6124	0.4270	0.4513	0.5217	0.6283	0.6379	0.6379	0.6124	0.6310	0.6437	0.6437
73	0.0151	0.0157	0.0164	0.0158	0.0143	0.0146	0.0195	0.0133	0.0135	0.0202	**0.0218**	0.0136	0.0139	0.0206
74	0.3401	0.4368	0.3601	0.4247	0.1149	0.3371	0.4035	0.4489	0.4492	**0.7292**	0.6465	0.4324	0.4486	**0.7292**
75	0.3620	0.2618	0.3252	0.2613	**0.6241**	0.6072	0.5015	0.6107	0.6184	0.6055	0.4192	0.5966	0.5978	0.5886
76	0.0000	**0.0140**	0.0071	0.0137	0.0000	0.0000	0.0000	0.0017	0.0017	0.0010	0.0059	0.0014	0.0014	0.0009
77	0.0243	**0.0252**	0.0216	0.0232	0.0004	0.0033	0.0215	0.0028	0.0028	0.0028	0.0203	0.0024	0.0023	0.0024
78	0.1201	0.1526	0.1488	0.1557	0.2447	0.2353	0.1290	**0.2715**	0.2672	0.2667	0.1674	0.2566	0.2593	0.2614
79	0.9583	0.9583	0.9583	0.9583	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**
80	0.0085	0.0472	0.0432	0.0473	0.1534	0.1429	0.0171	0.1917	0.2333	**0.2500**	0.0770	0.1806	0.1806	0.1917
81	0.0394	0.0473	0.0456	0.0442	0.0560	0.0619	0.0412	0.0806	0.0909	0.1053	0.0472	0.0779	0.0906	**0.1071**
82	0.0009	0.0007	0.0011	0.0007	0.0017	0.0009	**0.0141**	0.0018	0.0017	0.0037	0.0094	0.0015	0.0015	0.0044
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.4676	0.3282	0.2099	0.3281	0.3712	0.3834	**0.4735**	0.1742	0.1650	0.1773	0.3293	0.1741	0.1772	0.1904
85	0.3833	0.3631	0.3747	0.3599	**0.5047**	0.4967	0.3408	0.4715	0.4507	0.3825	0.3199	0.4458	0.4417	0.3798
86	0.1238	0.0641	0.0597	0.0653	0.0778	0.0826	0.1455	0.1481	0.1772	**0.1885**	0.0712	0.1396	0.1604	0.1686
87	0.0296	0.0519	0.0502	0.0516	0.0045	0.0047	0.0307	0.0040	0.0040	0.0040	**0.0527**	0.0123	0.0123	0.0124
88	0.0219	0.1514	0.1473	0.1431	0.0028	0.0027	0.1329	0.1099	0.1081	**0.2529**	0.2509	0.1109	0.1089	0.2431
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0011**	0.0000	0.0000	0.0000	0.0010
90	**0.1134**	0.1091	0.0828	0.1096	0.0975	0.0977	**0.1134**	0.0838	0.0851	0.0851	0.1096	0.0878	0.0907	0.0907
91	0.1667	0.1700	0.1343	0.1700	0.0625	0.0714	0.1711	0.0648	0.0648	0.0709	**0.1793**	0.0641	0.0641	0.0701
92	0.0424	0.0401	0.0435	0.0406	0.0066	0.0053	0.0444	0.0445	0.0451	0.0489	0.0436	0.0487	0.0495	**0.0543**
93	0.3111	0.8198	0.8000	0.8190	0.8476	0.8518	0.3607	0.9093	0.9084	0.8937	0.8315	**0.9105**	**0.9105**	0.8954
94	0.1168	0.2566	0.2035	0.2561	0.1520	0.1479	0.1260	0.0558	0.0558	0.0579	**0.3156**	0.0834	0.0833	0.0858
95	0.4780	0.4944	0.4941	0.5012	0.6010	0.6021	0.4668	0.5985	0.5985	0.5748	0.4845	0.6004	**0.6029**	0.5797
96	0.2702	0.3003	0.2903	0.3015	0.2682	0.2668	0.2799	0.2921	0.2922	0.2954	**0.3077**	0.2892	0.2900	0.2937
97	0.0454	0.0451	0.0466	0.0455	0.0458	0.0436	0.0461	**0.0530**	0.0464	0.0468	0.0458	0.0499	0.0453	0.0456
98	0.1027	0.1712	0.1359	0.1767	0.0357	0.0348	0.1080	0.0458	0.3230	0.3067	0.1787	0.0467	**0.3248**	0.3092
99	0.2168	0.1911	0.1567	0.1931	**0.3519**	0.3443	0.2169	0.1978	0.2110	0.2110	0.1931	0.2061	0.2221	0.2221
100	0.0125	0.0873	0.0564	0.0816	0.0505	0.0493	0.0122	**0.1186**	0.1100	0.1025	0.0781	0.1167	0.1091	0.1019
all	0.2386	0.2742	0.2619	0.2728	0.2692	0.2722	0.2485	0.2908	0.2959	**0.3082**	0.2833	0.2906	0.2936	0.3066

Table C.9: BM25 per query results (TDN. title, description and narrative tags used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6040	0.6502	0.6495	0.6386	0.6011	0.6112	0.6056	0.6421	0.6504	**0.6761**	0.6231	0.6618	0.6524	0.6756
02	0.0201	0.0513	0.0149	0.0310	0.0063	0.0065	0.0381	0.0064	0.0064	0.0143	**0.0594**	0.0068	0.0073	0.0160
03	0.2857	0.2340	0.2484	0.2276	0.4127	**0.4281**	0.2870	0.3024	0.3750	0.3757	0.2278	0.3193	0.3841	0.3849
04	0.3285	0.3165	**0.3355**	0.3060	0.3159	0.3205	0.2611	0.2817	0.2776	0.2213	0.2515	0.2877	0.2817	0.2215
05	0.6571	0.6621	0.6820	0.6676	0.7115	0.7094	0.6651	0.7092	0.7133	**0.7245**	0.6738	0.7127	0.7075	0.7217
06	0.3507	0.4498	0.3943	**0.4569**	0.2255	0.2657	0.3182	0.3359	0.3151	0.2801	0.4552	0.2934	0.2984	0.3030
07	**0.3607**	0.1354	0.1561	0.1354	0.2958	0.3254	0.3009	0.0614	0.0607	0.0561	0.1183	0.0742	0.0745	0.0665
08	0.0573	0.0421	0.0458	0.0395	0.0360	0.0350	**0.0898**	0.0333	0.0329	0.0501	0.0504	0.0344	0.0337	0.0518
09	0.3870	0.3894	0.2681	0.2793	0.4164	0.3973	**0.4275**	0.4246	0.3505	0.3832	0.3203	0.4078	0.3516	0.3877
10	0.7405	0.8476	0.7201	0.8503	0.7655	0.7623	0.8554	0.7968	0.7986	0.9123	0.9018	0.7987	0.8031	**0.9223**
11	0.0674	0.0466	0.0468	0.0465	**0.1070**	0.1062	0.0696	0.0876	0.0879	0.0918	0.0517	0.0838	0.0839	0.0882
12	0.0589	0.2257	0.2257	0.2296	0.0658	0.0677	0.0571	0.2471	0.2508	0.2680	0.2413	0.2471	0.2514	**0.2720**
13	0.4972	0.4756	0.4699	0.4770	0.4086	0.4853	**0.5066**	0.4270	0.4277	0.4330	0.4903	0.4431	0.4509	0.4563
14	0.2734	0.2598	0.2956	0.2683	0.6104	0.6134	0.2613	0.6229	**0.6267**	0.4343	0.3339	0.6225	0.6265	0.4368
15	0.6339	0.6764	0.6286	0.6740	0.7951	0.7943	0.6949	0.8083	0.8130	**0.8146**	0.7230	0.8057	0.8107	0.8119
16	0.8407	0.8968	0.8658	0.8967	0.8912	0.8847	0.8651	**0.9306**	0.9245	0.9277	0.9090	0.9244	0.9197	0.9272
17	0.4051	0.3737	**0.4237**	0.3734	0.4030	0.4044	0.4007	0.3447	0.3496	0.3450	0.3696	0.3515	0.3523	0.3477
18	0.2415	0.3146	0.2969	0.3248	0.1316	0.1382	0.2738	0.3005	0.3022	0.3232	0.3466	0.3423	0.3281	**0.3607**
19	0.1383	0.1313	0.1305	0.1412	0.1970	0.1800	0.1774	0.1777	0.1850	**0.2242**	0.1874	0.1718	0.1804	0.2194
20	0.1898	0.1967	**0.2603**	0.2088	0.0304	0.0331	0.0147	0.0318	0.0256	0.0024	0.0054	0.0411	0.0352	0.0023
21	0.3949	0.4914	0.5380	0.5377	0.4084	0.4086	0.4014	0.4012	0.3862	0.6171	0.5262	0.4183	0.4106	**0.6294**
22	0.3075	0.2902	0.2605	0.2951	0.3308	0.3401	0.3204	0.2869	0.2771	**0.3831**	0.3823	0.2851	0.2789	0.3751
23	0.0370	0.0510	0.0601	0.0570	0.0161	0.0173	0.0638	0.0488	0.0496	**0.1799**	0.1249	0.0458	0.0503	0.1791
24	0.5060	0.5082	0.4899	0.5073	0.5304	0.5327	0.5104	0.5635	0.5616	0.5645	0.5114	0.5633	0.5622	**0.5657**
25	0.3167	0.3076	0.3354	0.3076	**0.7917**	0.7778	0.3167	0.4048	0.3492	0.3492	0.3076	0.4048	0.4048	0.4048
26	0.0605	0.2092	0.2122	0.2161	0.0649	0.0708	0.0736	0.2689	0.2709	0.2945	0.2634	0.2715	0.2720	**0.2975**
27	0.0362	0.0236	0.0201	0.0192	0.0108	0.0121	**0.0374**	0.0077	0.0077	0.0082	0.0213	0.0083	0.0081	0.0092
28	0.0085	**0.0588**	0.0286	0.0554	0.0008	0.0007	0.0085	0.0077	0.0076	0.0076	0.0574	0.0064	0.0064	0.0064
29	0.1951	0.2546	**0.2650**	0.2501	0.1806	0.1735	0.1822	0.1951	0.1936	0.1857	0.2308	0.1918	0.1920	0.1861
30	0.3004	0.9167	0.3695	0.9167	0.4690	0.4224	0.4636	**1.0000**	**1.0000**	0.8345	0.8068	**1.0000**	**1.0000**	0.8346
31	0.3213	0.3295	0.3237	0.3347	0.2411	0.2439	0.3383	0.2817	0.2777	0.2850	**0.3519**	0.2822	0.2841	0.2912
32	0.8634	0.9383	0.8687	0.9402	0.9460	0.9472	0.8319	0.9659	0.9659	0.9365	0.9096	0.9669	**0.9673**	0.9378
33	0.0079	0.0058	0.0028	0.0055	0.0014	0.0024	**0.0210**	0.0020	0.0028	0.0101	0.0162	0.0023	0.0024	0.0101
34	0.3909	0.3440	0.2768	0.3441	0.4103	0.4242	**0.6693**	0.1997	0.2119	0.6678	0.6689	0.2108	0.2108	0.6678
35	0.0411	0.0372	0.0392	0.0363	0.0245	0.0252	**0.0440**	0.0228	0.0229	0.0248	0.0385	0.0229	0.0229	0.0248
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0025	**0.0031**	0.0025	0.0027	0.0007	0.0007	0.0025	0.0012	0.0011	0.0012	0.0027	0.0012	0.0011	0.0011
38	0.0095	0.0127	0.0139	0.0132	0.0040	0.0039	0.0103	0.0050	0.0055	0.0056	**0.0145**	0.0050	0.0055	0.0056
39	0.2817	0.3064	0.3002	0.3069	0.2912	0.3069	0.2820	0.4259	0.4268	0.4268	0.3076	0.4304	**0.4306**	**0.4306**
40	0.2416	**0.2776**	0.2574	0.2752	0.2574	0.2588	0.2416	0.2456	0.2407	0.2407	0.2752	0.2453	0.2442	0.2442
41	0.0003	0.0008	0.0008	0.0007	0.0003	0.0003	0.0007	**0.0324**	0.0276	0.0147	0.0007	0.0267	0.0254	0.0143
42	0.3214	0.6250	0.7000	0.6429	0.7000	0.6667	0.3500	**0.8333**	**0.8333**	**0.8333**	0.6429	**0.8333**	**0.8333**	**0.8333**
43	0.0056	0.0207	0.0110	0.0198	0.0068	0.0068	0.0009	0.0230	0.0225	0.0074	0.0073	**0.0234**	0.0226	0.0073
44	0.1765	0.1545	0.1597	0.1538	0.1725	0.1759	**0.1868**	0.1246	0.1353	0.1356	0.1585	0.1480	0.1403	0.1406
45	0.2867	0.2807	0.2746	0.2783	0.6952	0.6926	0.2908	0.7140	**0.7148**	**0.7148**	0.2828	0.6944	0.6931	0.6931
46	0.8095	0.7101	0.7222	0.7121	0.8333	0.8095	**0.9167**	0.3547	0.7333	0.7576	0.7500	0.5833	0.7143	0.7576
47	0.0309	0.0225	0.0299	0.0225	0.0469	**0.0537**	0.0310	0.0304	0.0313	0.0299	0.0230	0.0377	0.0391	0.0382
48	0.7935	0.7992	0.7729	0.7992	0.8878	0.8945	0.7862	0.9022	0.9024	0.8738	0.7989	0.9027	**0.9034**	0.8721
49	0.5625	**0.6111**	**0.6111**	**0.6111**	0.5625	0.5625	0.5000	**0.6111**	**0.6111**	0.5000	0.5000	0.6000	0.6000	0.5000
50	0.1153	0.2605	0.2602	0.2628	0.1447	0.1460	0.1203	0.2857	0.2855	**0.2978**	0.2722	0.2836	0.2838	0.2961
51	0.7512	0.7211	0.7384	0.7215	0.7624	0.7658	**0.7792**	0.7726	0.7707	0.7785	0.7559	0.7694	0.7673	**0.7792**
52	0.0057	0.0015	0.0035	0.0016	0.0053	0.0052	0.0073	0.0008	0.0009	0.0153	0.0068	0.0007	0.0008	**0.0156**
53	0.1265	0.2009	0.2075	0.2032	0.1254	0.1263	0.1247	0.2504	0.2506	0.2325	0.1895	0.2517	**0.2520**	0.2349
54	0.0051	0.0204	0.0100	0.0205	0.0010	0.0015	0.0052	0.0019	0.0019	0.0019	**0.0206**	0.0019	0.0019	0.0020
55	0.0958	0.1669	0.2193	0.1679	0.2131	0.2132	0.0956	0.2069	**0.2267**	0.2176	0.1724	0.2059	0.2194	0.2104
56	0.1303	0.2878	0.2424	**0.2903**	0.0874	0.0817	0.1303	0.2231	0.2358	0.2358	**0.2903**	0.2059	0.2121	0.2121
57	0.4510	0.4442	0.4354	0.4450	**0.4805**	0.4751	0.3670	0.4209	0.4266	0.3798	0.3702	0.4295	0.4221	0.3777
58	0.1211	0.1180	0.1187	0.1168	0.1304	0.1256	0.1336	0.1238	0.1231	**0.1342**	0.1217	0.1299	0.1153	0.1234
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7809	0.7802	0.7804	0.7813	0.7696	0.7731	0.7941	0.7798	0.7823	0.7434	**0.8017**	0.7835	0.7889	0.7503
61	0.0876	0.1079	0.1067	0.1064	0.0165	0.0159	0.1453	0.0347	0.0365	0.4854	0.2294	0.0362	0.0413	**0.4951**
62	0.0466	**0.0506**	0.0497	0.0486	0.0379	0.0394	0.0467	0.0435	0.0403	0.0405	0.0486	0.0442	0.0426	0.0427
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0209	0.1280	0.0757	0.1270	0.0287	0.0213	0.0326	0.0686	0.0742	0.0780	**0.1383**	0.0632	0.0674	0.0710
65	0.2591	0.4402	0.4272	0.4337	0.3790	0.3725	0.3160	0.4488	0.4515	0.4532	0.4593	0.4616	0.4647	**0.4651**
66	0.0341	0.1768	0.0889	**0.1776**	0.0064	0.0046	0.0523	0.0439	0.0440	0.1464	0.1398	0.0395	0.0396	0.1488
67	0.3997	0.3998	0.3875	0.3992	0.4212	**0.4237**	0.1892	0.4010	0.3975	0.1775	0.1886	0.3996	0.3949	0.1840
68	0.3733	0.7391	0.7076	0.7378	0.5681	0.5780	0.3733	0.7558	**0.7584**	**0.7584**	0.7378	0.7562	0.7583	0.7583
69	0.3220	0.2156	0.3362	0.2138	**0.3461**	0.3432	0.2556	0.2877	0.2895	0.2217	0.1976	0.2828	0.2921	0.2214
70	0.0194	0.0121	0.0113	0.0122	0.0055	0.0062	**0.0979**	0.0051	0.0044	0.0669	0.0775	0.0040	0.0035	0.0648
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.5307	0.6255	**0.6552**	0.6195	0.4261	0.4533	0.5307	0.6289	0.6206	0.6206	0.6195	0.6313	0.6350	0.6350
73	0.0133	0.0147	0.0155	0.0149	0.0136	0.0141	0.0187	0.0124	0.0126	0.0189	**0.0209**	0.0128	0.0131	0.0194
74	0.3414	0.4535	0.3657	0.4372	0.1158	0.3377	0.4068	0.4493	0.4494	**0.7255**	0.6389	0.4487	0.4489	**0.7255**
75	0.3792	0.3135	0.3590	0.3081	**0.6393**	0.6191	0.5068	0.6269	0.6287	0.6077	0.4289	0.6108	0.6012	0.5883
76	0.0000	**0.0145**	0.0075	0.0143	0.0000	0.0000	0.0000	0.0018	0.0018	0.0011	0.0065	0.0015	0.0015	0.0010
77	0.0243	**0.0254**	0.0227	0.0230	0.0004	0.0036	0.0213	0.0028	0.0028	0.0028	0.0198	0.0025	0.0024	0.0025
78	0.1207	0.1313	0.1400	0.1345	0.2411	0.2356	0.1282	**0.2654**	0.2644	0.2642	0.1432	0.2554	0.2559	0.2575
79	0.9583	0.9583	0.9583	0.9583	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**
80	0.0096	0.0517	0.0470	0.0518	0.1833	0.1742	0.0173	**0.6250**	**0.6250**	**0.6250**	0.0767	0.3333	0.3409	0.3409
81	0.0324	0.0399	0.0396	0.0381	0.0568	0.0570	0.0339	0.0629	0.0833	**0.0943**	0.0401	0.0645	0.0676	0.0761
82	0.0011	0.0009	0.0014	0.0009	0.0034	0.0028	**0.0152**	0.0024	0.0021	0.0039	0.0104	0.0019	0.0019	0.0049
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.4537	0.3286	0.2179	0.3286	0.4097	0.4300	**0.4611**	0.1420	0.1429	0.1531	0.3296	0.1519	0.1535	0.1596
85	0.3863	0.3607	0.3748	0.3580	**0.5073**	0.5010	0.3446	0.4774	0.4728	0.3965	0.3186	0.4480	0.4445	0.3798
86	0.1292	0.0653	0.0598	0.0640	0.0748	0.0804	0.1459	0.1734	0.2166	**0.2270**	0.0721	0.1654	0.1650	0.1740
87	0.0248	0.0407	0.0392	0.0405	0.0045	0.0049	0.0250	0.0079	0.0079	0.0079	**0.0408**	0.0086	0.0085	0.0089
88	0.0238	0.1587	0.1519	0.1559	0.0027	0.0026	0.1403	0.1108	0.1101	0.2534	**0.2596**	0.1122	0.1121	0.2489
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0013**	0.0000	0.0000	0.0000	0.0012
90	**0.1245**	0.1094	0.0825	0.1101	0.0977	0.1020	**0.1245**	0.0819	0.0831	0.0831	0.1101	0.0878	0.0904	0.0904
91	0.1667	0.1710	0.1352	0.1711	0.0625	0.0714	0.1717	0.0656	0.0656	0.0720	**0.1814**	0.0646	0.0646	0.0707
92	0.0430	0.0411	0.0455	0.0418	0.0070	0.0055	0.0449	0.0453	0.0069	0.0081	0.0457	**0.0499**	0.0078	0.0079
93	0.3071	0.8132	0.7954	0.8128	0.8483	0.8530	0.3499	0.9076	0.9068	0.8929	0.8249	**0.9078**	0.9076	0.8938
94	0.1167	0.2163	0.2022	0.2230	0.1521	0.1464	0.1270	0.1071	0.0999	0.1065	**0.2613**	0.1115	0.1139	0.1189
95	0.4656	0.4783	0.4806	0.4788	0.5941	**0.5966**	0.4543	0.5928	0.5932	0.5701	0.4616	0.5924	0.5936	0.5707
96	0.2636	0.2958	0.2878	0.3008	0.2621	0.2584	0.2725	0.2950	0.2960	0.2988	**0.3062**	0.2923	0.2931	0.2964
97	0.0465	0.0458	0.0476	0.0462	0.0465	0.0442	0.0472	**0.0539**	0.0450	0.0454	0.0465	0.0500	0.0442	0.0445
98	0.1094	0.1790	0.1521	0.1881	0.1596	0.1442	0.1118	0.3076	**0.3221**	0.3043	0.1845	0.3123	0.3198	0.3029
99	0.2226	0.2291	0.1912	0.2308	**0.3561**	0.3480	0.2227	0.2594	0.2425	0.2425	0.2308	0.2726	0.2578	0.2578
100	0.0150	0.0851	0.0572	0.0858	0.0551	0.0518	0.0144	**0.1093**	**0.1093**	0.1019	0.0817	0.1083	0.1081	0.1002
all	0.2390	0.2749	0.2624	0.2744	0.2714	0.2743	0.2481	0.2908	0.2943	**0.3062**	0.2826	0.2907	0.2916	0.3044

Table C.10: InL2 per query results (TDN. title,description and narrative tags used).

T	base	S	L	LS	Bo1	KL	G	SB1	LSB1	LSB1G	LSG	SKL	LSKL	LSKLG
01	0.6061	0.6345	0.6313	0.6034	0.5975	0.6176	0.6052	0.6455	0.6415	**0.6684**	0.6052	0.6508	0.6426	0.6645
02	0.0207	0.0499	0.0151	0.0306	0.0060	0.0063	0.0375	0.0049	0.0067	0.0145	**0.0573**	0.0087	0.0075	0.0159
03	0.2559	0.1780	0.2393	0.1732	0.3013	0.2898	0.2563	0.3038	0.2642	0.2648	0.1735	0.3111	0.3161	**0.3166**
04	0.3187	0.3156	**0.3222**	0.2939	0.3061	0.3073	0.2563	0.2778	0.2642	0.2113	0.2470	0.2805	0.2667	0.2105
05	0.6391	0.6707	0.6668	0.6735	0.7051	0.7091	0.6486	0.7042	0.7030	0.7114	0.6808	0.7036	0.7028	**0.7118**
06	0.3443	0.4218	0.3618	**0.4373**	0.1630	0.1827	0.3139	0.2716	0.2737	0.2909	0.4361	0.2713	0.2772	0.2924
07	**0.3436**	0.1232	0.1451	0.1240	0.3014	0.3279	0.2859	0.0334	0.0335	0.0337	0.1103	0.0449	0.0436	0.0419
08	0.0588	0.0439	0.0467	0.0418	0.0396	0.0387	**0.0913**	0.0377	0.0326	0.0481	0.0533	0.0384	0.0335	0.0510
09	0.3786	0.4014	0.2577	0.2803	0.4119	0.3943	0.4170	**0.4209**	0.3548	0.3866	0.3214	0.4087	0.3541	0.3889
10	0.7574	0.8496	0.7354	0.8551	0.7737	0.7994	0.8463	0.8076	0.8207	0.9177	0.9016	0.8114	0.8177	**0.9355**
11	0.0725	0.0706	0.0745	0.0709	0.1095	0.1091	0.0748	0.1627	0.1627	**0.1667**	0.0761	0.1623	0.1625	0.1650
12	0.0675	0.2399	0.2375	0.2399	0.2073	0.2229	0.0647	0.2089	0.2552	0.2723	0.2515	0.2192	0.2529	**0.2727**
13	0.4795	0.4603	0.4652	0.4669	0.4096	0.4718	**0.4871**	0.4287	0.4286	0.4329	0.4795	0.4562	0.4628	0.4678
14	0.2861	0.2624	0.3055	0.2712	0.5951	0.6051	0.2648	0.6202	0.6227	0.4369	0.3317	0.6202	**0.6250**	0.4390
15	0.6239	0.6672	0.6196	0.6648	0.7929	0.7949	0.6928	**0.8197**	0.8167	0.8182	0.7170	0.8164	0.8145	0.8155
16	0.8488	0.8987	0.8665	0.8987	0.8944	0.8827	0.8714	0.9252	0.9262	**0.9319**	0.9112	0.9171	0.9176	0.9277
17	0.4311	0.4068	**0.4529**	0.4070	0.4241	0.4246	0.4263	0.3875	0.3901	0.3858	0.4027	0.3912	0.3925	0.3875
18	0.2576	0.3240	0.2942	0.3208	0.1525	0.1974	0.2852	0.3503	0.3244	0.3349	0.3452	**0.3730**	0.3406	0.3589
19	0.1480	0.1396	0.1432	0.1536	**0.2353**	0.2263	0.1864	0.1722	0.1923	0.2280	0.1987	0.1627	0.1850	0.2231
20	0.1917	0.1830	**0.2593**	0.2077	0.0337	0.0327	0.0120	0.0310	0.0265	0.0022	0.0047	0.0386	0.0352	0.0021
21	0.4069	0.4988	0.5456	0.5283	0.4146	0.4121	0.4153	0.4319	0.3974	0.6288	0.5303	0.4411	0.4100	**0.6311**
22	0.3377	0.3165	0.2785	0.3268	0.3648	0.3769	0.3391	0.3650	0.3846	**0.4454**	0.3987	0.3608	0.3806	0.4450
23	0.0376	0.0493	0.0581	0.0560	0.0214	0.0219	0.0643	0.0495	0.0537	0.1780	0.1184	0.0481	0.0533	**0.1828**
24	0.4847	0.4786	0.4691	0.4787	0.5213	0.5235	0.4909	0.5577	0.5565	0.5599	0.4816	**0.5626**	0.5609	0.5624
25	0.3278	0.3167	0.3167	0.3167	**0.8333**	0.7778	0.3278	0.5909	0.4333	0.4333	0.3167	0.4242	0.4242	0.4242
26	0.0758	0.2303	0.2376	0.2312	0.1000	0.1100	0.0924	0.3181	0.3188	0.3436	0.2785	0.3308	0.3338	**0.3643**
27	0.0335	0.0229	0.0204	0.0192	0.0695	**0.0712**	0.0346	0.0334	0.0348	0.0351	0.0220	0.0344	0.0615	0.0618
28	0.0079	**0.0593**	0.0291	0.0570	0.0008	0.0007	0.0080	0.0083	0.0082	0.0082	0.0576	0.0063	0.0062	0.0063
29	0.1953	0.2492	**0.2623**	0.2432	0.1785	0.1723	0.1820	0.1924	0.1916	0.1851	0.2267	0.1909	0.1914	0.1869
30	0.1721	0.9306	0.3310	0.9306	0.3985	0.3225	0.3712	**1.0000**	**1.0000**	0.8345	0.8068	**1.0000**	**1.0000**	0.8346
31	0.3298	0.3309	0.3251	0.3363	0.3193	0.3128	0.3444	0.4406	0.4560	**0.4639**	0.3520	0.4190	0.4374	0.4452
32	0.8588	0.9449	0.8751	0.9413	0.9497	0.9477	0.8282	0.9669	0.9669	0.9365	0.9107	**0.9699**	**0.9699**	0.9395
33	0.0076	0.0062	0.0030	0.0060	0.0013	0.0026	**0.0192**	0.0024	0.0029	0.0100	0.0153	0.0024	0.0028	0.0100
34	0.4111	0.3426	0.2677	0.3427	0.4333	0.4889	**0.6692**	0.2123	0.2124	0.6678	0.6688	0.2110	0.2269	0.6678
35	0.0400	0.0352	0.0395	0.0359	0.0242	0.0248	**0.0431**	0.0216	0.0217	0.0234	0.0383	0.0226	0.0227	0.0245
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	0.0026	**0.0033**	0.0027	0.0029	0.0008	0.0007	0.0027	0.0012	0.0012	0.0012	0.0029	0.0012	0.0012	0.0012
38	0.0093	0.0120	0.0149	0.0137	0.0037	0.0038	0.0101	0.0044	0.0048	0.0049	**0.0152**	0.0044	0.0048	0.0049
39	0.2954	0.3165	0.3197	0.3169	0.3604	0.3564	0.2957	0.4498	0.4495	0.4495	0.3176	0.4436	**0.4547**	**0.4547**
40	0.2290	0.2626	0.2519	**0.2633**	0.2553	0.2582	0.2290	0.2409	0.2404	0.2404	**0.2633**	0.2400	0.2401	0.2401
41	0.0003	0.0007	0.0008	0.0007	0.0003	0.0000	0.0007	**0.0214**	0.0058	0.0027	0.0007	0.0203	0.0046	0.0027
42	0.3214	0.6111	0.7000	0.6111	0.6667	0.6667	0.3500	**0.8333**	**0.8333**	**0.8333**	0.6111	**0.8333**	**0.8333**	**0.8333**
43	0.0055	0.0215	0.0106	0.0208	0.0090	0.0082	0.0010	0.0232	0.0208	0.0072	0.0077	**0.0237**	0.0205	0.0071
44	0.1839	0.1455	0.1685	0.1580	0.1718	0.1807	**0.1947**	0.1371	0.1244	0.1248	0.1641	0.1458	0.1436	0.1440
45	0.2761	0.2694	0.2567	0.2737	**0.7611**	0.7502	0.2796	0.7564	0.7546	0.7546	0.2783	0.7514	0.7514	0.7514
46	0.8095	0.7037	0.7193	0.7083	0.8095	0.8095	**0.9167**	0.3547	0.5833	0.6000	0.7436	0.7436	0.7436	0.7576
47	0.0304	0.0214	0.0298	0.0214	0.0461	**0.0528**	0.0305	0.0386	0.0386	0.0387	0.0230	0.0458	0.0460	0.0466
48	0.7831	0.7900	0.7667	0.7903	0.8859	0.8928	0.7763	0.8997	**0.9001**	0.8717	0.7905	0.8995	0.8999	0.8693
49	0.5588	0.6000	0.5909	0.6000	0.5588	0.5556	0.5000	**0.6111**	**0.6111**	0.5000	0.5000	0.6000	0.6000	0.5000
50	0.1197	0.2757	0.2646	0.2690	0.1586	0.1640	0.1314	0.3054	0.3062	0.3197	0.2952	0.3049	0.3061	**0.3215**
51	0.7462	0.7129	0.7304	0.7122	0.7485	0.7590	0.7717	0.7695	0.7672	**0.7746**	0.7438	0.7639	0.7615	0.7733
52	0.0064	0.0017	0.0039	0.0018	0.0055	0.0055	0.0086	0.0010	0.0010	0.0119	0.0071	0.0008	0.0009	**0.0124**
53	0.1314	0.2079	0.2079	0.2102	0.1472	0.1478	0.1315	0.2570	0.2576	0.2392	0.1965	0.2550	**0.2578**	0.2403
54	0.0058	0.0212	0.0121	0.0213	0.0013	0.0008	0.0075	0.0017	0.0017	0.0017	**0.0217**	0.0017	0.0018	0.0018
55	0.0781	0.1461	0.2034	0.1459	0.2039	0.2064	0.0777	0.2021	**0.2250**	0.2161	0.1533	0.2032	0.2200	0.2121
56	0.1955	0.3197	0.2579	0.3028	0.1254	0.1159	0.1955	0.3436	**0.4325**	**0.4325**	0.3028	0.3338	0.4122	0.4122
57	0.4420	0.4422	0.4402	0.4445	**0.4803**	0.4616	0.3626	0.4395	0.4496	0.4148	0.3655	0.4339	0.4402	0.4034
58	0.1239	0.1138	0.1263	0.1099	0.1263	0.1187	0.1380	0.1240	0.1496	**0.1586**	0.1156	0.1555	0.1499	0.1583
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	0.7799	0.7777	0.7776	0.7787	0.7636	0.7664	0.7940	0.7758	0.7753	0.7374	**0.7960**	0.7827	0.7870	0.7483
61	0.0828	0.1015	0.1066	0.1031	0.0167	0.0162	0.1521	0.0325	0.0338	0.4043	0.2159	0.0347	0.0354	**0.4177**
62	0.0451	**0.0496**	0.0467	0.0478	0.0361	0.0376	0.0453	0.0416	0.0407	0.0408	0.0478	0.0429	0.0421	0.0423
63	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0187	0.1135	0.0604	0.1118	0.0203	0.0208	0.0291	0.0526	0.0530	0.0581	**0.1213**	0.0472	0.0481	0.0544
65	0.2590	0.4299	0.4174	0.4263	0.3306	0.3343	0.3135	0.3449	0.3883	0.3897	**0.4567**	0.3594	0.4001	0.4015
66	0.0674	0.3446	0.1177	**0.3464**	0.0079	0.0061	0.0530	0.0462	0.0464	0.2058	0.1447	0.0415	0.0417	0.2083
67	0.3907	0.3939	0.3856	0.3944	0.3822	0.3816	0.1874	**0.4017**	0.3992	0.1764	0.1862	0.4010	0.3986	0.1820
68	0.3692	0.7404	0.7048	0.7373	0.5598	0.5688	0.3692	0.7602	0.7602	0.7601	0.7373	**0.7610**	0.7608	0.7607
69	0.3198	0.2032	0.3437	0.2023	**0.3470**	**0.3470**	0.2731	0.2751	0.2806	0.2212	0.1944	0.2778	0.2736	0.2163
70	0.0219	0.0125	0.0123	0.0126	0.0059	0.0062	**0.0998**	0.0053	0.0053	0.0641	0.0766	0.0046	0.0046	0.0624
71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	0.5185	0.6021	0.6231	0.5899	0.5365	0.5481	0.5185	0.6597	**0.6675**	**0.6675**	0.5899	0.6563	0.6655	0.6655
73	0.0186	0.0179	0.0189	0.0181	0.0225	0.0240	0.0238	0.0145	0.0147	0.0214	**0.0242**	0.0149	0.0152	0.0217
74	0.3400	0.4549	0.3607	0.4550	0.1147	0.3368	0.4084	0.4352	0.4228	**0.7576**	**0.7576**	0.4343	0.4219	0.7500
75	0.3648	0.2713	0.3281	0.2683	**0.6173**	0.6028	0.4969	0.6036	0.6013	0.5879	0.4030	0.5718	0.5719	0.5647
76	0.0000	**0.0116**	0.0059	0.0115	0.0000	0.0000	0.0000	0.0015	0.0015	0.0009	0.0051	0.0006	0.0006	0.0008
77	0.0226	**0.0239**	0.0209	0.0217	0.0073	0.0082	0.0201	0.0030	0.0028	0.0031	0.0190	0.0032	0.0024	0.0026
78	0.1207	0.1636	0.1455	0.1667	0.2432	0.2375	0.1280	0.2737	**0.2755**	0.2753	0.1759	0.2650	0.2626	0.2644
79	0.9583	0.9583	0.9583	0.9583	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**	0.9583	**0.9762**	**0.9762**	**0.9762**
80	0.0084	0.0728	0.0512	0.0729	0.1484	0.1082	0.0157	0.5769	**0.5909**	**0.5909**	0.1045	0.5769	0.5769	0.5833
81	0.0332	0.0427	0.0420	0.0421	0.0664	0.0691	0.0350	0.0811	0.0804	**0.0926**	0.0448	0.0690	0.0685	0.0780
82	0.0008	0.0007	0.0010	0.0007	0.0011	0.0009	**0.0118**	0.0021	0.0021	0.0043	0.0088	0.0017	0.0017	0.0042
83	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	0.4710	0.2782	0.2139	0.3331	0.3414	0.3634	**0.4778**	0.1586	0.1586	0.1750	0.3338	0.1627	0.1634	0.1750
85	0.3868	0.3633	0.3770	0.3577	**0.5138**	0.5008	0.3427	0.4764	0.4663	0.3926	0.3160	0.4608	0.4508	0.3853
86	0.1281	0.0669	0.0571	0.0759	0.2290	0.2209	0.1450	0.1639	0.2216	**0.2328**	0.0819	0.1528	0.1652	0.1742
87	0.0275	0.0458	0.0426	0.0456	0.0042	0.0046	0.0277	0.0076	0.0076	0.0077	**0.0459**	0.0087	0.0087	0.0088
88	0.0212	0.1673	0.1544	0.1647	0.0028	0.0028	0.1391	0.1169	0.1129	0.2523	**0.2544**	0.1156	0.1138	0.2385
89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0012**	0.0000	0.0000	0.0000	0.0011
90	**0.1214**	0.1110	0.0832	0.1120	0.0969	0.0994	**0.1214**	0.0855	0.0863	0.0863	0.1120	0.0915	0.0928	0.0928
91	0.1667	0.1703	0.1347	0.1703	0.0556	0.0714	0.1710	0.0650	0.0581	0.0638	**0.1797**	0.0642	0.0574	0.0634
92	0.0434	0.0344	0.0373	0.0351	0.0057	0.0052	0.0455	0.0430	0.0436	0.0472	0.0382	0.0465	0.0472	**0.0515**
93	0.3173	0.8263	0.8083	0.8251	0.8472	0.8517	0.3671	0.9061	0.9049	0.8885	0.8338	0.9095	**0.9104**	0.8924
94	0.1178	0.2333	0.2287	0.2359	0.0549	0.0622	0.1285	0.1425	0.1335	0.1378	**0.2871**	0.1420	0.1407	0.1433
95	0.4679	0.4886	0.4850	0.4983	0.5953	0.5954	0.4568	0.5949	0.5948	0.5711	0.4815	0.5975	**0.6000**	0.5760
96	0.2715	0.3001	0.2864	0.3011	0.2751	0.2719	0.2803	0.2939	0.2938	0.2966	**0.3067**	0.2920	0.2908	0.2939
97	0.0456	0.0446	0.0461	0.0450	0.0476	0.0453	0.0462	**0.0539**	0.0472	0.0475	0.0452	0.0510	0.0454	0.0457
98	0.1076	0.1780	0.1419	**0.1842**	0.0356	0.0354	0.1117	0.0484	0.0488	0.0473	0.1841	0.0491	0.0498	0.0486
99	0.2070	0.1925	0.1604	0.1938	**0.3432**	0.3366	0.2071	0.2176	0.2189	0.2189	0.1938	0.2362	0.2376	0.2376
100	0.0151	0.0929	0.0677	0.0940	0.0537	0.0524	0.0147	0.1140	**0.1172**	0.1101	0.0909	0.1135	0.1137	0.1071
all	0.2387	0.2753	0.2613	0.2750	0.2732	0.2764	0.2478	0.2947	0.2967	0.3092	0.2830	0.2973	0.2987	**0.3116**

APPENDIX D

Annex D: GeoQA Questions

This annex shows the test set of the GeoQA experiments. This set contain 62 questions about the Spanish geography in Spanish.

1 ¿A qué comunidad autónoma pertenece el Puigcampana?

To what autonomous community does the Puigcampana belongs?

2 ¿A qué comunidad pertenece El Ferrol?

To what autonomous community does El Ferrol belongs?

3 ¿A qué comunidad pertenece la isla La Gomera?

To what community belongs the island La Gomera?

4 ¿A qué mar desemboca la ría de Betanzos?

To what sea leads the ria of Betanzos?

5 ¿Cuál es el sistema de la comunidad autónoma Canaria?

What is the mountain range of tha Canary autonomous community?

6 ¿Cuál es el capital de Andalucía?

What is the capital of Andalusia?

7 ¿Cuál es el nombre de la comunidad autónoma en la que se encuentra Cullera?

What is the name of the autonomous community in which Cullera is located?

8 ¿Cuál es la capital Navarra?

What is the capital of Navarre?

9 ¿Cuál es la capital de las islas Las Canarias?

What is the capital of the Canary Islands?

10 ¿Cuál es la comunidad en la que desemboca el Guadalentín?

What is the community in which Guadalentín ends?

11 ¿Cuál es la extensión de la comunidad Madrileña?

What is the extension of the Madrid community?

12 ¿Cuál es la extensión de la comunidad autónoma donde está el golfo de Vizcaya?

What is the extension of the autonomous community in which the Bay of Biscay is located?

13 ¿Cuál es la extensión de la comunidad de Castilla y León?

What is the extension of the community of Castilla y León?

14 ¿Cuántos habitantes tiene la comunidad autónoma de Castilla?

How many inhabitants has the autonomous community of Castile?

15 ¿Cómo se llama la capital de la comunidad autónoma de La Rioja?

What is the capital of the autonomous community of La Rioja called?

16 ¿Cómo se nombra el río que pasa por Granada?

How is the river that passes through Granada named?

17 Dime a qué sistema pertenece el pico Teide?

Tell me which system belongs the peak Teide?

18 Dime a qué comunidad pertenece el cabo de La Nao?

Tell me which community is the Cape of La Nao?

19 Dime a qué comunidad pertenece la ría de Vigo?

Tell me which community is the Vigo estuary?

20 Dime el mar en que desemboca el Llobregat?

Tell me the sea where the Llobregat flows?

21 Dime el mar que baña las islas Canarias?

Tell me the sea that bathes the Canary Islands?

22 Dime en qué sistema nace el río Aragón?

Tell me in what system is the river Aragón born?

23 Dime en qué comunidad autónoma se encuentra Manacor?

Tell me in what autonomous community is Manacor?

24 Dime en qué comunidad autónoma se encuentra la ciudad de Barbastro?

Tell me in what autonomous community is the city of Barbastro?

25 Dime en qué comunidad desemboca el Llobregat?

Tell me in what community does the Llobregat ends?

26 Dime en qué mar está la isla de Conejera?

Tell me in what sea is the island of Conejera?

27 Dime la población de la comunidad autónoma de Murcia?

Tell me the population of the autonomous community of Murcia?

28 Dime qué extensión tiene la isla de Hierro?

Tell me the extent of the island of Hierro?

29 ¿Dónde está la isla de Gran Canaria?

Where is the island of Gran Canaria?

30 ¿Dónde está la ría Ribadeo?

Where is the Ribadeo estuary?

31 ¿En qué archipiélago se encuentra Mallorca?

In what archipelago is Mallorca located?

32 ¿En qué ciudad desemboca el río Segura?

In what city does the Segura River ends?

33 ¿En qué comunidad autónoma está el Cantábrico?

In what autonomous community is the Cantabrian Sea located?

34 ¿En qué comunidad autónoma está el Mulhacén?

In what autonomous community is the Mulhacen located?

35 ¿En qué comunidad autónoma está el cabo Tarifa?

In what autonomous community is the Tarifa Cape located?

36 ¿En qué comunidad autónoma está situada la Sierra de Gũdar?

In what autonomous community is the Sierra of Gúdar located?

37 ¿En qué comunidad autónoma están los Picos de Europa?

In what autonomous community are the Picos the Europa located?

38 ¿En qué comunidad autónoma se encuentra la isla de La Gomera?

In what autonomous community is the island of La Gomera located?

39 ¿En qué comunidad autónoma se encuentra la Sierra del Maestrazgo?

In what autonomous community is the Sierra of Maestrazgo located?

40 ¿En qué comunidad está la sierra de Somosierra?

In what autonomous community is the sierra of Somosierra located?

41 ¿En qué comunidad nace el río Guadarrama?

In what autonomous community is the Guadarrama river located?

42 ¿En qué comunidad se encuentra el cabo San Adrián?

In what autonomous community is the San Adrián Cape located?

43 ¿En qué comunidad se encuentran los Pirineos?

In what autonomous community are the Pyrenees located?

44 ¿En qué mar está situado el golfo de Cádiz?

In what sea is the Gulf of Cádiz the located?

45 ¿En qué mar se encuentra la ría de Camariñas?

In what sea is the Camariñas estuary?

46 La comunidad en la que nace el río Guadalbullón?

The community in which the river Guadalbullón is born?

47 Me gustaría saber la extensión de la comunidad Vasca?

I would like to know the extension of the Basque community?

48 Nombre de la capital de Andalucía?

Name of the capital of Andalusia?

49 Nombre de la capital de la comunidad autónoma de Andalucía?

Name of the capital of the Autonomous Community of Andalusia?

50 Nombre de la comunidad donde nace el río Eresma?

Name of the community where the river Eresma is born?

51 Podría decirme el nũmero de habitantes de Figueras?

Can you tell me the number of inhabitants of Figueras?

52 Quiero que me digas la capital de la comunidad autónoma de Canarias?

I want you to tell me the capital of the autonomous community of the Canary Islands?

53 Quisiera saber el mar en donde está situada La Gomera?

I would like to know the sea where La Gomera is located?
54 ¿Qué capital tiene Castilla?
What capital does Castilla have?
55 ¿Qué extensión tiene La Gomera?
What is La Gomera extension?
56 ¿Qué extensión tiene la comunidad autónoma Asturiana?
What is the extension of the Asturian Autonomous Community?
57 ¿Qué mar baña el golfo de Onteniente?
What sea bathes the Gulf of Onteniente?
58 ¿Qué mar baña la comunidad autónoma Murciana?
What sea bathes the Murcian autonomous community?
59 ¿Qué mar es el que baña a la comunidad de Murcia?
What sea bathes the Murcian community?
60 ¿Qué nũmero de habitantes tiene Castilla la Mancha? *What is the number of inhabitants of Castilla la Mancha?*
61 ¿Qué nũmero de habitantes tiene Astorga?
What is the number of inhabitants of Astorga?
62 ¿Qué río pasa por Salamanca?
Which river passes through Salamanca?

APPENDIX E

Annex E: GikiCLEF Questions

This annex shows the GikiCLEF topics in English and Spanish used in the GikiCLEF 2009 evaluation benchmark.

```
<topic id="GC-2009-01">List the Italian places where Ernest Hemingway visited during his life.</topic>
<topic id="GC-2009-02"> Which countries have the white, green and red colors in their
national flag? </topic>
<topic id="GC-2009-03"> In which countries outside Bulgaria are there published opinions on
 Petar Dunov's (Beinsa Duno's) ideas? </topic>
<topic id="GC-2009-04"> Name Romanian poets who published volumes with ballads until 1941. </topic>
<topic id="GC-2009-05">Which written fictional works of non-Romanian authors have as subject
 the Carpathians mountains? </topic>
<topic id="GC-2009-06"> Which Dutch violinists held the post of concertmaster at the Royal
Concertgebouw Orchestra in the twentieth century? </topic>
<topic id="GC-2009-07"> What capitals of Dutch provinces received their town privileges
 before the fourteenth century? </topic>
<topic id="GC-2009-08"> Which authors were born in and write about the Bohemian Forest? </topic>
<topic id="GC-2009-09"> Name places where Goethe fell in love. </topic>
<topic id="GC-2009-10"> Which Flemish towns hosted a restaurant with two or
 three Michelin stars in 2008? </topic>
<topic id="GC-2009-11"> What Belgians won the Ronde van Vlaanderen exactly twice? </topic>
<topic id="GC-2009-12">Present monarchies in Europe headed by a woman.</topic>
<topic id="GC-2009-13"> Romantic and realist European novelists of the XIXth century who died
 of tuberculosis. </topic>
<topic id="GC-2009-14"> Name rare diseases with dedicated research centers in Europe. </topic>
<topic id="GC-2009-15"> List the basic elements of the cassata. </topic>
<topic id="GC-2009-16"> In which European countries is the bidet commonly used? </topic>
<topic id="GC-2009-17"> List the 5 Italian regions with a special statute.</topic>
<topic id="GC-2009-18"> In which Tuscan provinces is Chianti produced?</topic>
<topic id="GC-2009-19"> Name mountains in Chile with permanent snow. </topic>
<topic id="GC-2009-20"> List the name of the sections of the North-Western Alps. </topic>
<topic id="GC-2009-21"> List the left side tributaries of the Po river. </topic>
<topic id="GC-2009-22"> Which South American national football teams use the yellow color? </topic>
<topic id="GC-2009-23"> Name American museums which have any Picasso painting. </topic>
<topic id="GC-2009-24">Which countries have won a futsal European championship played in Spain?</topic>
<topic id="GC-2009-25"> Name Spanish drivers who have driven in Minardi. </topic>
<topic id="GC-2009-26"> Which Bulgarian fighters were awarded the "Diamond belt"? </topic>
<topic id="GC-2009-27"> Which Dutch bands are named after a Bulgarian footballer? </topic>
<topic id="GC-2009-28"> Find coastal states with Petrobras refineries.</topic>
<topic id="GC-2009-29"> Places above the Arctic circle with a population larger
 than 100,000 people </topic>
<topic id="GC-2009-30"> Which Japanese automakers companies have manufacturing or assembling
 factories in Europe? </topic>
<topic id="GC-2009-31"> Which countries have Italian as an official language? </topic>
<topic id="GC-2009-32"> Name Romanian writers who were living in USA in 2003. </topic>
<topic id="GC-2009-33"> What European Union countries have national parks in the Alps? </topic>
<topic id="GC-2009-34"> What eight-thousanders are at least partially in Nepal? </topic>
<topic id="GC-2009-35"> Which Romanian mountains are declared biosphere reserves? </topic>
<topic id="GC-2009-36"> Name Romanian caves where Paleolithic human fossil remains were found. </topic>
<topic id="GC-2009-37"> Which Norwegian musicians were convicted for burning churches? </topic>
<topic id="GC-2009-38"> Which Norwegian waterfalls are higher than 200m? </topic>
<topic id="GC-2009-39">National team football players from Scandinavia with sons who have played
 for English clubs. </topic>
<topic id="GC-2009-40">Which rivers in North Rhine Westphalia are approximately 10km long?</topic>
<topic id="GC-2009-41"> Chefs born in Austria who have received a Michelin Star. </topic>
<topic id="GC-2009-42"> Political parties in the National Council of Austria which were founded after
 the end of World War II </topic>
<topic id="GC-2009-43"> Austrian ski resorts with a total ski trail length of
 at least 100 km </topic>
<topic id="GC-2009-44"> Find Austrian grape varieties with a vineyard area below 100 ha. </topic>
<topic id="GC-2009-45"> Find Swiss casting show winners. </topic>
<topic id="GC-2009-46"> German writers who are Honorary Citizens in Switzerland. </topic>
<topic id="GC-2009-47"> Which cities in Germany have more than one university? </topic>
<topic id="GC-2009-48"> Which German-speaking movies have been nominated for an Oscar? </topic>
<topic id="GC-2009-49"> Formula One drivers who moved to Switzerland. </topic>
<topic id="GC-2009-50"> Which Swiss people were Olympic medalists in snowboarding at the Winter
 Olympic Games in 2006? </topic>
```

Figure E.1: List of GikiCLEF 2009 Questions in English.

```
<topic id=”GC-2009-01”>Nombre los lugares de Italia que haya visitado Ernest Hemingway
a lo largo de su vida.</topic>
<topic id=”GC-2009-02”> ¿Qué países tienen los colores blanco, verde y rojo en su bandera
nacional?</topic>
<topic id=”GC-2009-03”> ¿En qué países fuera de Bulgaria se han publicado opiniones sobre las ideas
de Petar Dunov?</topic>
<topic id=”GC-2009-04”> Nombre poetas rumanos que hayan publicado volúmenes de romances
antes de 1941.</topic>
<topic id=”GC-2009-05”> ¿Qué obras literarias de escritores no rumanos tienen por tema
los Cárpatos?</topic>
<topic id=”GC-2009-06”> ¿Qué violinistas holandeses ocuparon el puesto de concertino en
la Orquesta Real del Concertgebouw durante el siglo XX?</topic>
<topic id=”GC-2009-07”> ¿Qué capitales de provincias holandesas recibieron sus derechos de ciudad
antes del siglo XIV?</topic>
<topic id=”GC-2009-08”> ¿Qué autores que nacieron en el bosque de Bohemia han escrito
sobre él?</topic>
<topic id=”GC-2009-09”> Nombre lugares donde Goethe se haya enamorado.</topic>
<topic id=”GC-2009-10”> ¿Qué ciudades flamencas tenían en 2008 un restaurante con dos o
tres estrellas Michelín?</topic>
<topic id=”GC-2009-11”> ¿Qué belgas han ganado el Tour de Flandes exactamente dos veces?</topic>
<topic id=”GC-2009-12”>Monarquías europeas cuyo jefe de estado sea actualmente una mujer.</topic>
<topic id=”GC-2009-13”> Novelistas europeos románticos y realistas del siglo XIX que murieran
de tuberculosis.</topic>
<topic id=”GC-2009-14”> Nombre enfermedades raras que tengan dedicadas centros de investigación
en Europa.</topic>
<topic id=”GC-2009-15”> Liste los ingredientes principales de la Cassata.</topic>
<topic id=”GC-2009-16”> ¿En qué países europeos se suele usar el bidé?</topic>
<topic id=”GC-2009-17”> Nombre las cinco regiones italianas que tengan un estatuto especial.</topic>
<topic id=”GC-2009-18”> ¿En qué provincias de la Toscana se produce el Chianti?</topic>
<topic id=”GC-2009-19”> Nombre montañas chilenas que tengan nieves perpetuas.</topic>
<topic id=”GC-2009-20”> Liste los nombres de las secciones noroccidentales de los Alpes.</topic>
<topic id=”GC-2009-21”> Nombre los afluentes de la margen izquierda del río Po.</topic>
<topic id=”GC-2009-22”> ¿Qué selecciones sudamericanas de fútbol visten de color amarillo? </topic>
<topic id=”GC-2009-23”> Nombre museos americanos que tengan alguna obra de Picasso. </topic>
<topic id=”GC-2009-24”>¿Qué países han ganado un campeonato de Europa de fútbol sala
celebrado en España?</topic>
<topic id=”GC-2009-25”> Nombre pilotos españoles que hayan corrido en Minardi.</topic>
<topic id=”GC-2009-26”> ¿Qué luchadores búlgaros han sido galardonados con el cinturón
de diamantes?</topic>
<topic id=”GC-2009-27”> ¿Qué bandas holandesas deben su nombre al de un futbolista búlgaro?</topic>
<topic id=”GC-2009-28”> Nombre estados costeros que tengan refinerías de Petrobras.</topic>
<topic id=”GC-2009-29”> Poblaciones sobre el círculo polar ártico con más de 100.000 habitantes.</topic>
<topic id=”GC-2009-30”> ¿Qué fabricantes japoneses de automóviles tienen cadenas de producción o
de montaje en Europa?</topic>
<topic id=”GC-2009-31”> ¿Qué países tienen el italiano como lengua oficial?</topic>
<topic id=”GC-2009-32”> Nombre escritores rumanos que estuvieran viviendo en 2003
en Estados Unidos.</topic>
<topic id=”GC-2009-33”> ¿Qué países de la Unión Europea tienen parques nacionales
en los Alpes?</topic>
<topic id=”GC-2009-34”> ¿Qué ochomiles pertenecen al menos parcialmente a Nepal?</topic>
<topic id=”GC-2009-35”> ¿Qué montañas rumanas están declaradas reserva de la biosfera?</topic>
<topic id=”GC-2009-36”> Nombre cuevas rumanas donde se hayan encontrado restos fósiles humanos
del Paleolítico.</topic>
<topic id=”GC-2009-37”> ¿Qué músicos noruegos fueron encarcelados por quemar iglesias?</topic>
<topic id=”GC-2009-38”> ¿Qué cataratas noruegas tienen una altura superior a 200 metros?</topic>
<topic id=”GC-2009-39”> Jugadores de fútbol de equipos nacionales escandinavos que tengan hijos que
hayan jugado en equipos ingleses.</topic>
<topic id=”GC-2009-40”> ¿Qué ríos de Renania del Norte-Westfalia tienen aproximadamente
10 km de longitud?</topic>
<topic id=”GC-2009-41”> Cocineros nacidos en Austria que hayan recibido una estrella Michelin.</topic>
<topic id=”GC-2009-42”> Partidos políticos del consejo nacional austriaco que hayan sido fundados
después del fin de la segunda guerra mundial.</topic>
<topic id=”GC-2009-43”> Estaciones de esquí austriacas cuya longitud total de pistas sea de
al menos 100 km.</topic>
<topic id=”GC-2009-44”> Encuentre variedades de uva austriaca cuya área de viñedos sea menor
a 100 hectáreas.</topic>
<topic id=”GC-2009-45”> Encuentre ganadores de programas suizos de casting.</topic>
<topic id=”GC-2009-46”> Escritores alemanes que sean ciudadanos honorarios de Suiza.</topic>
<topic id=”GC-2009-47”> ¿Qué ciudades alemanas tienen más de una universidad?</topic>
<topic id=”GC-2009-48”> ¿Qué películas de habla alemana han recibido una nominación
a los óscar?</topic>
<topic id=”GC-2009-49”> Pilotos de Fórmula 1 que se hayan mudado a Suiza.</topic>
<topic id=”GC-2009-50”> ¿ Qué suizos ganaron en snowboard medallas olímpicas durante
los Juegos Olímpicos de Invierno de 2006?</topic>
```

Figure E.2: List of GikiCLEF 2009 Questions in Spanish.

APPENDIX F

Annex F: Geographical Data Mappings

This annex describes shows different geographical data mappings developed in this thesis.

F.1 Mapping of GNS and GNIS gazetteers to the ADL Feature Type Thesaurus)

This section contains the mapping of GeoNet Names Server (GNS) and GNIS gazetteers feature types to Alexandria Digital Library Feature Type Thesaurus (ADLFTT). The feature type thesaurus of the Geographical KB for the GIR task is the *ADL Feature Type Thesaurus (ADLFTT)*. The *ADL Feature Type Thesaurus* is a hierarchical set of geographical terms used to type named geographic places in English (Hill, 2000). Both *GNIS* and *GNS* gazetteers have been mapped to the *ADLFTT*, with a resulting set of 575 geographical types. Our *GNIS* mapping is similar to the one exposed in Hill (2000).

GNS or GNIS	ADLFTT
ADM1	administrative__areas@@political__areas@@countries__1st__order__divisions
ADM2	administrative__areas@@political__areas@@countries__2nd__order__divisions
ADM3	administrative__areas@@political__areas@@countries__3rd__order__divisions
ADM4	administrative__areas@@political__areas@@countries__4th__order__divisions
ADMD	administrative__areas@@administrative__divisions
ADMF	manmade__features@@buildings@@administrative__facility
AGRC	administrative__areas@@populated__places
AGRF	manmade__features@@agricultural__sites
AIRB	manmade__features@@transportation__features@@airport__features@@airbases
AIRF	manmade__features@@transportation__features@@airport__features@@airfields
AIRH	manmade__features@@transportation__features@@airport__features@@heliports
AIRP	manmade__features@@transportation__features@@airport__features@@airports
airport	manmade__features@@transportation__features@@airport__features@@airports
AIRQ	manmade__features@@transportation__features@@airport__features@@airfields@@abandoned__airfield
AIRS	manmade__features@@transportation__features@@airport__features@@seaplane__bases@@seaplane__landing__areas
AMTH	manmade__features@@recreational__facilities@@performance__sites@@amphitheatres
ANCH	manmade__features@@hydrographic__structures@@harbors@@anchorages
ANS	manmade__features@@historical__sites@@archaeological__sites@@ancient__sites

APNU	physiographic__features@@aprons__
arch	physiographic__features@@natural__rock__formations@@arches__
ARCH	physiographic__features@@natural__rock__formations@@arches__
ARCU	physiographic__features@@natural__rock__formations@@arches__
AREA	others
area	physiographic__features@@areas
arroyo	physiographic__features@@arroyos__
ARRU	physiographic__features@@seafloor__features
ASPH	physiographic__features@@basins@@asphalt__lakes
ASTR	administrative__areas@@reference__locations@@astronomical__stations
ASYL	manmade__features@@buildings@@institutional__sites@@medical__facilities@@asylums
ATHF	manmade__features@@recreational__facilities@@sports__facilities@@athletic__fields
ATOL	regions@@land__regions@@islands@@atolls
bar	physiographic__features@@bars__
BAR	physiographic__features@@bars
basin	physiographic__features@@basins
bay	hydrographic__features@@bays
BAY	hydrographic__features@@bays
BAYS	hydrographic__features@@bays
BCH	physiographic__features@@beaches
BCHS	physiographic__features@@beaches
BCN	administrative__areas@@reference__locations@@beacons
BDG	manmade__features@@transportation__features@@bridges
BDGQ	manmade__features@@transportation__features@@bridges@@ruined__bridge
BDLD	physiographic__features@@badlands
BDLU	physiographic__features@@seafloor__features@@continental__margins@@borderlands__
beach	physiographic__features@@beaches
bench	physiographic__features@@seafloor__features@@benches__
bend	hydrographic__features@@streams@@rivers@@bends__
BGHT	physiographic__features@@bights
BKSU	physiographic__features@@seafloor__features@@banks__
BLDG	manmade__features@@buildings
BLDR	physiographic__features@@boulder__fields
BLHL	physiographic__features@@blowholes
BLOW	physiographic__features@@blowouts
BNCH	physiographic__features@@seafloor__features@@benches__
BNCU	physiographic__features@@seafloor__features@@benches__
BNK	physiographic__features@@seafloor__features@@banks__
BNKR	physiographic__features@@seafloor__features@@banks__@@stream__banks
BNKU	physiographic__features@@seafloor__features@@banks__
BNKX	physiographic__features@@seafloor__features@@banks__@@section__of__bank
BOG	regions@@biogeographic__regions@@wetlands@@bogs
BP	administrative__areas@@reference__locations@@boundary__markers
bridge	manmade__features@@transportation__features@@bridges
BRKS	manmade__features@@buildings@@residential__sites@@barracks
BRKW	manmade__features@@hydrographic__structures@@breakwaters
BSND	hydrographic__features@@drainage__basins
BSNP	physiographic__features@@mineral__deposit__areas@@petroleum__basins
BSNU	physiographic__features@@basins
BSTN	manmade__features@@buildings@@baling__station
BTL	manmade__features@@historical__sites@@battlefields
BTYD	manmade__features@@hydrographic__structures@@harbors
building	manmade__features@@buildings
BUR	manmade__features@@historical__sites@@archaeological__sites@@burial__caves
BUSH	regions@@biogeographic__regions@@shrublands@@bushes
BUTE	physiographic__features@@mesas@@buttes
canal	manmade__features@@hydrographic__structures@@canals
cape	physiographic__features@@capes
CAPE	physiographic__features@@capes
CAPG	hydrographic__features@@ice__masses@@icecaps
CARN	manmade__features@@landmarks@@cairns
cave	physiographic__features@@caves
CAVE	physiographic__features@@caves
CDAU	physiographic__features@@mountains@@cordilleras
cemetery	manmade__features@@cemeteries
CFT	physiographic__features@@cliffs@@clefts
channel	hydrographic__features@@channels
CH	manmade__features@@buildings@@institutional__sites@@religious__facilities@@churches
CHN	hydrographic__features@@channels
CHNL	hydrographic__features@@channels@@lake__channels
CHNM	hydrographic__features@@channels@@marine__channels
CHNN	hydrographic__features@@channels@@navigation__channels
church	manmade__features@@buildings@@institutional__sites@@religious__facilities@@churches
civil	administrative__areas@@political__areas
CLDA	physiographic__features@@craters
CLF	physiographic__features@@cliffs
CLG	regions@@biogeographic__regions@@grasslands
cliff	physiographic__features@@cliffs
CMN	administrative__areas@@parks@@commons
CMPLA	manmade__features@@buildings@@institutional__sites@@correctional__facilities@@labor__camps
CMPL	manmade__features@@recreational__facilities@@camps@@logging__camps
CMP	manmade__features@@recreational__facilities@@camps
CMPMN	manmade__features@@mine__sites@@mining__camps
CMPO	manmade__features@@recreational__facilities@@camps@@oil__camps
CMPQ	manmade__features@@recreational__facilities@@camps@@abandoned__camp
CMPRF	manmade__features@@recreational__facilities@@camps@@refugee__camps
CMTY	manmade__features@@cemeteries
CNFL	hydrographic__features@@streams@@confluences
CNLA	manmade__features@@transportation__features@@aqueducts
CNLB	manmade__features@@hydrographic__structures@@canals@@canal__bends
CNLD	manmade__features@@hydrographic__structures@@canals@@drainage__canals
CNLI	manmade__features@@hydrographic__structures@@canals@@irrigation__canals
CNL	manmade__features@@hydrographic__structures@@canals
CNLN	manmade__features@@hydrographic__structures@@canals@@navigation__canals
CNLQ	manmade__features@@hydrographic__structures@@canals@@abandoned__canal
CNLSB	manmade__features@@hydrographic__structures@@canals@@underground__irrigation__canals
CNLX	manmade__features@@hydrographic__structures@@canals@@section__of__canal
CNS	administrative__areas@@concession__areas
CNSU	physiographic__features@@valleys@@canyons
CNYN	physiographic__features@@valleys@@canyons

CNYU	physiographic__features@@valleys@@canyons
COLF	physiographic__features@@mineral__deposit__areas@@coalfields
COMC	manmade__features@@telecommunication__features@@communication__centers
CONE	physiographic__features@@cones__
COVE	hydrographic__features@@bays@@coves
crater	physiographic__features@@craters
CRDR	manmade__features@@transportation__features@@corridors
CRKT	hydrographic__features@@streams@@tidal__creeks
CRNT	hydrographic__features@@seas@@oceans@@ocean__currents@@currents
crossing	manmade__features@@transportation__features@@crossings
CRQ	physiographic__features@@cirques
CRQS	physiographic__features@@cirques
CRRL	manmade__features@@agricultural__sites@@corrals
CRSU	physiographic__features@@seafloor__features@@continental__margins@@continental__rises
CRTR	physiographic__features@@craters
CSNO	manmade__features@@recreational__facilities
CSTL	manmade__features@@fortifications@@castles
CSTM	manmade__features@@buildings@@customs__houses
CST	regions@@coastal__zones@@coasts
CSWY	manmade__features@@transportation__features@@roadways@@causeways
CSWYQ	manmade__features@@transportation__features@@roadways@@causeways@@former__causeway
CTHSE	manmade__features@@buildings@@court__houses
CTRA	manmade__features@@buildings@@research__facilities@@atomic__centers
CTRB	manmade__features@@buildings@@commercial__sites@@business__centers
CTRCM	manmade__features@@buildings@@community__centers
CTRF	manmade__features@@buildings@@facility__centers
CTRM	manmade__features@@buildings@@institutional__sites@@medical__facilities@@medical__centers
CTRR	manmade__features@@buildings@@institutional__sites@@religious__facilities@@religious__centers
CTRS	manmade__features@@buildings@@research__facilities@@space__centers
CUET	physiographic__features@@mountains@@ridges@@cuestas
CULT	manmade__features@@agricultural__sites@@cultivated__areas
CUTF	hydrographic__features@@channels@@cutoffs__
CVNT	manmade__features@@buildings@@institutional__sites@@religious__facilities@@convents
dam	manmade__features@@hydrographic__structures@@dam__sites@@dams
DAM	manmade__features@@hydrographic__structures@@dam__sites@@dams
DAMQ	manmade__features@@hydrographic__structures@@dam__sites@@dams@@ruined__dam
DAMSB	manmade__features@@hydrographic__structures@@dam__sites@@sub__surface__dams
DARY	manmade__features@@agricultural__sites@@dairies
DCKB	manmade__features@@hydrographic__structures@@harbors@@docking__basins
DCKD	manmade__features@@hydrographic__structures@@harbors@@dry__docks
DCK	manmade__features@@hydrographic__structures@@piers@@docks
DCKY	manmade__features@@hydrographic__structures@@harbors@@dockyards
DEPU	physiographic__features@@seafloor__features@@ocean__trenches@@deeps__
DEVH	manmade__features@@buildings@@residential__sites@@housing__areas@@housing__developments
DIKE	manmade__features@@hydrographic__structures@@levees@@dikes___
DLTA	hydrographic__features@@deltas
DOMG	hydrographic__features@@ice__masses@@icecap__domes
DPOF	manmade__features@@transportation__features@@fuel__depots
DPRG	hydrographic__features@@ice__masses@@icecap__depressions
DPR	physiographic__features@@basins@@depressions
DSRT	regions@@biogeographic__regions@@deserts
DTCHD	manmade__features@@hydrographic__structures@@canals@@drainage__ditches
DTCHI	manmade__features@@hydrographic__structures@@canals@@irrigation__ditches
DTCH	manmade__features@@hydrographic__structures@@canals@@ditches
DTCHM	manmade__features@@hydrographic__structures@@canals@@ditch__mouths
DUNE	physiographic__features@@dunes
DVD	others@@divide
EDGU	physiographic__features@@seafloor__features@@continental__margins@@shelf__edges__
ERG	regions@@biogeographic__regions@@deserts@@sandy__deserts
ESCU	physiographic__features@@cliffs
ESTB	manmade__features@@agricultural__sites@@banana__plantations
ESTC	manmade__features@@agricultural__sites@@cotton__plantations
EST	land__parcels@@estates
ESTO	manmade__features@@agricultural__sites@@oil__palm__plantations
ESTR	manmade__features@@agricultural__sites@@rubber__plantations
ESTSG	manmade__features@@agricultural__sites@@sugar__plantations
ESTSL	manmade__features@@agricultural__sites@@sisal__plantations
ESTT	manmade__features@@agricultural__sites@@tea__plantations
ESTX	land__parcels@@estates@@section__of__estate
ESTY	hydrographic__features@@estuaries
falls	hydrographic__features@@streams@@rivers@@waterfalls@@falls
FAN	physiographic__features@@alluvial__fans@@fans__
FANU	physiographic__features@@alluvial__fans@@fans__
FCL	manmade__features@@buildings
FISH	manmade__features@@recreational__facilities@@fishing__areas
FJD	hydrographic__features@@bays@@fjords
FJDS	hydrographic__features@@bays@@fjords
flat	physiographic__features@@flats
FLDI	manmade__features@@agricultural__sites@@irrigated__fields
FLD	manmade__features@@recreational__facilities@@sports__facilities@@fields
FLLS	hydrographic__features@@streams@@rivers@@waterfalls
FLLSX	hydrographic__features@@streams@@rivers@@waterfalls@@section__of__waterfall(s)
FLTM	regions@@biogeographic__regions@@wetlands@@mud__flats
FLTT	regions@@biogeographic__regions@@wetlands@@tidal__flats
FLTU	physiographic__features@@flats
FNDY	manmade__features@@buildings
FORD	manmade__features@@transportation__features@@fords__
forest	regions@@biogeographic__regions@@forests
FRKU	physiographic__features@@valleys@@forks__
FRM	manmade__features@@agricultural__sites@@farms
FRMQ	manmade__features@@agricultural__sites@@farms@@abandoned__farm
FRMS	manmade__features@@agricultural__sites@@farms
FRMT	manmade__features@@agricultural__sites@@farmsteads
FRSTF	regions@@biogeographic__regions@@forests@@petrified__forests@@fossilized__forests
FRST	regions@@biogeographic__regions@@forests
FRSU	physiographic__features@@valleys@@forks__
FRZU	physiographic__features@@tectonic__features@@faults@@fracture__zones
FSR	physiographic__features@@fissures
FT	administrative__areas@@military__areas@@forts
FURU	physiographic__features@@furrows

FY	manmade__features@@transportation__features@@ferries
gap	physiographic__features@@gaps
GAP	physiographic__features@@gaps
GAPU	physiographic__features@@gaps
GASF	manmade__features@@oil__fields@@gasfields
GATE	manmade__features@@transportation__features@@gates__
GDN	administrative__areas@@parks@@gardens
geyser	hydrographic__features@@thermal__features@@geysers
GHSE	manmade__features@@buildings@@residential__sites@@housing__areas@@guest__houses
glacier	hydrographic__features@@ice__masses@@glacier__features@@glaciers
GLCR	hydrographic__features@@ice__masses@@glacier__features@@glaciers
GLYU	physiographic__features@@arroyos@@gullies
GOSP	manmade__features@@buildings@@gas__oil__separator__plant
GRAZ	regions@@biogeographic__regions@@grasslands@@grazing__areas
GRGE	physiographic__features@@valleys@@canyons@@gorges
GRSLD	regions@@biogeographic__regions@@grasslands
GRVC	regions@@biogeographic__regions@@forests@@woods@@coconut__groves
GRVE	manmade__features@@cemeteries@@graves
GRVO	regions@@biogeographic__regions@@forests@@woods@@olive__groves
GRVPN	regions@@biogeographic__regions@@forests@@woods@@pine__groves
GRVP	regions@@biogeographic__regions@@forests@@woods@@palm__groves
GULF	hydrographic__features@@gulfs
gut	hydrographic__features@@guts
GVL	others@@gravel__area
GYSR	hydrographic__features@@thermal__features@@geysers
harbor	manmade__features@@hydrographic__structures@@harbors
HBR	manmade__features@@hydrographic__structures@@harbors
HBRX	manmade__features@@hydrographic__structures@@harbors@@section__of__harbor
HDLD	physiographic__features@@capes@@headlands
HERM	manmade__features@@buildings@@residential__sites@@housing__areas@@hermitages
HLL	physiographic__features@@mountains@@hills
HLLS	physiographic__features@@mountains@@hills
HLLU	physiographic__features@@mountains@@hills
HLSU	physiographic__features@@mountains@@hills
HLT	manmade__features@@transportation__features@@halting__places__
HMCK	physiographic__features@@hammocks
HMDA	regions@@biogeographic__regions@@deserts@@rock__deserts
HOLU	physiographic__features@@seafloor__features@@holes__
hospital	manmade__features@@buildings@@institutional__sites@@medical__facilities@@hospitals
HSEC	manmade__features@@buildings@@residential__sites@@housing__areas@@country__houses
HSE	manmade__features@@buildings@@residential__sites@@housing__areas@@houses
HSPC	manmade__features@@buildings@@institutional__sites@@medical__facilities@@clinics
HSPD	manmade__features@@buildings@@institutional__sites@@medical__facilities@@dispensaries__
HSPL	manmade__features@@buildings@@institutional__sites@@medical__facilities@@leprosarium
HSP	manmade__features@@buildings@@institutional__sites@@medical__facilities@@hospitals
HSTS	manmade__features@@historical__sites
HTH	regions@@biogeographic__regions@@shrublands@@heaths
HTL	manmade__features@@buildings@@hotels
HUT	manmade__features@@buildings@@residential__sites@@housing__areas@@huts
HUTS	manmade__features@@buildings@@residential__sites@@housing__areas@@huts
INDS	manmade__features@@buildings@@commercial__sites@@industrial__sites@@industrial__areas
INLT	hydrographic__features@@channels@@inlets
INLTQ	hydrographic__features@@channels@@inlets@@former__inlet
INSM	administrative__areas@@military__areas@@military__installations
INTF	physiographic__features@@plains
island	regions@@land__regions@@islands
ISLF	manmade__features@@hydrographic__structures@@offshore__platforms@@artificial__islands
ISLM	regions@@land__regions@@islands@@mangrove__islands
ISL	regions@@land__regions@@islands
ISLS	regions@@land__regions@@islands
ISLT	regions@@land__regions@@islands@@land__tied__islands
ISLX	regions@@land__regions@@islands@@section__of__island
isthmus	physiographic__features@@isthmuses
ISTH	physiographic__features@@isthmuses
ITTR	manmade__features@@buildings@@research__facilities@@research__institutes
JTY	manmade__features@@hydrographic__structures@@piers@@jetties
KNLU	physiographic__features@@mountains@@knolls
KNSU	physiographic__features@@mountains@@knolls
KRST	physiographic__features@@karst__areas
lake	hydrographic__features@@lakes
LAND	regions@@biogeographic__regions@@tundras@@Arctic__land
lava	physiographic__features@@volcanic__features@@lava__fields
LAVA	physiographic__features@@volcanic__features@@lava__fields@@lava__areas
LBED	hydrographic__features@@lakes@@lake__beds
LCTY	administrative__areas@@populated__places@@localities
LDGU	physiographic__features@@ledges
LDNG	hydrographic__structures@@boat__landings
LEPC	manmade__features@@buildings@@residential__sites@@leper__colonies
levee	manmade__features@@hydrographic__structures@@levees
LEV	manmade__features@@hydrographic__structures@@levees
LEVU	manmade__features@@hydrographic__structures@@levees
LGN	hydrographic__features@@lakes@@lagoons
LGNS	hydrographic__features@@lakes@@lagoons
LGNX	hydrographic__features@@lakes@@lagoons@@section__of__lagoon
LKC	physiographic__features@@volcanic__features@@crater__lakes
LK	hydrographic__features@@lakes
LKI	hydrographic__features@@lakes@@intermittent__lakes
LKN	hydrographic__features@@lakes@@salt__lakes
LKNI	hydrographic__features@@lakes@@intermittent__salt__lakes
LKO	hydrographic__features@@lakes@@oxbow__lakes
LKOI	hydrographic__features@@lakes@@intermittent__oxbow__lakes
LKSB	physiographic__features@@caves@@underground__lakes
LKSC	physiographic__features@@volcanic__features@@crater__lakes
LKS	hydrographic__features@@lakes
LKSI	hydrographic__features@@lakes@@intermittent__lakes
LKSN	hydrographic__features@@lakes@@salt__lakes
LKSNI	hydrographic__features@@lakes@@intermittent__salt__lakes
LKX	hydrographic__features@@lakes@@section__of__lake
locale	administrative__areas@@populated__places@@locales
LOCK	manmade__features@@transportation__features@@locks

LTER	administrative__areas@@leased__areas__
LTHSE	manmade__features@@reference__locations@@lighthouse
MAR	manmade__features@@hydrographic__structures@@harbors@@marinas
MDVU	physiographic__features@@valleys@@median__valleys
MDW	regions@@biogeographic__regions@@grasslands@@meadows
MESA	physiographic__features@@mesas
MESU	physiographic__features@@mesas
MFGB	manmade__features@@buildings@@commercial__sites@@breweries
MFGC	manmade__features@@buildings@@commercial__sites@@industrial__sites@@canneries
MFGCU	manmade__features@@buildings@@commercial__sites@@industrial__sites@@copper__works
MFGLM	manmade__features@@buildings@@commercial__sites@@industrial__sites@@limekilns
MFG	manmade__features@@buildings@@commercial__sites@@industrial__sites@@factories
MFGM	manmade__features@@buildings@@commercial__sites@@industrial__sites@@munitions__plants
MFGN	hydrographic__features@@lakes@@salt__evaporation__ponds
MFGPH	manmade__features@@buildings@@commercial__sites@@industrial__sites@@phosphate__works
MFGQ	manmade__features@@buildings@@commercial__sites@@industrial__sites@@factories@@abandoned__factory
MFGSG	manmade__features@@buildings@@commercial__sites@@industrial__sites@@sugar__refineries
MGV	regions@@biogeographic__regions@@wetlands@@mangrove__swamps
MILB	administrative__areas@@military__areas@@military__bases
military	administrative__areas@@military__areas
mine	manmade__features@@mine__sites@@mines
MKT	manmade__features@@buildings@@commercial__sites@@markets
ML	manmade__features@@buildings@@commercial__sites@@industrial__sites@@mills
MLM	manmade__features@@buildings@@commercial__sites@@industrial__sites@@ore__treatment__plants
MLO	manmade__features@@buildings@@commercial__sites@@industrial__sites@@olive__oil__mills
MLSG	manmade__features@@buildings@@commercial__sites@@industrial__sites@@sugar__mills
MLSGQ	manmade__features@@buildings@@commercial__sites@@industrial__sites@@sugar__mills@@former__sugar__mill
MLSW	manmade__features@@buildings@@commercial__sites@@industrial__sites@@sawmills
MLWND	manmade__features@@windmills
MLWTR	manmade__features@@hydrographic__structures@@water__mills
MNA	manmade__features@@mine__sites@@mining__areas
MNAU	manmade__features@@mine__sites@@gold__mines
MNC	manmade__features@@mine__sites@@coal__mines
MNCR	manmade__features@@mine__sites@@chrome__mines
MNCU	manmade__features@@mine__sites@@copper__mines
MND	physiographic__features@@mountains@@mounds
MNDT	manmade__features@@mine__sites@@diatomite__mines
MNDU	physiographic__features@@mountains@@mounds
MNFE	manmade__features@@mine__sites@@iron__mines
MN	manmade__features@@mine__sites@@mines
MNMT	manmade__features@@monuments
MNNI	manmade__features@@mine__sites@@nickel__mines
MNN	manmade__features@@mine__sites@@salt__mines
MNPB	manmade__features@@mine__sites@@lead__mines
MNPL	manmade__features@@mine__sites@@placer__mines
MNQ	manmade__features@@mine__sites@@mines@@abandoned__mine
MNQR	anmade__features@@mine__sites@@mines@@quarries
MNSN	manmade__features@@mine__sites@@tin__mines
MOLE	manmade__features@@hydrographic__structures@@breakwaters@@moles__
MOOR	regions@@biogeographic__regions@@shrublands@@moors
MOTU	physiographic__features@@seafloor__features@@moats__
MRN	physiographic__features@@moraines
MRSHN	regions@@biogeographic__regions@@wetlands@@salt__marshes
MRSH	regions@@biogeographic__regions@@wetlands@@marshes
MSQE	manmade__features@@buildings@@institutional__sites@@religious__facilities@@mosques
MSSN	manmade__features@@buildings@@missions
MSSNQ	manmade__features@@buildings@@missions@@abandoned__mission
MSTY	manmade__features@@buildings@@residential__sites@@monasteries
MT	physiographic__features@@mountains
MTS	physiographic__features@@mountains
MTSU	physiographic__features@@mountains
MTU	physiographic__features@@mountains
MUS	manmade__features@@buildings@@museum__buildings@@museums
MVA	administrative__areas@@military__areas@@maneuver__areas
NKM	physiographic__features@@meander__necks
NOV	manmade__features@@buildings@@residential__sites@@novitiates
NRWS	hydrographic__features@@channels@@narrows__
NSY	manmade__features@@buildings@@institutional__sites@@medical__facilities@@nurseries
NTK	hydrographic__features@@ice__masses@@glacier__features@@nunataks
NTKS	physiographic__features@@mountains@@mountain__summits@@nunataks
NVB	administrative__areas@@military__areas@@naval__bases
OAS	others@@oases
OBPT	administrative__areas@@parks@@observation__points
OBS	manmade__features@@buildings@@research__facilities@@observatories
OBSR	manmade__features@@buildings@@research__facilities@@radio__observatories
OCH	manmade__features@@agricultural__sites@@orchards
OCN	hydrographic__features@@seas@@oceans
oilfield	manmade__features@@oil__fields@@oilfields
OILF	manmade__features@@oil__fields@@oilfields
OILJ	manmade__features@@transportation__features@@pipelines@@oil__pipeline__junctions
OILP	manmade__features@@transportation__features@@pipelines@@oil__pipelines
OILQ	manmade__features@@wells@@oil__wells@@abandoned__oil__well
OILR	manmade__features@@buildings@@commercial__sites@@industrial__sites@@oil__refineries
OILT	manmade__features@@buildings@@commercial__sites@@industrial__sites@@tank__farms
OILW	manmade__features@@wells@@oil__wells
other	others
OVF	hydrographic__features@@overfalls
PAL	manmade__features@@buildings@@residential__sites@@palaces
PAN	physiographic__features@@basins@@pans__
PANS	physiographic__features@@basins@@pans__
park	administrative__areas@@parks
PASS	physiographic__features@@gaps@@passes
PCL	administrative__areas@@political__areas@@political__entities
PCLD	administrative__areas@@political__areas@@dependent__political__entities
PCLF	administrative__areas@@political__areas@@freely__associated__states
PCLI	administrative__areas@@political__areas@@independent__political__entities
PCLIX	administrative__areas@@political__areas@@independent__political__entities@@section__of__independent__political__entity
PCLS	administrative__areas@@political__areas@@semi__independent__political__entities
PEAT	regions@@biogeographic__regions@@wetlands@@peat__cutting__areas
PEN	physiographic__features@@capes@@peninsulas

PENX	physiographic_features@@capes@@peninsulas@@section_of_peninsula
PGDA	manmade_features@@towers@@pagodas
PIER	manmade_features@@hydrographic_structures@@piers
pillar	physiographic_features@@natural_rock_formations@@pillars_
PKLT	manmade_features@@transportation_features@@parking_sites@@parking_lots
PK	physiographic_features@@mountains@@mountain_summits@@peaks
PKS	physiographic_features@@mountains@@mountain_summits@@peaks
PKSU	physiographic_features@@mountains@@mountain_summits@@peaks
PKU	physiographic_features@@mountains@@mountain_summits@@peaks
plain	physiographic_features@@plains
PLAT	physiographic_features@@@@plateaus
PLATX	physiographic_features@@plateaus@@section_of_plateau
PLDR	physiographic_features@@polders
PLFU	manmade_features@@hydrographic_structures@@offshore_platforms@@platforms_
PLN	physiographic_features@@plains
PLNU	physiographic_features@@plains
PLNX	physiographic_features@@plains@@section_of_plain
PLTU	physiographic_features@@plateaus
PMPO	manmade_features@@transportation_features@@pipelines@@oil_pumping_stations
PMPW	manmade_features@@transportation_features@@pipelines@@water_pumping_stations
PND	hydrographic_features@@lakes@@ponds
PNDI	hydrographic_features@@lakes@@intermittent_ponds
PNDN	hydrographic_features@@lakes@@salt_ponds
PNDNI	hydrographic_features@@lakes@@intermittent_salt_ponds
PNDSF	hydrographic_features@@lakes@@fishponds
PNDS	hydrographic_features@@lakes@@ponds
PNDSI	hydrographic_features@@lakes@@intermittent_ponds
PNDSN	hydrographic_features@@lakes@@salt_ponds
PNLU	physiographic_features@@natural_rock_formations@@pinnacles___
po	manmade_features@@buildings@@post_office_buildings
PO	manmade_features@@buildings@@post_office_buildings@@post_offices
POOLI	hydrographic_features@@lakes@@intermittent_pools
POOL	manmade_features@@recreational_facilities@@sports_facilities@@pools
Populated	Place_administrative_areas@@populated_places
Post	Office___manmade_features@@buildings@@post_office_buildings
PPLA	administrative_areas@@populated_places@@cities@@capitals
ppl	administrative_areas@@populated_places
PPL	administrative_areas@@populated_places
PPLC	administrative_areas@@populated_places@@cities@@capitals
PPLL	administrative_areas@@populated_places@@populated_localities
PPLQ	administrative_areas@@populated_places@@abandoned_populated_place
PPLR	manmade_features@@buildings@@institutional_sites@@religious_facilities@@religious_populated_places
PPLS	administrative_areas@@populated_places
PPLW	administrative_areas@@populated_places@@destroyed_populated_place
PPLX	administrative_areas@@populated_places@@section_of_populated_place
PP	manmade_features@@buildings@@police_posts
PPQ	manmade_features@@buildings@@police_posts@@abandoned_police_post
PRK	administrative_areas@@parks
PRKGT	administrative_areas@@parks@@park_gates
PRKHQ	manmade_features@@buildings@@park_headquarters
PRMN	manmade_features@@transportation_features@@trails@@promenades
PRNJ	manmade_features@@buildings@@institutional_sites@@correctional_facilities@@reformatories
PRN	manmade_features@@buildings@@institutional_sites@@correctional_facilities@@prisons
PRNQ	manmade_features@@buildings@@institutional_sites@@correctional_facilities@@prisons@@abandoned_prison
PROM	physiographic_features@@capes
PRSH	administrative_areas@@political_areas@@countries_2nd_order_divisions
PRT	manmade_features@@hydrographic_structures@@harbors@@ports
PRVU	administrative_areas@@political_areas@@countries_1st_order_divisions
PSH	manmade_features@@buildings@@commercial_sites@@industrial_sites@@power_generation_sites@@hydroelectric_power_stations
PS	manmade_features@@buildings@@commercial_sites@@industrial_sites@@power_generation_sites@@power_stations
PSTB	manmade_features@@buildings@@border_posts
PSTC	manmade_features@@buildings@@customs_posts
PSTP	manmade_features@@buildings@@patrol_posts
PTGE	manmade_features@@transportation_features@@portages
PT	physiographic_features@@capes@@points_
PTS	physiographic_features@@capes@@points_
PYR	manmade_features@@monuments@@pyramids
PYRS	manmade_features@@monuments@@pyramids
QCKS	physiographic_features@@bars
QUAY	manmade_features@@hydrographic_structures@@piers@@quays
range	physiographic_features@@mountains@@mountain_ranges@@ranges_
rapids	hydrographic_features@@streams@@rivers@@rapids
RAVU	physiographic_features@@valleys@@canyons@@ravines
RCH	hydrographic_features@@channels@@reaches_
RDA	manmade_features@@transportation_features@@roadways@@ancient_roads
RDB	manmade_features@@transportation_features@@roadways@@road_bends
RDCUT	manmade_features@@transportation_features@@roadways@@road_cuts
RDGB	physiographic_features@@mountains@@ridges@@beach_ridges
RDGE	physiographic_features@@mountains@@ridges
RDGG	physiographic_features@@mountains@@ridges@@icecap_ridges
RDGU	physiographic_features@@mountains@@ridges
RDJCT	manmade_features@@transportation_features@@roadways@@road_junctions
RD	manmade_features@@transportation_features@@roadways@@roads
RDST	manmade_features@@hydrographic_structures@@harbors@@roadsteads_
RDSU	physiographic_features@@mountains@@ridges
RECG	manmade_features@@recreational_facilities@@sports_facilities@@golf_courses
RECR	manmade_features@@recreational_facilities@@sports_facilities@@racetracks
REG	regions@@biogeographic_regions@@deserts@@stony_deserts
REP	administrative_areas@@political_areas@@countries@@republics
RESA	administrative_areas@@reserves@@agricultural_reserves
RES	administrative_areas@@reserves
reserve	administrative_areas@@reserves
reservoir	manmade_features@@hydrographic_structures@@reservoirs
RESF	administrative_areas@@reserves@@forest_reserves
RESH	administrative_areas@@reserves@@hunting_reserves
RESN	administrative_areas@@reserves@@nature_reserves
RESP	administrative_areas@@reserves@@palm_tree_reserves
RESV	administrative_areas@@tribal_areas@@reservations_
RESW	administrative_areas@@reserves@@wildlife_reserves
RFC	physiographic_features@@reefs@@coral_reefs

RF	physiographic__features@@reefs
RFSU	physiographic__features@@reefs
RFU	physiographic__features@@reefs
RFX	physiographic__features@@reefs@@section__of__reef
RGNE	regions@@economic__regions
RGNL	regions@@land__regions@@lake__regions
RGN	regions
RHSE	manmade__features@@transportation__features@@resthouses
ridge	physiographic__features@@mountains@@ridges
RISU	physiographic__features@@seafloor__features@@rises__
RJCT	manmade__features@@transportation__features@@railroad__features@@railroad__junctions
RKFL	physiographic__features@@rockfall
RK	physiographic__features@@natural__rock__formations@@rocks
RKRY	regions@@biogeographic__regions@@habitats@@rookeries
RKS	physiographic__features@@natural__rock__formations@@rocks
RLG	manmade__features@@buildings@@institutional__sites@@religious__facilities@@religious__sites
RLGR	manmade__features@@buildings@@institutional__sites@@religious__facilities@@retreats__
RMPU	physiographic__features@@seafloor__features@@ramps__
RNCH	manmade__features@@agricultural__sites@@ranches
RNGA	administrative__areas@@military__areas@@artillery__ranges
RNGU	physiographic__features@@mountains@@mountain__ranges@@ranges__
RPDS	hydrographic__features@@streams@@rivers@@rapids
RR	manmade__features@@transportation__features@@railroad__features
RRQ	manmade__features@@transportation__features@@railroad__features@@abandoned__railroad
RSD	manmade__features@@transportation__features@@railroad__features@@railroad__siding
RSGNL	manmade__features@@transportation__features@@railroad__features@@railroad__signal
RSRT	manmade__features@@recreational__facilities@@resorts
RSTN	manmade__features@@transportation__features@@railroad__features@@railroad__stations
RSTNQ	manmade__features@@transportation__features@@railroad__features@@railroad__stations@@abandoned__railroad__station
RSTP	manmade__features@@transportation__features@@railroad__features@@railroad__stops
RSTPQ	manmade__features@@transportation__features@@railroad__features@@railroad__stops@@abandoned__railroad__stop
RSVI	manmade__features@@hydrographic__structures@@reservoirs@@intermittent__reservoirs
RSV	manmade__features@@hydrographic__structures@@reservoirs
RSVT	manmade__features@@storage__structures@@water__tanks
RTE	manmade__features@@transportation__features@@caravan__routes
RUIN	manmade__features@@historical__sites@@ruins
RVN	physiographic__features@@valleys@@canyons@@ravines
RYD	manmade__features@@transportation__features@@railroad__features@@railroad__yards
SALT	physiographic__features@@mineral__deposit__areas@@salt__areas
SAND	physiographic__features@@bars@@sand__area
SBED	hydrographic__features@@streams@@dry__stream__bed
SBKH	physiographic__features@@flats@@sabkhas
SCHA	manmade__features@@buildings@@institutional__sites@@educational__facilities@@agricultural__schools
SCHC	manmade__features@@buildings@@institutional__sites@@educational__facilities@@colleges
SCH	manmade__features@@buildings@@institutional__sites@@educational__facilities@@schools
SCHM	manmade__features@@buildings@@institutional__sites@@educational__facilities@@military__schools
SCHN	manmade__features@@buildings@@institutional__sites@@educational__facilities@@schools
school	manmade__features@@buildings@@institutional__sites@@educational__facilities@@schools
SCNU	physiographic__features@@seafloor__features@@seachannels__
SCRB	regions@@biogeographic__regions@@shrublands@@scrublands
SCRP	physiographic__features@@cliffs@@escarpments
SCSU	physiographic__features@@seafloor__features@@seachannels__
SD	hydrographic__features@@channels@@sounds__
SDL	physiographic__features@@gaps@@saddles__
SDLU	physiographic__features@@gaps@@saddles__
sea	hydrographic__features@@seas
SEA	hydrographic__features@@seas
SHFU	physiographic__features@@seafloor__features@@continental__margins@@shelves
SHLU	physiographic__features@@bars@@shoals
SHOL	physiographic__features@@bars@@shoals
SHOR	regions@@coastal__zones@@shores
SHPF	manmade__features@@agricultural__sites@@sheepfolds
SHRN	manmade__features@@monuments@@shrines
SHSE	manmade__features@@storage__structures@@storehouse
SHSU	physiographic__features@@bars@@shoals
SHVU	physiographic__features@@valleys@@shelf__valleys__
SILL	physiographic__features@@gaps@@sills__
SILU	physiographic__features@@gaps@@sills__
SINK	physiographic__features@@sinkholes
SLCE	manmade__features@@hydrographic__structures@@sluices
SLID	physiographic__features@@slides__
slope	physiographic__features@@slopes
SLP	physiographic__features@@slopes
SLPU	physiographic__features@@slopes
SMSU	physiographic__features@@seafloor__features@@seamounts
SMU	physiographic__features@@seafloor__features@@seamounts
SNOW	regions@@biogeographic__regions@@snow__regions@@snowfields
SNTR	manmade__features@@buildings@@institutional__sites@@medical__facilities@@sanitariums
SPA	manmade__features@@recreational__facilities@@spas
SPIT	physiographic__features@@bars@@spits
SPLY	manmade__features@@hydrographic__structures@@dam__sites@@spillways
SPNG	hydrographic__features@@streams@@springs__
SPNS	hydrographic__features@@streams@@springs@@sulphur__springs
SPNT	hydrographic__features@@thermal__features@@hot__springs
spring	hydrographic__features@@streams@@springs__
SPRU	physiographic__features@@mountains@@ridges@@spurs__
SPUR	physiographic__features@@mountains@@ridges@@spurs__
SQR	administrative__areas@@parks@@squares
STBL	manmade__features@@recreational__facilities@@stables
STDM	manmade__features@@recreational__facilities@@stadium
STKR	manmade__features@@transportation__features@@stock__routes
STLMT	administrative__areas@@populated__places@@Israeli__settlement
STMA	hydrographic__features@@streams@@anabranch
ST	manmade__features@@transportation__features@@roadways@@streets
STMB	hydrographic__features@@streams@@stream__bends
STMC	manmade__features@@hydrographic__structures@@canals@@canalized__streams
STMD	hydrographic__features@@streamsdistributaries
STMH	hydrographic__features@@drainage__basins@@headwaters
STM	hydrographic__features@@streams
STMI	hydrographic__features@@streams@@intermittent__streams

STMIX	hydrographic__features@@streams@@intermittent__streams@@section__of__intermittent__stream
STMM	hydrographic__features@@streams@@stream__mouths
STMQ	hydrographic__features@@streams@@abandoned__watercourse
STMSB	hydrographic__features@@streams@@lost__rivers
STMS	hydrographic__features@@streams
STMX	hydrographic__features@@streams@@section__of__stream
STNB	manmade__features@@buildings@@research__facilities@@scientific__research__bases
STNC	manmade__features@@buildings@@coast__guard__stations
STNE	manmade__features@@buildings@@research__facilities@@experiment__stations
STNF	manmade__features@@buildings@@forest__stations
STNI	manmade__features@@buildings@@inspection__stations
STNM	manmade__features@@buildings@@research__facilities@@data__collection__facilities@@meteorological__stations
STNR	manmade__features@@telecommunication__features@@radio__stations
STNS	manmade__features@@buildings@@research__facilities@@data__collection__facilities@@satellite__stations
STNW	manmade__features@@buildings@@commercial__sites@@industrial__sites@@whaling__stations
STPS	manmade__features@@transportation__features@@trails@@steps__
stream	hydrographic__features@@streams
STRT	hydrographic__features@@channels@@straits
summit	physiographic__features@@mountains@@mountain__summits@@summits
swamp	regions@@biogeographic__regions@@wetlands@@swamps
SWMP	regions@@biogeographic__regions@@wetlands@@swamps
SYSI	manmade__features@@hydrographic__structures@@canals@@irrigation__systems
TAL	physiographic__features@@talus__slopes
TERR	administrative__areas@@political__areas@@countries__1st__order__divisions@@territory
TERU	physiographic__features@@terraces__
TMB	manmade__features@@monuments@@tombs
TMPL	manmade__features@@buildings@@institutional__sites@@religious__facilities@@temples
TMSU	physiographic__features@@seafloor__features@@seamounts@@tablemounts
TMTU	physiographic__features@@seafloor__features@@seamounts@@tablemounts
TNGU	physiographic__features@@seafloor__features@@tongues__
TNKD	manmade__features@@agricultural__sites@@cattle__dipping__tanks
TNLC	manmade__features@@transportation__features@@tunnels@@canal__tunnels
TNL	manmade__features@@transportation__features@@tunnels
TNLN	physiographic__features@@caves@@natural__tunnels
TNLRD	manmade__features@@transportation__features@@tunnels@@road__tunnels
TNLRR	manmade__features@@transportation__features@@tunnels@@railroad__tunnels
TNLS	manmade__features@@transportation__features@@tunnels
tower	manmade__features@@towers
TOWR	manmade__features@@towers
trail	manmade__features@@transportation__features@@trails
TRB	administrative__areas@@tribal__areas
TREE	regions@@biogeographic__regions@@forests@@woods@@trees
TRGD	physiographic__features@@dunes@@interdune__troughs
TRGU	physiographic__features@@seafloor__features@@ocean__trenches@@troughs__
TRIG	administrative__areas@@reference__locations@@triangulation__stations
TRL	manmade__features@@transportation__features@@trails
TRMO	manmade__features@@transportation__features@@pipelines@@oil__pipeline__terminals
TRNU	physiographic__features@@seafloor__features@@ocean__trenches@@trenches__
TRR	physiographic__features@@terraces__
TUND	regions@@biogeographic__regions@@tundras
tunnel	manmade__features@@transportation__features@@tunnels
UPLD	physiographic__features@@uplands
USGE	manmade__features@@buildings@@United__States__Government__Establishment
VALG	physiographic__features@@valleys@@hanging__valleys
valley	physiographic__features@@valleys
VAL	physiographic__features@@valleys
VALS	physiographic__features@@valleys
VALU	physiographic__features@@valleys
VALX	physiographic__features@@valleys@@section__of__valley
VETF	manmade__features@@buildings@@veterinary__facility
VIN	manmade__features@@agricultural__sites@@vineyards
VINS	manmade__features@@agricultural__sites@@vineyards
VLC	physiographic__features@@volcanic__features@@volcanoes
VLSU	physiographic__features@@valleys
WADB	physiographic__features@@arroyos@@wadi__bends
WADJ	physiographic__features@@arroyos@@wadi__junctions
WADM	physiographic__features@@arroyos@@wadi__mouths
WAD	physiographic__features@@arroyos@@wadis
WADS	physiographic__features@@arroyos@@wadis
WADX	physiographic__features@@arroyos@@wadis@@section__of__wadi
WALLA	manmade__features@@buildings@@walls@@ancient__wall
WALL	manmade__features@@buildings@@walls
WEIR	manmade__features@@hydrographic__structures@@dam__sites@@weirs
well	manmade__features@@wells
WHRF	manmade__features@@hydrographic__structures@@piers@@wharves
WHRL	hydrographic__features@@whirlpools
WLL	manmade__features@@wells
WLLQ	manmade__features@@wells@@abandoned__well
WLLS	manmade__features@@wells
woods	regions@@biogeographic__regions@@forests@@woods
WRCK	manmade__features@@historical__sites@@wrecks
WTLDI	regions@@biogeographic__regions@@wetlands@@intermittent__wetlands
WTLD	regions@@biogeographic__regions@@wetlands
WTRC	hydrographic__features@@streams@@watercourses
WTRH	hydrographic__features@@waterholes
WTRW	manmade__features@@hydrographic__structures@@waterworks
ZNB	administrative__areas@@buffer__zones
ZNF	administrative__areas@@free__trade__zones
ZNL	administrative__areas@@leased__zones__
ZN	others@@zone
ZOO	administrative__areas@@parks@@zoos
ZZZZZ	others@@master__source__holdings__list

Table F.1: Manual Mapping from GNS and GNIS to the ADLFTT feature types..

APPENDIX G

Annex G: Web Person Search experiments at WePS-3

This annex presents a set of experiments performed by this thesis author at the Third Web People Search Workshop (WePS-3) task for clustering web people search documents in English (Ferrés and Rodríguez, 2010). The WePS-3 workshop was an evaluation Task under the scope of TebleCLEF (Artiles et al., 2010). Its aim was to evaluate systems which cluster and extract information from web people searches in English. According to Andogah (2010) PhD thesis Very Important People (VIPs) (e.g., political leaders) mentioned in a document can be used to determine the geographical coverage of a document. For this reason technologies related with Person Name Recognition and Disambiguation can be potentially useful for Geographical Information Access tasks.

The approaches presented are Lingo, Hierachical Agglomerative Clustering (HAC), and a 2-step HAC algorithm. The best results were obtained with with HAC and 2-step HAC algorithms. This annex describes the experiments with the WePS-3 development and test data, results and initial conclusions in the context of the WePS-3 Task 1 and the comparison with the official results of all the participants.

G.1 Development and Test Data at WePS-3

The development data used for WePS-3 is based on the test data of WePS-2 Clustering Task (Artiles et al., 2009). Test data for WePS-2 is composed of 30 ambiguous names: 10 name sets from the 1990 US Census, 10 from participants in ACL'08 and 10 from Wikipedia. Each name is made of two tokens, a first name and a last name. See more details of the WePS-2 data set in Artiles et al. (2009). Around 100 documents have been downloaded from the top ranked search results.

The test data for WePS-3 was composed of 300 person names and 200 web documents for each name. As the WePS organizers did in WePS-2, some person names were obtained from the following sources: US Census (50), Wikipedia (50) and Computer Science Program Committee lists (50). In addition to that, the organizers provided names for which at least one person is an attorney (50), corporate executive (50) or realtor (50). For each name the top 200 web search results from Yahoo! were provided (URL, HTML pages, search snippets and ranking information).

G.2 System Description

The system architecture has two phases that are performed sequentially: HTML Cleaning and Clustering. The HTML cleaning phase consists in to convert HTML documents into plain text. We used the existing HTMLParser[1] (version 1.6) open-source software to perform this task. For Clustering phase the following algorithms were used (separately): Lingo, Hierarchical Agglomerative Clustering (HAC), and 2-steps HAC.

G.2.1 Lingo

Lingo is an algorithm that combines Phrase Discovery (detection of topics and phrases) and Latent Semantic Indexing to organize web search results in groups based on their content Osinski and Weiss, 2005. The approach of Lingo tries to seek short and clear labels with useful meanings that could cover most of the topics of the input text collection. Lingo gets phrases with semantic content to use them as labels in the clusters, then documents are assigned to the labels to create the groups. Lingo is implemented in the Carrot2 Project[2]. Carrot2 is an Open Source Clustering software that can group automatically small collections of documents or web search results in thematic categories.

Lingo uses the Vector Space Model and Singular Value Decomposition to find the labels of the clusters. It uses 3 methods of Natural Language Processing: Stemming, stop-words, and textual segmentation heuristics. Using stemming and stop-words according Lingo developers is important when we are working with small textual information and some noise (like working with snippets).

The most used parameters for the tuning of the Lingo algorithm are the Cluster assigment threshold and the Cluster candidate label threshold. The Cluster Assignment Threshold (tcA) controls the assignments of documents to the clusters. This threshold is based on the Cosine similarity between a label and a document and its common range is from 0.15 (default) to 0.3. The Cluster Candidate Label Threshold (tcL) controls the number of clusters (labels created). This threshold is based on the Cosine similarity between a candidate cluster label and the basis vectors of the SVD decomposition. This threshold default value is 0.775 and its common value range is from 0.70 to 0.90.

G.2.2 Hierarchical Agglomerative Clustering

The Hierarchical Agglomerative Clustering method used is agglomerative, it starts at the leaves and successively merges clusters together. HAC can be stopped by distance criterion and number of clusters criterion. The Lemur[3] Information Retrieval software includes an implementation of Hierarchial Agglomerative Clustering. The clustering algorithms implemented for Lemur and used in this paper are described in Steinbach et al. (2000). These algorithms use cosine similarity in the vector space model as their metric. Stemming is

[1] HTMLParser. `http://htmlparser.sourceforge.net/`

[2] Carrot2 Project. `http://project.carrot2.org`

[3] Lemur Project. `http://www.lemurproject.org`

used using the Porter algorithm. The HAC algorithm implemented in Lemur was used in WePS2 with good results (Balog et al., 2009). The parameters accepted by Cluster are: 1) Type of cluster to use, either agglomerative or centroid (centroid is agglomerative using mean as a scoring method). 2) The scoring method to use for the agglomerative cluster over documents in a cluster maximum (max), minimum (min), average (avg), mean (mean). 3) The threshold, the minimum score for adding a document to an existing cluster.

G.2.3 2-step Clustering with Agglomerative Clustering

This is a two step algorithm that consists to cluster the results of an initial clustering process. The process follows these steps: 1) initial clustering with an agglomerative clustering algorithm that produces a set of clusters, 2) merging the content of each cluster in one new document by merging all the documents that pertain to a cluster into a one representative document for the whole cluster, 3) a second clustering step does agglomerative clustering (centroid or agglomerative configurations) over the collection of representative documents for the initial clusters.

G.3 Development experiments with WePS-3 trial data

For the WePS-3 trial evaluation a set of several experiments that consist in applying different baseline configurations (see Table G.1) to the WePS-3 trial data (WePS-2 test data) were designed.

The baseline runs were designed changing the parameters of the algorithms and the Clustering method. We did experiments with the three algorithms described before: Lingo, HAC, and 2-step HAC. We present here a set of these experiments. The experiments with Lingo share the same parameters (tcL=0.15, tcA=0.7), and differ in the kind of input to use as source documents. The following four experiments were done: full documents (1), snippets and title (2), context of the person name with 100 and 500 chars (3) (4). The experiments with agglomerative clustering differ with the type of cluster (agglomerative or centroid), type of scoring (minimum or maximum), and threshold. We did the following experiments: (5) aglomerative (agglo) with maximum score (max) and 0.07 as threshold, (6) centroid (cent) with minimum score (min) and 0.20 as threshold, (7) centroid (cent) with max and 0.05 as threshold. The experiments with 2-step clustering were in four types, i) centroid (first step) & centroid (second step) (8), ii) centroid (first step) & agglomerative (second step) (9) iii) agglomerative (first step) & centroid (second step) (10) and iv) agglomerative (first step) & agglomerative (second step): experiments from (11) to (15).

G.4 Test experiments with WePS-3 test data

For the WePS-3 evaluation with the test data the author designed a set of five experiments that consist in applying different baseline configurations to the development set data(see Table G.2). The first run (TALP_1) uses agglomerative clustering and the second run (TALP_2) uses a 2-step clustering approach with both Agglomerative clustering algorithms of Lemur. The first step does agglomerative clustering and the second step does again agglomerative clustering with the output of the first step. The third run (TALP_3) applies the algorithm Lingo for clustering. The fourth run (TALP_4) uses the centroid algorithm

Table G.1: Results with WePS-3 trial (development) data using B-Cubed measures

	Macro-averaged Scores			
	F-measures		B-Cubed	
Algorithm & Parameters	alfa=0,5	alfa=0,2	Prec.	Rec.
(10) Agglo(0.07;max)+Centroid(0.20;min)	**0,58**	0,63	0,55	0,67
(12) Agglo(0.07;max)+Agglo(0.15;max)	**0,58**	0,63	0,53	0,70
(11) Agglo (0.07;max)+Agglo(0.20;max)	**0,58**	0,62	0,55	0,66
(13) Agglo (0.07;max)+Agglo(0.10;max)	0,57	0,65	0,49	0,75
(14) Agglo (0.07;max)+Agglo (0.20;min)	0,57	0,60	0,56	0,63
(8) Centroid (0.20;min)+Cent (0.20;min)	0,56	0,67	0,47	0,80
(15) Agglo (0.07;max)+Agglo (0.07;max)	0,55	0,65	0,45	0,79
(7) Centroid (0.05;max)	0,54	0,58	0,54	0,62
(5) Agglo (0.07;max)	0,54	0,55	0,58	0,56
(6) Centroid (0.20;min)	0,53	0,52	0,61	0,52
(9) Centroid/0.07;min)+Agglo(0.03;min)	0,52	0,57	0,49	0,62
(baseline) ALL_IN_ONE	0,53	0,66	0,43	1,00
(baseline) CHEAT_SYS	0,52	0,65	0,43	1,00
(1) Lingo (Full document)	0,45	0,54	0,39	0,64
(2) Lingo (Snippets + Title)	0,43	0,44	0,47	0,46
(3) Lingo (context 500 chars)	0,42	0,42	0,51	0,43
(4) Lingo (context 100 chars)	0,43	0,42	0,53	0,42
(baseline) ONE_IN_ONE	0,34	0,27	1,00	0,24

from the Lemur. The fifth run (TALP_5) used a 2 step clustering, the first step applies the centroid algorithm of lemur and the second step applies agglomerative clustering.

Table G.2: Results with the WePS-3 Test data Task evaluated with BCubed mesures.

run	Algorithm	Parameters.	avgPrec.	avgRec.	avgF-m.(0,5)
TALP_1	Agglo	(t=0.10;max)	0.56	0.41	0.42
TALP_2	Agglo + Agglo	(t=0.10;max)	0.38	0.70	0.43
TALP_3	Lingo	(tcl=0.15;tca=0.7)	0.40	0.49	0.39
TALP_4	Centroid	(t=0.10;max)	0.60	0.41	0.43
TALP_5	**Centroid + Agglo**	(t=0.10;max)	0.40	0.66	**0.44**

G.5 Conclusions

The author presented three clustering algorithms (Lingo, HAC, and 2-step HAC) to perform the task of clustering web people search in the context of the WePS-3 Task-1.

In the preprocessing of documents HTML filtering is a crucial step to avoid noise. It is convenient to avoid noise from input documents to achieve better results, specially in the Lingo algorithm. Input noise as broken sentences and random strings could have afected the results of the clustering algorithms, specially Lingo and its cluster labels. The best results are achieved with the 2-step HAC and Agglomerative Clustering which deliver better performance than Lingo. Limited NLP processing was used only in with lingo, the other runs used Porter Stemmer before indexing and clustering. Further improvements include the use of NLP techniques for Part-of-Speech Tagging, Named Entity Recognition and Classification, and Information Extraction.

G.6 Official Team Results at WePS-3

Table G.3: Clustering Results at WePS-3: official team ranking.

		Macro-averaged Scores		
		F-measure	B-Cubed	
rank	run	*alpha*=0.5	Pre.	Rec.
1	**YHBJ_2 _unofficial**	**0.55**	0.61	0.60
2	AXIS_2	0.50	0.69	0.46
3	TALP_5	0.44	0.40	0.66
4	RGAI_AE_1	0.40	0.38	0.61
5	WOLVES_1	0.40	0.31	0.80
6	DAEDALUS_3	0.39	0.29	0.84
7	BYU	0.38	0.52	0.39
	one_in_one_baseline	0.35	1.00	0.23
8	HITSGS	0.35	0.26	0.81
	all_in_one_baseline	0.32	0.22	1.00

Bibliography

Ahlers, D. (2013). “Assessment of the Accuracy of GeoNames Gazetteer Data”. In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*. GIR ’13. Orlando, Florida: ACM, pp. 74–81. ISBN: 978-1-4503-2241-6. DOI: 10.1145/2533888.2533938. URL: http://doi.acm.org/10.1145/2533888.2533938.

Amati, G. (2003). “Probability Models for Information Retrieval Based on Divergence From Randomness”. PhD thesis. University of Glasgow.

Amitay, E., N. Har’el, R. Sivan, and A. Soffer (2005). “Web-a-where: Geotagging Web Content”. In: *SIGIR ’04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, pp. 273–280. ISBN: 1581138814.

Andogah, G. (2006). “GIR Experimentation.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Andogah, G. (2010). *Geographically Constrained Information Retrieval*. University of Groningen, The Netherlands: PhD Thesis.

Andogah, G. and G. Bouma (2007). “University of Groningen at GeoCLEF 2007”. In: *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary.

Andogah, G. and G. Bouma (2008). “Relevance Measures Using Geographic Scopes and Types”. English. In: *Advances in Multilingual and Multimodal Information Retrieval*. Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 794–801. ISBN: 978-3-540-85759-4. DOI: 10.1007/978-3-540-85760-0100. URL: http://dx.doi.org/10.1007/978-3-540-85760-0100.

Andogah, G., G. Bouma, J. Nerbonne, and E. Koster (2008). “Geographical Scope Resolution”. In: *Methodologies and Resources for Processing Spatial Language, Workshop at the 6th Conference on Language Resources and Evaluation (LREC)*. Marrakech, Marocco, pp. 4–10.

Andre, E., G. Bosch, G. Herzog, and T. Rist (1986). “Characterizing trajectories of moving objects using natural language path descriptions”. In: *Proceedings of the 7th ECAI*. Vol. 2. Brighton, UK, pp. 1–8. URL: citeseer.ist.psu.edu/andre86characterizing.html.

Artiles, J., A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó (2010). “WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks”. In: *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy.* Ed. by M. Braschler, D. Harman, and E. Pianta. Vol. 1176. CEUR Workshop Proceedings. CEUR-WS.org. ISBN: 978-88-904810-0-0. URL: http://ceur-ws.org/Vol-1176/CLEF2010wn-WePS-ArtilesEt2010.pdf.

Artiles, J., J. Gonzalo, and S. Sekine (2009). “WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task”. In: *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.*

Asadi, S., C.-Y. Chang, X. Zhou, and J. Diederich (2005). “Searching the World Wide Web for Local Services and Facilities: A Review on the Patterns of Location-Based Queries”. In: *Advances in Web-Age Information Management.* Ed. by W. Fan, Z. Wu, and J. Yang. Vol. 3739. Lecture Notes in Computer Science. 10.1007/115639529. Springer Berlin / Heidelberg, pp. 91–101. URL: http://dx.doi.org/10.1007/115639529.

Atserias, J., B. Casas, E. Comelles, M. González, L. Padró, and M. Padró (2006). “FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library”. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC’06)*, pp. 48–55.

Baba, Y., F. Ishikawa, and S. Honiden (2010). “Extraction of Places Related to Flickr Tags”. In: *ECAI.* Ed. by H. Coelho, R. Studer, and M. Wooldridge. Vol. 215. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 523–528. ISBN: 978-1-60750-605-8.

Baeza-Yates, R. A. and B. A. Ribeiro-Neto (1999). *Modern Information Retrieval.* ACM Press / Addison-Wesley. ISBN: 0-201-39829-X.

Balog, K., J. He, K. Hofmann, V. Jijkoun, C. Monz, M. Tsagkias, W. Weerkamp, and M. de Rijke (2009). “The University of Amsterdam at WePS2”. In: *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.*

Behrangi, E., H. Ghasemzadeh, K. S. Esmaili, and B. M. Bidgoli (2007). “A GEOGRAPHICAL QUESTION ANSWERING SYSTEM”. In: *Proceedings of the Third International Conference on Web Information Systems and Technologies*, pp. 308–314. ISBN: 978-972-8865-78-8. DOI: 10.5220/0001287203080314.

Benamara, F. (2004). “Cooperative Question Answering in Restricted Domains: the WEB-COOP Experiment”. In: *Proceedings of the Workshop Question Answering in Restricted Domains, within ACL-2004.*

Bensalem, I. and M.-K. Kholladi (2010). “Toponym Disambiguation by Arborescent Relationships ”. In: *Journal of Computer Science 6 (6)*, pp. 653–659.

Bischoff, K., T. Mandl, and C. Womser-Hacker (2006). “Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Bolettieri, P., A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti (2009). “CoPhIR: a Test Collection for Content-Based Image Retrieval”. In: *CoRR* abs/0905.4627v2. URL: http://cophir.isti.cnr.it.

Brants, T. (2000). “TnT – A Statistical Part-Of-Speech Tagger”. In: *Proceedings of the 6th Applied NLP Conference (ANLP-2000).* Seattle, WA, United States. URL: citeseer.ist.psu.edu/brants00tnt.html.

Braschler, M., D. Harman, and E. Pianta, eds. (2014). *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy.* Vol. 1176. CEUR Workshop Proceedings. CEUR-WS.org. ISBN: 978-88-904810-0-0. URL: `http://ceur-ws.org/Vol-1176`.

Breck, E., J. Burger, L. Ferro, D. House, M. Light, and I. Mani (1999). "A Sys Called Qanda". In: *Proceedings of the Eighth Text REtrieval Conference (TREC-8).* Ed. by E. M. Voorhees and D. K. Harman. NIST Special Publication, Gaithersburg, Md., November 1999., pp. 499–506.

Burger, J., C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, C.-Y. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Shrihari, T. Strzalkowski, E. Voorhees, and R. Weishedel (2000). "Issues, Tasks, and Program Structures to Roadmap Research in Question & Answering (Q&A)". In: URL: `http://www-nlpir.nist.gov/projects/duc/roadmapping.html`.

Buscaldi, D. (2010). "Toponym Disambiguation in Information Retrieval". PhD thesis. Universitat Politècnica de València.

Buscaldi, D. and P. Rosso (2007). "The UPV at GeoCLEF 2007". In: *Working Notes for CLEF 2007 Workshop co-located with the 11th European Conference on Digital Libraries (ECDL 2007), Budapest, Hungary, September 19-21, 2007.* Ed. by A. Nardi, C. Peters, and N. Ferro. Vol. 1173. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1173/CLEF2007wn-GeoCLEF-BuscaldiEt2007.pdf`.

Buscaldi, D. and P. Rosso (2008a). "On the Relative Importance of Toponyms in GeoCLEF". English. In: *Advances in Multilingual and Multimodal Information Retrieval.* Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 815–822. ISBN: 978-3-540-85759-4. DOI: `10.1007/978-3-540-85760-0103`. URL: `http://dx.doi.org/10.1007/978-3-540-85760-0103`.

Buscaldi, D. and P. Rosso (2008b). "The UPV at GeoCLEF 2008: The GeoWorSE System". In: *Working Notes for the CLEF 2008 Workshop.* Aarhus, Denmark.

Buscaldi, D. and P. Rosso (2011). "Explicit Query Diversification for Geographical Information Retrieval". In: *ECIR 2011 - the 33rd European Conference on Information Retrieval.* Ireland, pp. 73–80. URL: `https://hal.archives-ouvertes.fr/hal-00596899`.

Buscaldi, D., P. Rosso, and E. S. Arnal (2005). "Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task". In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 939–946. ISBN: 978-3-540-45697-1.

Buscaldi, D., P. Rosso, and E. Sanchis (2006). "WordNet-based Index Terms Expansion for Geographical Information Retrieval." In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-X.

Buscaldi, D. and P. Rosso (2008c). "A Conceptual Density-based Approach for the Disambiguation of Toponyms". In: *International Journal of Geographical Information Science* 22.3, pp. 301–313. DOI: `10.1080/13658810701626251`. eprint: `http://dx.doi.org/10.1080/13658810701626251`. URL: `http://dx.doi.org/10.1080/13658810701626251`.

Cabrio, E., J. Cojan, A. P. Aprosio, B. Magnini, A. Lavelli, and F. G (2012). "QAKiS: an open domain QA system based on relational patterns". In: *In Proc. of the 11th International Semantic Web Conference (ISWC 2012), demo paper.*

Cao, J. (2013). "Photo Set Refinement and Tag Segmentation in Georeferencing Flickr Photos". In: *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013.* Ed. by M. A. Larson, X. Anguera, T. Reuter, G. J. F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani. Vol. 1043. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1043/mediaeval2013submission49.pdf`.

Cao, J., Z. Huang, Y. Yang, and H. T. Shen (2014). "UQ-DKE's Participation at MediaEval 2014 Placing Task". In: *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.* Ed. by M. A. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. F. E. Sutcliffe, J. Choi, and G. J. F. Jones. Vol. 1263. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1263/mediaeval2014submission71.pdf`.

Carbonell, J., D. Harman, E. Hovy, S. Maiorano, J. Prange, and K. Sparck-Jones (2000). "Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization". In: *NIST draft paper.*

Cardoso, N., D. Cruz, M. Chaves, and M. J. Silva (2007). "The University of Lisbon at GeoCLEF 2007". In: *Working Notes for the CLEF 2007 Workshop.* Budapest, Hungary.

Cardoso, N., B. Martins, M. Chaves, L. Andrade, and M. J. Silva (2005). "The XLDB Group at GeoCLEF 2005". In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 997–1006. ISBN: 978-3-540-45697-1.

Carreras, X., L. Màrquez, and L. Padró (2003a). "A Simple Named Entity Extractor using AdaBoost". In: *Proceedings of CoNLL-2003.* Ed. by W. Daelemans and M. Osborne. Edmonton, Canada, pp. 152–155.

Carreras, X., L. Màrquez, and L. Padró (2003b). "Learning a Perceptron-Based Named Entity Chunker via Online Recognition Feedback". In: *Proceedings of CoNLL-2003.* Ed. by W. Daelemans and M. Osborne. Edmonton, Canada, pp. 156–159.

Chang, H.-w., D. Lee, M. Eltaher, and J. Lee (2012). "@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage". In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012).* ASONAM '12. Washington, DC, USA: IEEE Computer Society, pp. 111–118. ISBN: 978-0-7695-4799-2. DOI: `10.1109/ASONAM.2012.29`. URL: `http://dx.doi.org/10.1109/ASONAM.2012.29`.

Chen, W. (2014). "Developing a Framework for Geographic Question Answering Systems Using GIS, Natural Language Processing, Machine Learning, and Ontologies." PhD thesis. Ohio State University.

Chieu, H. L. and H. T. Ng (2003). "Named Entity Recognition with a Maximum Entropy Approach". In: *Proceedings of CoNLL-2003.* Ed. by W. Daelemans and M. Osborne. Edmonton, Canada, pp. 160–163.

Chinchor, N. and P. Robinson (1997). "MUC-7 Named Entity Task Definition (Version 3.5)". In: *Proceedings of the 7th Message Understanding Conference (MUC-7).* URL: `http://www.itl.nist.gov/iaui/894.02/relatedprojects/muc/`.

Choi, J., G. Friedland, V. N. Ekambaram, and K. Ramchandran (2012). "Multimodal Location Estimation of Consumer Media: Dealing with Sparse Training Data". In: *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, ICME 2012, Melbourne, Australia, July 9-13, 2012*. IEEE Computer Society, pp. 43–48. DOI: 10.1109/ICME.2012.141. URL: http://dx.doi.org/10.1109/ICME.2012.141.

Choi, J., C. Hauff, O. Van Laere, and B. Thomee (2015). "The Placing Task at MediaEval 2015". In: *Working Notes Proceedings of the MediaEval 2015 Workshop*.

Choi, J., C. Hauff, O. Van Laere, and B. Thomee (2016). "The Placing Task at MediaEval 2016". In: *Working Notes Proceedings of the MediaEval 2016 Workshop*.

Choi, J. and X. Li (2014). "The 2014 ICSI/TU Delft Location Estimation System". In: *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014*. Ed. by M. A. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. F. E. Sutcliffe, J. Choi, and G. J. F. Jones. Vol. 1263. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-1263/mediaeval2014submission84.pdf.

Choi, J., B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, and D. Poland (2014). "The Placing Task: A Large-Scale Geo-Estimation Challenge for Social-Media Videos and Images". In: *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*. GeoMM '14. Orlando, Florida, USA: ACM, pp. 27–31. ISBN: 978-1-4503-3127-2. DOI: 10.1145/2661118.2661125. URL: http://doi.acm.org/10.1145/2661118.2661125.

Chung, H., Y. Song, K. Han, D. Yoon, J. Lee, H. Rim, and S. Kim (2004). "A Practical QA System in Restricted Domains". In: *Proceedings of the Workshop Question Answering in Restricted Domains, within ACL-2004*.

Cleverdon, C. W. and M. Keen (1996). *Cranfield CERES: Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 2, Test results*. Tech. rep. URL: https://dspace.lib.cranfield.ac.uk/handle/1826/863.

Collins, M. (1999). "Head-Driven Statistical Models for Natural Language Parsing." PhD thesis. University of Pennsylvania.

Collins, M. (2002). "Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms". In: *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 1–8.

Crandall, D. J., L. Backstrom, D. Huttenlocher, and J. Kleinberg (2009). "Mapping the World's Photos". In: *Proceedings of the 18th international conference on World wide web*. WWW '09. Madrid, Spain: ACM, pp. 761–770. ISBN: 978-1-60558-487-4. DOI: http://doi.acm.org/10.1145/1526709.1526812. URL: http://doi.acm.org/10.1145/1526709.1526812.

Cumbreras, M. A. G., J. M. Perea-Ortega, M. G. Vega, and L. A. U. López (2009). "Information retrieval with geographical references. Relevant documents filtering vs. query expansion". In: *Inf. Process. Manage.* 45.5, pp. 605–614.

Curran, J. R. and S. Clark (2003). "Language Independent NER using a Maximum Entropy Tagger". In: *Proceedings of CoNLL-2003*. Ed. by W. Daelemans and M. Osborne. Edmonton, Canada, pp. 164–167.

De Longueville, B., R. S. Smith, and G. Luraschi (2009). ""OMG, from Here, I Can See the Flames!": A Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires". In: *Proceedings of the 2009 International Workshop on*

Location Based Social Networks. LBSN '09. Seattle, Washington: ACM, pp. 73–80. ISBN: 978-1-60558-860-5. DOI: 10.1145/1629890.1629907. URL: http://doi.acm.org/10.1145/1629890.1629907.

DeLozier, G., J. Baldridge, and L. London (2015). "Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Diaz, J., A. Rubio, A. Peinado, E. Segarra, N. Prieto, and F. Casacuberta (1998). "Development of Task-Oriented Spanish Speech Corpora". In: *Procceedings of the First International Conference on Language Resources and Evaluation*. ELDA. Granada, Spain, pp. 497–501.

Duong-Trung, N., M. Wistuba, L. R. Drumond, and L. Schmidt-Thieme (2015). "Geo_ML @ MediaEval Placing Task 2015". In: *MediaEval*. Vol. 1436. CEUR Workshop Proceedings. CEUR-WS.org.

Ellis, P. D. (2012). *The essential guide to effect sizes : statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press. ISBN: 9780521194235; 0521194237; 9780521142465; 0521142466.

Fellbaum, C., ed. (1998). *WordNet: An Electronic Lexical Database*. pub-MIT. URL: http://www.cogsci.princeton.edu/~wn/.

Ferrández, O., Z. Kozareva, A. Toral, E. Noguera, A. Montoyo, R. Muñoz, and F. Llopis (2005). "University of Alicante at GeoCLEF 2005". In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers*. Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 924–927. ISBN: 978-3-540-45697-1.

Ferrés, D., A. Ageno, and H. Rodríguez (2005a). "The GeoTALP-IR System at GeoCLEF 2005: Experiments Using a QA-Based IR System, Linguistic Analysis, and a Geographical Thesaurus". In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers*. Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 947–955. ISBN: 978-3-540-45697-1.

Ferrés, D., S. Kanaan, A. Ageno, E. González, H. Rodríguez, M. Surdeanu, and J. Turmo (2004a). "The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints." In: *CLEF*. Ed. by C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini. Vol. 3491. Lecture Notes in Computer Science. Springer, pp. 557–568. ISBN: 3-540-27420-0.

Ferrés, D., S. Kanaan, D. Domínguez-Sal, E. González, A. Ageno, M. Fuentes, H. Rodríguez, M. Surdeanu, and J. Turmo (2005b). "TALP-UPC at TREC 2005: Experiments Using Voting Scheme Among Three Hetereogeneous QA Systems". In: *Proceedings of the Fourteenth TREC Conference (TREC 2005)*. Gaithersburg, MD, USA.

Ferrés, D., S. Kanaan, E. González, A. Ageno, H. Rodríguez, M. Surdeanu, and J. Turmo (2005c). "TALP-QA System at TREC 2004: Structural and Hierarchical Relaxation Over Semantic Constraints". In: *Proceedings of the Text Retrieval Conference (TREC-2004)*.

Ferrés, D., M. Massot, M. Padró, H. Rodríguez, and J. Turmo (2004b). "Automatic Classification of Geographical Named Entities". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC*. Lisbon, Portugal.

Ferrés, D. and H. Rodríguez (2006a). "Experiments Adapting an Open-Domain Question Answering System to the Geographical Domain Using Scope-Based Resources." In: *Proceedings of the Multilingual Question Answering Workshop of the EACL 2006.* Trento, Italy, pp. 69–76. ISBN: 2-9524532-4-1.

Ferrés, D. and H. Rodríguez (2006b). "TALP at GeoCLEF-2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus." In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Ferrés, D. and H. Rodríguez (2007a). "TALP at GeoCLEF 2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus". English. In: *Evaluation of Multilingual and Multi-modal Information Retrieval.* Ed. by C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber. Vol. 4730. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 962–969. ISBN: 978-3-540-74998-1. DOI: 10.1007/978-3-540-74999-8124. URL: http://dx.doi.org/10.1007/978-3-540-74999-8124.

Ferrés, D. and H. Rodríguez (2007b). "TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering". In: *Working Notes for CLEF 2007 Workshop co-located with the 11th European Conference on Digital Libraries (ECDL 2007), Budapest, Hungary, September 19-21, 2007.* Ed. by A. Nardi, C. Peters, and N. Ferro. Vol. 1173. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-1173/CLEF2007wn-GeoCLEF-FerresEt2007a.pdf.

Ferrés, D. and H. Rodríguez (2008a). "TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier". English. In: *Advances in Multilingual and Multimodal Information Retrieval.* Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 830–833. ISBN: 978-3-540-85759-4. DOI: 10.1007/978-3-540-85760-0105. URL: http://dx.doi.org/10.1007/978-3-540-85760-0105.

Ferrés, D. and H. Rodríguez (2008b). "TALP at GeoQuery 2007: Linguistic and Geographical Analysis for Query Parsing". English. In: *Advances in Multilingual and Multimodal Information Retrieval.* Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 834–837. ISBN: 978-3-540-85759-4. DOI: 10.1007/978-3-540-85760-0106. URL: http://dx.doi.org/10.1007/978-3-540-85760-0106.

Ferrés, D. and H. Rodríguez (2010a). "TALP at GikiCLEF 2009". English. In: *Multilingual Information Access Evaluation I. Text Retrieval Experiments.* Ed. by C. Peters, G. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda. Vol. 6241. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 322–325. ISBN: 978-3-642-15753-0. DOI: 10.1007/978-3-642-15754-738. URL: http://dx.doi.org/10.1007/978-3-642-15754-738.

Ferrés, D. and H. Rodríguez (2010b). "TALP at MediaEval 2010 Placing Task: Geographical Focus Detection of Flickr Textual Annotations". In: *Working Notes of the MediaEval 2010 Workshop, October 24, 2010.* Pisa, Italy.

Ferrés, D. and H. Rodríguez (2011a). "Georeferencing Textual Annotations and Tagsets with Geographical Knowledge and Language Models". In: *Actas de la SEPLN 2011.* Huelva, Spain.

Ferrés, D. and H. Rodríguez (2011b). "TALP at MediaEval 2011 Placing Task: Georeferencing Flickr Videos with Geographical Knowledge and Information Retrieval". In: *Working*

Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011. Ed. by M. Larson, A. Rae, C. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones. Vol. 807. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-807/FerresUPCPlacingme11wn.pdf`.

Ferrés, D. and H. Rodríguez (2014). "TALP-UPC at MediaEval 2014 Placing Task: Combining Geographical Knowledge Bases and Language Models for Large-Scale Textual Georeferencing". In: *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.* Ed. by M. A. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. F. E. Sutcliffe, J. Choi, and G. J. F. Jones. Vol. 1263. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1263/mediaeval2014submission77.pdf`.

Ferrés, D. and H. Rodríguez (2015a). "Evaluating Geographical Knowledge Re-Ranking, Linguistic Processing and Query Expansion Techniques for Geographical Information Retrieval". In: *String Processing and Information Retrieval - 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015. Proceedings.* (to appear). Springer.

Ferrés, D. and H. Rodríguez (2015b). "Knowledge-Based and Data-Driven Approaches for Georeferencing of Informal Documents." In: *8th International Conference on Text, Speech and Dialogue TSD2015. Plzen, Czech Republic, September, 2015.*

Ferrés, D. and H. Rodríguez (2010). "TALP at WePS-3 2010". In: *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy.* Ed. by M. Braschler, D. Harman, and E. Pianta. Vol. 1176. CEUR Workshop Proceedings. CEUR-WS.org. ISBN: 978-88-904810-0-0. URL: `http://ceur-ws.org/Vol-1176/CLEF2010wn-WePS-FerresEt2010.pdf`.

Ferrucci, D. (2012). "Introduction to "This is Watson"". In: *IBM Journal of Research and Development* 56.3.4, 1:1–1:15. ISSN: 0018-8646. DOI: `10.1147/JRD.2012.2184356`.

Florian, R., A. Ittycheriah, H. Jing, and T. Zhang (2003). "Named Entity Recognition through Classifier Combination". In: *Proceedings of CoNLL-2003.* Ed. by W. Daelemans and M. Osborne. Edmonton, Canada, pp. 168–171.

Fonseca, F., M. Egenhofer, P. Agouris, and G. Camara (2002). "Using ontologies for integrated geographic information systems". In: *Transactions in Geographic Information Systems* 6.3.

Frew, J., M. Freeston, N. Freitas, L. L. Hill, G. Janee, K. Lovette, R. Nideffer, T. R. Smith, and Q. Zheng (1998). "The Alexandria Digital Library Architecture." In: *ECDL.* Ed. by C. Nikolaou and C. Stephanidis. Vol. 1513. Lecture Notes in Computer Science. Springer, pp. 61–73. ISBN: 3-540-65101-2.

Friedland, G., J. Choi, H. Lei, and A. Janin (2011). "Multimodal Location Estimation on Flickr Videos". In: *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media.* WSM '11. Scottsdale, Arizona, USA: ACM, pp. 23–28. ISBN: 978-1-4503-0989-9. DOI: `10.1145/2072609.2072619`. URL: `http://doi.acm.org/10.1145/2072609.2072619`.

Gaio, M., C. Sallaberry, P. Etcheverry, C. Marquesuzaa, and J. Lesbegueries (2008). "A Global Process to Access Documents' Contents from a Geographical Point of View". In: *J. Vis. Lang. Comput.* 19.1, pp. 3–23. ISSN: 1045-926X. DOI: `10.1016/j.jvlc.2007.08.010`. URL: `http://dx.doi.org/10.1016/j.jvlc.2007.08.010`.

Gale, W. A., K. W. Church, and D. Yarowsky (1992). “One Sense per Discourse”. In: *HLT ’91: Proceedings of the Workshop on Speech and Natural Language*. Harriman, New York: Association for Computational Linguistics, pp. 233–237. ISBN: 1-55860-272-0.

Gan, Q., J. Attenberg, A. Markowetz, and T. Suel (2008). “Analysis of Geographic Queries in a Search Engine Log”. In: *Proceedings of the First International Workshop on Location and the Web*. LOCWEB ’08. Beijing, China: ACM, pp. 49–56. ISBN: 978-1-60558-160-6. DOI: `10.1145/1367798.1367806`. URL: `http://doi.acm.org/10.1145/1367798.1367806`.

Garbin, E. and I. Mani (2005). “Disambiguating Toponyms in News.” In: *HLT/EMNLP*. The Association for Computational Linguistics.

García-Vega, M., M. García-Cumbreras, L. Ureña-López, and J. Perea-Ortega (2006a). “SINAI at GeoCLEF 2006: Expanding the Topics with Geographical Information and Thesaurus.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

García-Vega, M., M. Á. García-Cumbreras, L. A. Ureña-López, J. M. Perea-Ortega, F. J. Ariza-López, O. Ferrández, A. Toral, Z. Kozareva, E. Noguera, A. Montoyo, R. Muñoz, D. Buscaldi, and P. Rosso (2006b). “R2D2 at GeoCLEF 2006: a Mixed Approach.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Gaussier, É. and F. Yvon (2012). *Textual Information Access: Statistical Models*. John Wiley & Sons, p. 448.

Gey, F. C., R. R. Larson, N. Kando, J. Machado, and T. Sakai (2010). “NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search”. In: *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-8, National Center of Sciences, Tokyo, Japan, June 15-18, 2010*. Ed. by N. Kando, K. Kishida, and M. Sugimoto. National Institute of Informatics (NII), pp. 147–153. URL: `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-GeoTime-GeyF.pdf`.

Gey, F. C., R. R. Larson, J. Machado, and M. Yoshioka (2011). “NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2”. In: *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-9, National Center of Sciences, Tokyo, Japan, December 6-9, 2011*. Ed. by N. Kando, D. Ishikawa, and M. Sugimoto. National Institute of Informatics (NII). URL: `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-GEOTIME-GeyF.pdf`.

Gey, F., R. Larson, M. Sanderson, K. Bischoff, T. Mandl, C. Womser-Hacker, D. Santos, P. Rocha, G. Di Nunzio, and N. Ferro (2006). “GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Gey, F., R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras (2005). “GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview.” In: *CLEF*. Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 908–919. ISBN: 978-3-540-45697-1.

Gigerenzer, G. (2004). “Mindless statistics”. In: *Journal of Behavioral and Experimental Economics (formerly The Journal of Socio-Economics)* 33.5, pp. 587–606. URL: http://EconPapers.repec.org/RePEc:eee:soceco:v:33:y:2004:i:5:p:587-606.

Giménez, J. and L. Márquez (2004). “SVMTool: A general POS tagger generator based on Support Vector Machines”. In: *Proceedings of the 4th LREC*.

Gonzalo, J., F. Verdejo, I. Chugur, and J. Cigarran (1998). “Indexing with WordNet synsets can improve Text Retrieval”. In: *Proceedings of the COLING/ACL ’98 Workshop on Usage of WordNet for NLP*. Montreal, Canada, pp. 38–44. URL: citeseer.ist.psu.edu/gonzalo98indexing.html.

Graesser, A. C., N. Person, and J. Huber (1992). “Mechanisms that Generate Questions”. In: Lawrence Erlbaum Associates. Chap. 9, pp. 167–187.

Greenwood, M. A. (2004). “Using Pertainyms to Improve Passage Retrieval for Questions Requesting Information About a Location.” In: *Proceedings of the Workshop on Information Retrieval for Question Answering (SIGIR 2004)*.

Guillén, R. (2005). “CSUSM Experiments in GeoCLEF2005: Monolingual and Bilingual Tasks”. In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers*. Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 987–996. ISBN: 978-3-540-45697-1.

Guillén, R. (2006). “Monolingual and Bilingual Experiments in GeoCLEF2006.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Guillén, R. (2007). “GeoCLEF2007 Experiments in Query Parsing and Cross-language GIR”. In: *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary.

Guillén, R. (2008a). “GeoParsing Web Queries”. English. In: *Advances in Multilingual and Multimodal Information Retrieval*. Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 781–785. ISBN: 978-3-540-85759-4. DOI: 10.1007/978-3-540-85760-098. URL: http://dx.doi.org/10.1007/978-3-540-85760-098.

Guillén, R. (2008b). “Multi-lingual Geographical Information Retrieval”. In: *Working Notes for the CLEF 2008 Workshop*. Aarhus, Denmark.

Guttman, A. (1984). “R-Trees: A Dynamic Index Structure for Spatial Searching.” In: *SIGMOD Conference*. Ed. by B. Yormark. ACM Press, pp. 47–57.

Habib, M. B. and D. M. van Keulen (2013). “A hybrid approach for robust multilingual toponym extraction and disambiguation”. In: *International Conference on Language Processing and Intelligent Information Systems, LP&IIS 2013*. Lecture notes in computer science. Berlin, Germany: Springer.

Han, B., P. Cook, and T. Baldwin (2014). “Text-based Twitter User Geolocation Prediction”. In: *J. Artif. Int. Res.* 49.1, pp. 451–500. ISSN: 1076-9757. URL: http://dl.acm.org/citation.cfm?id=2655713.2655726.

Harabagiu, S. M., D. I. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. C. Bunescu, R. Girju, V. Rus, and P. Morarescu (2000). “FALCON: Boosting Knowledge for Answer Engines.” In: *TREC*.

Harabagiu, S., D. Moldovan, C. Clark, M. Bowden, A. Hickl, and P. Wang (2005). “Employing Two Question Answering Systems in TREC 2005.” In: *Proceedings of the Text Retrieval Conference (TREC-2005)*.

Hauff, C. and G. Houben (2011). “WISTUD at MediaEval 2011: Placing Task”. In: *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*. Ed. by M. Larson, A. Rae, C. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones. Vol. 807. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-807/HauffWISTUDPlacingme11wn.pdf`.

Hauff, C., D. Trieschnigg, and H. Rode (2006). “University of Twente at GeoCLEF 2006: Geofiltered Document Retrieval.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Hauptmann, A. G., E. G. Hauptmann, and A. M. Olligschlaeger (1999). “Using Location Information From Speech Recognition Of Television News Broadcasts”. In: *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*. http://www.scientificcommons.org/43058913: University of Cambridge, pp. 102–106.

Hays, J. and A. A. Efros (2008). “im2gps: Estimating Geographic Information from a Single Image”. In: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Hiemstra, D. (2001). “Using Language Models for Information Retrieval”. PhD thesis. Enschede. URL: `http://doc.utwente.nl/36473/`.

Hill, L. L. (2000). “Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints”. In: *ECDL 00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*. London, UK: Springer-Verlag, pp. 280–290. ISBN: 3-540-41023-6.

Hill, L. L. (2006). *Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing)*. The MIT Press. ISBN: 026208354X.

Hoffart, J., M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum (2011). “Robust disambiguation of named entities in text”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 782–792.

Hovy, E. H., L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin (2000). “Question Answering in Webclopedia.” In: *TREC*.

Hu, Y.-H. and L. Ge (2006). “UNSW at GeoCLEF 2006.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Hughes, B. (2005). “NICTA i2d2 in GeoCLEF 2005”. In: *Working Notes for CLEF 2005 Workshop co-located with the 9th European Conference on Digital Libraries (ECDL 2005), Wien, Austria, September 21-22, 2005*.

Intagorn, S. and K. Lerman (2014). “Placing User-generated Content on the Map with Confidence”. In: *ACM GIS*. URL: `http://www.isi.edu/integration/people/lerman/papers/Intagorn14acmgis.pdf`.

Ittycheriah, A., M. Franz, and S. Roukos (2001). “IBMś Statistical Question Answering System - TREC-10.” In: *TREC*.

Jijkoun, V., G. Mishne, M. de Rijke, S. Schlobach, D. Ahn, and K. M̈ller (2004). “The University of Amsterdam at QA@CLEF 2004”. In: *Results of the CLEF 2004 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2004 Workshop*. Ed. by C. Peters and F. Borri. Bath, England, pp. 321–324.

Jones, C. B., A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid (2004). “The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing”. In: *Proceedings of the Geographic Information Science: Third International Conference, GIScience 2004*. Ed. by M. J. Egenhofer, C. Freksa, and H. J. Miller. Vol. 3234 / 2004. Springer Berlin / Heidelberg. Chap. p. 125. URL: `http://www.geo-spirit.org/publications/cbjone-giscience04.pdf`.

Jones, C., R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel (2002). “Spatial Information Retrieval and Geographical Ontologies – an Overview of the SPIRIT Project”. In: *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*. URL: `citeseer.ist.psu.edu/jones02spatial.html`.

Jones, C. and R. Purves (2005). *GIR 05: Proceedings of the 2005 Workshop on Geographic Information Retrieval*. Bremen, Germany. URL: `http://portal.acm.org/citation.cfm?id=1096985`.

Jones, C. B. and R. S. Purves (2008). “Geographical Information Retrieval”. In: *International Journal of Geographical Information Science* 22.3, pp. 219–228. DOI: `10.1080/13658810701626343`. URL: `http://dx.doi.org/10.1080/13658810701626343`.

Jones, R., W. Zhang, B. Rey, P. Jhala, and E. Stipp (2008). “Geographic Intention and Modification in Web Search”. In: *Int. J. Geogr. Inf. Sci.* 22.3, pp. 229–246. ISSN: 1365-8816. DOI: `10.1080/13658810701626186`. URL: `http://dx.doi.org/10.1080/13658810701626186`.

Jones, T., A. Turpin, S. Mizzaro, F. Scholer, and M. Sanderson (2014). “Size and Source Matter: Understanding Inconsistencies in Test Collection-Based Evaluation”. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. CIKM 14. Shanghai, China: ACM, pp. 1843–1846. ISBN: 978-1-4503-2598-1. DOI: `10.1145/2661829.2661945`. URL: `http://doi.acm.org/10.1145/2661829.2661945`.

Jorg Tiedemann (2004). “A Comparison of off-the-shelf IR Engines for Question Answering”. In: *CLIN 2004*. Leiden, The Netherlands.

Karney, C. F. F. (2013). “Algorithms for geodesics”. In: *Journal of Geodesy* 87.1, pp. 43–55. ISSN: 1432-1394. DOI: `10.1007/s00190-012-0578-z`. URL: `http://dx.doi.org/10.1007/s00190-012-0578-z`.

Katz, B., J. J. Lin, and S. Felshin (2002). “The START Multimedia Information System: Current Technology and Future Directions.” In: *Multimedia Information Systems*. Arizona State University, pp. 117–123.

Kelm, P., S. Schmiedeke, and L. Goldmann (2015). “Imcube @ MediaEval 2015 Placing Task: Hierarchical Approach for Geo-referencing Large-Scale Datasets”. In: *MediaEval*. Vol. 1436. CEUR Workshop Proceedings. CEUR-WS.org.

Kelm, P., S. Schmiedeke, and T. Sikora (2010). “Video2GPS: Geotagging Using Collaborative Systems, Textual and Visual Features”. In: *Working Notes of the MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy*. Pisa, Italy.

Klein, D., J. Smarr, H. Nguyen, and C. D. Manning (2003). “Named Entity Recognition with Character-Level Models”. In: *Proceedings of CoNLL-2003*. Ed. by W. Daelemans and M. Osborne. Edmonton, Canada, pp. 180–183.

Kölle, R., B. Heuwing, T. Mandl, and C. Womser-Hacker (2007). “Monolingual Retrieval Experiments with Spatial Restrictions at GeoCLEF 2007”. In: *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary.

Kölle, R., B. Heuwing, T. Mandl, and C. Womser-Hacker (2008). "Mono-and Crosslingual Retrieval Experiments with Spatial Restrictions at GeoCLEF 2007". English. In: *Advances in Multilingual and Multimodal Information Retrieval*. Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 850–855. ISBN: 978-3-540-85759-4. DOI: `10.1007/978-3-540-85760-0109`. URL: `http://dx.doi.org/10.1007/978-3-540-85760-0109`.

Kordopatis-Zilos, G., G. Orfanidis, S. Papadopoulos, and Y. Kompatsiaris (2014). "SocialSensor at MediaEval Placing Task 2014". In: *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014*. Ed. by M. A. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. F. E. Sutcliffe, J. Choi, and G. J. F. Jones. Vol. 1263. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1263/mediaeval2014submission44.pdf`.

Kordopatis-Zilos, G., S. Papadopoulos, and Y. Kompatsiaris (2015). "Geotagging Social Media Content with a Refined Language Modelling Approach". English. In: *Intelligence and Security Informatics*. Ed. by M. Chau, G. A. Wang, and H. Chen. Vol. 9074. Lecture Notes in Computer Science. Springer International Publishing, pp. 21–40. ISBN: 978-3-319-18454-8. DOI: `10.1007/978-3-319-18455-52`. URL: `http://dx.doi.org/10.1007/978-3-319-18455-52`.

Kordopatis-Zilos, G., A. Popescu, S. Papadopoulos, and Y. Kompatsiaris (2015). "CERTH/CEA LIST at MediaEval Placing Task 2015". In: *MediaEval*. Vol. 1436. CEUR Workshop Proceedings. CEUR-WS.org.

Kordopatis-Zilos, G., A. Popescu, S. Papadopoulos, and Y. Kompatsiaris (2016). "Placing Images with Refined Language Models and Similarity Search with PCA-reduced VGG Features". In: *MediaEval*. Vol. 1739. CEUR Workshop Proceedings. CEUR-WS.org.

Kornai, A. (2005). "Evaluating Geographic Information Retrieval". In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers*. Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 928–938. ISBN: 978-3-540-45697-1.

Krippner, F., G. Meier, J. Hartmann, and R. Knauf (2011). "Placing media items using the Xtrieval Framework". In: *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*. Ed. by M. Larson, A. Rae, C. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones. Vol. 807. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-807/KrippnerCUTPlacingme11wn.pdf`.

Lana-Serrano, S., J. Goñi-Menoyo, and J. González-Cristóbal (2006a). "MIRACLE at GeoCLEF 2005: First Experiments in Geographical IR". English. In: *Accessing Multilingual Information Repositories*. Ed. by C. Peters, F. Gey, J. Gonzalo, H. Müller, G. Jones, M. Kluck, B. Magnini, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 920–923. ISBN: 978-3-540-45697-1. DOI: `10.1007/11878773102`. URL: `http://dx.doi.org/10.1007/11878773102`.

Lana-Serrano, S., J. Goñi-Menoyo, and J. González-Cristóbal (2006b). "Report of MIRACLE Team for Geographical IR in CLEF 2006." In: *Working Notes of the Cross-Lingual*

Evaluation Forum (CLEF) 2006. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Lana-Serrano, S., J. Villena-Román, J. González-Cristóbal, and J. Goñi-Menoyo (2008). “MIRACLE at GeoCLEF Query Parsing 2007: Extraction and Classification of Geographical Information”. English. In: *Advances in Multilingual and Multimodal Information Retrieval*. Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 786–793. ISBN: 978-3-540-85759-4. DOI: 10.1007/978-3-540-85760-099. URL: http://dx.doi.org/10.1007/978-3-540-85760-099.

Larson, M., M. Soleymani, P. Serdyukov, V. Murdock, and G. Jones, eds. (2010). Pisa, Italy. ISBN: 9789081348904. URL: http://www.multimediaeval.org/mediaeval2010/2010worknotes/.

Larson, M. A., B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. F. E. Sutcliffe, J. Choi, and G. J. F. Jones, eds. (2014). *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014*. Vol. 1263. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-1263.

Larson, M., P. Kelm, A. Rae, C. Hauff, B. Thomee, M. Trevisiol, J. Choi, O. Van Laere, S. Schockaert, G. Jones, P. Serdyukov, V. Murdock, and G. Friedland (2015). “The Benchmark as a Research Catalyst: Charting the Progress of Geo-prediction for Social Multimedia”. English. In: *Multimodal Location Estimation of Videos and Images*. Ed. by J. Choi and G. Friedland. Springer International Publishing, pp. 5–40. ISBN: 978-3-319-09860-9. DOI: 10.1007/978-3-319-09861-62. URL: http://dx.doi.org/10.1007/978-3-319-09861-62.

Larson, M., A. Rae, C. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones, eds. (2011). *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*. Vol. 807. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-807.

Larson, R. R. (2007). “Cheshire at GeoCLEF 2007: Retesting Text Retrieval Baselines”. In: *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary.

Larson, R. R. (2008). “Cheshire at GeoCLEF 2008: Text and Fusion Approaches for GIR”. In: *Working Notes for the CLEF 2008 Workshop*. Aarhus, Denmark.

Larson, R., F. Gey, and V. Petras (2006). “Berkeley at GeoCLEF: Logistic Regression and Fusion for Geographic Information Retrieval”. English. In: *Accessing Multilingual Information Repositories*. Ed. by C. Peters, F. Gey, J. Gonzalo, H. Müller, G. Jones, M. Kluck, B. Magnini, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 963–976. ISBN: 978-3-540-45697-1. DOI: 10.1007/11878773108. URL: http://dx.doi.org/10.1007/11878773108.

Larson, R. and F. Gey (2006). “GeoCLEF Text Retrieval and Manual Expansion Approaches.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Lehnert, W. G. (1978). *The Process of Question Answering*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Leidner, J. (2007). *Toponym Resolution: a Comparison and Taxonomy of Heuristics and Methods*. University of Edinburgh: PhD Thesis.

Leidner, J. “Re-Ranking for Geo-Relevance With Non-Contextual Heuristics at GeoCLEF 2007”. In:

Leidner, J. (2005). “Experiments with Geo-Filtering Predicates for IR”. In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 987–996. ISBN: 978-3-540-45697-1.

Leidner, J. (2006). *Toponym Resolution: A First Large-Scale Comparative Evaluation.* Research Report EDI–INF–RR–0839. Edinburgh, Scotland, UK: School of Informatics, University of Edinburgh.

Leidner, J., G. Sinclair, and B. Webber (2003). “Grounding Spatial Named Entities for Information Extraction and Question Answering”. In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references.* Ed. by A. Kornai and B. Sundheim. Morristown, NJ, USA: Association for Computational Linguistics, pp. 31–38. DOI: `http://dx.doi.org/10.3115/1119394.1119399`.

Leveling, J., S. Hartrumpf, and D. Veiel (2005). “Using Semantic Networks for Geographic Information Retrieval”. In: *CLEF.* Ed. by C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke. Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer, pp. 977–986. ISBN: 978-3-540-45697-1.

Leveling, J. and D. Veiel (2006). “University of Hagen at GeoCLEF2006: Experiments with metonymy recognition in documents”. In: *CLEF.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Lewis, D. D., Y. Yang, T. G. Rose, and F. Li (2004). “RCV1: A New Benchmark Collection for Text Categorization Research”. In: *J. Mach. Learn. Res.* 5, pp. 361–397. ISSN: 1532-4435. URL: `http://dl.acm.org/citation.cfm?id=1005332.1005345`.

Li, H., K. Srihari, C. Niu, and W. Li (2003). “InfoXtract location normalization: a hybrid approach to geographic references in information extraction”. In: *HLT-NAACL 2003 Workshop: Analysis of Geographic References.* Ed. by A. Kornai and B. Sundheim. Edmonton, Alberta, Canada: Association for Computational Linguistics, pp. 39–44.

Li, H., R. Srihari, C. Niu, and W. Li (2002). “Location Normalization for Information Extraction.” In: *COLING.*

Li, L. T., J. Almeida, and R. Torres (2011). “RECOD Working Notes for Placing Task MediaEval 2011”. In: *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011.* Ed. by M. Larson, A. Rae, C. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones. Vol. 807. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-807/LiUNICAMPPlacingme11wn.pdf`.

Li, L. T., J. A. V. Muñoz, J. Almeida, R. T. Calumby, O. A. B. Penatti, Í. C. Dourado, K. Nogueira, P. R. Mendes-Junior, L. A. M. Pereira, D. C. G. Pedronette, J. A. dos Santos, M. A. Gonçalves, and R. da Silva Torres (2015). “RECOD @ Placing Task of MediaEval 2015”. In: *MediaEval.* Vol. 1436. CEUR Workshop Proceedings. CEUR-WS.org.

Li, L., O. Penatti, J. Almeida, G. Chiachia, R. T. Calumby, P. R. Mendes-Junior, D. C. G. Pedronette, and R. da Silva Torres (2014). “Multimedia Geocoding: The RECOD 2014 Approach”. In: *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.* Ed. by M. A. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. F. E. Sutcliffe, J. Choi, and G. J. F. Jones. Vol. 1263. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1263/mediaeval2014submission81.pdf`.

Li, X. and D. Roth (2002). "Learning Question Classifiers". In: *Proceedings of the 19th International Conference on Computational Linguistics, 2002*. URL: `citeseer.ist.psu.edu/article/li02learning.html`.

Li, X. and D. Roth (2004). "Learning Question Classifiers: The Role of Semantic Information". In: *Natural Language Engineering* 1.1.

Li, Z., C. Wang, X. Xie, and W. Ma (2006). "MSRA Columbus at GeoCLEF 2006." In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Li, Z., C. Wang, X. Xie, and W. Ma (2007a). "MSRA Columbus at GeoCLEF2007". In: *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary.

Li, Z., C. Wang, X. Xie, and W. Ma (2007b). "Query Parsing Task for GeoCLEF2007 Report". In: *This volume*. Ed. by A. Nardi and C. Peters. Budapest, Hungary. ISBN: 2-912335-23-x.

Li, Z., C. Wang, X. Xie, and W. Ma (2007c). "Query Parsing Task for GeoCLEF2007 Report". In: *Working Notes for CLEF 2007 Workshop co-located with the 11th European Conference on Digital Libraries (ECDL 2007), Budapest, Hungary, September 19-21, 2007*. Ed. by A. Nardi, C. Peters, and N. Ferro. Vol. 1173. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1173/CLEF2007wn-GeoCLEF-LiEt2007a.pdf`.

Lieberman, M., H. Samet, and J. Sankaranarayanan (2010a). "Geotagging with local lexicons to build indexes for textually-specified spatial data". In: *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pp. 201–212. DOI: `10.1109/ICDE.2010.5447903`.

Lieberman, M. D. and H. Samet (2011). "Multifaceted Toponym Recognition for Streaming News". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '11. Beijing, China: ACM, pp. 843–852. ISBN: 978-1-4503-0757-4. DOI: `10.1145/2009916.2010029`. URL: `http://doi.acm.org/10.1145/2009916.2010029`.

Lieberman, M. D. and H. Samet (2012). "Adaptive Context Features for Toponym Resolution in Streaming News". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. Portland, Oregon, USA: ACM, pp. 731–740. ISBN: 978-1-4503-1472-5. DOI: `10.1145/2348283.2348381`. URL: `http://doi.acm.org/10.1145/2348283.2348381`.

Lieberman, M., H. Samet, J. Sankaranarayanan, and J. Sperling (2007). "STEWARD: Architecture of a Spatio-textual Search Engine". In: *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*. GIS '07. Seattle, Washington: ACM, 25:1–25:8. ISBN: 978-1-59593-914-2. DOI: `10.1145/1341012.1341045`. URL: `http://doi.acm.org/10.1145/1341012.1341045`.

Lieberman, M., H. Samet, and J. Sankaranayananan (2010b). "Geotagging: Using Proximity, Sibling, and Prominence Clues to Understand Comma Groups". In: *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*. Zurich, Switzerland: ACM, pp. 1–8. ISBN: 978-1-60558-826-1. DOI: `10.1145/1722080.1722088`. URL: `http://dx.doi.org/10.1145/1722080.1722088`.

Lin, C.-Y. and E. Hovy (2000). "The Automated Acquisition of Topic Signatures for Text Summarization." In: *COLING*. Morgan Kaufmann, pp. 495–501.

Lin, J. and B. Katz (2003). "Question Answering From the Web Using Knowledge Annotation and Knowledge Mining Techniques". In: *CIKM '03: Proceedings of the twelfth* *international conference on Information and knowledge management*. New Orleans, LA, USA: ACM Press, pp. 116–123. ISBN: 1-58113-723-0. DOI: http://doi.acm.org/10.1145/956863.956886.

Liu, S., F. Liu, C. Yu, and W. Meng (2004). "An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases". In: *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. Sheffield, United Kingdom: ACM Press, pp. 266–272. ISBN: 1-58113-881-4. DOI: http://doi.acm.org/10.1145/1008992.1009039.

Lo, K. K. and W. Lam (2006). "Using Semantic Relations with World Knowledge for Question Answering". In: *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Luque, J., D. Ferrés, J. Hernando, J. B. Mariño, and H. Rodríguez (2006). "GeoVAQA: A Voice Activated Geographical Question Answering System." In: *Actas de las IV Jornadas en Tecnología del Habla (4JTH)*. Zaragoza, Spain.

Magnini, B., S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. de Rijke (2003). "Creating the DISEQuA Corpus: A Test Set for Multilingual Question Answering." In: *CLEF*. Ed. by C. Peters, J. Gonzalo, M. Braschler, and M. Kluck. Vol. 3237. Lecture Notes in Computer Science. Springer, pp. 487–500. ISBN: 3-540-24017-9.

Mahmud, J., J. Nichols, and C. Drews (2014). "Home Location Identification of Twitter Users". In: *ACM Trans. Intell. Syst. Technol.* 5.3, 47:1–47:21. ISSN: 2157-6904. DOI: 10.1145/2528548. URL: http://doi.acm.org/10.1145/2528548.

Mandl, T., P. Carvalho, G. Di Nunzio, F. Gey, R. Larson, D. Santos, and C. Womser-Hacker (2009). "GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview". English. In: *Evaluating Systems for Multilingual and Multimodal Information Access*. Ed. by C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. Jones, M. Kurimo, T. Mandl, A. Peñas, and V. Petras. Vol. 5706. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 808–821. ISBN: 978-3-642-04446-5. DOI: 10.1007/978-3-642-04447-2106. URL: http://dx.doi.org/10.1007/978-3-642-04447-2106.

Mandl, T., F. Gey, G. Di Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie (2008). "GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview". English. In: *Advances in Multilingual and Multimodal Information Retrieval*. Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 745–772. ISBN: 978-3-540-85759-4. DOI: 10.1007/978-3-540-85760-096. URL: http://dx.doi.org/10.1007/978-3-540-85760-096.

Mandl, T., F. Gey, G. D. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie (2007). "GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview". In: *Results of the CLEF 2007 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2007 Workshop*. Ed. by A. Nardi and C. Peters. Budapest, Hungary. ISBN: 2-912335-23-x.

Mani, I., J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner. "SpatialML: Annotation Scheme, Corpora, and Tools". In:

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press. ISBN: 0521865719, 9780521865715.

Manov, D., A. Kiryakov, B. Popov, K. Bontcheva, D. Maynard, and H. Cunningham (2003). " Experiments with geographic knowledge for information extraction ". In: *HLT-NAACL 2003 Workshop: Analysis of Geographic References.* Ed. by A. Kornai and B. Sundheim. Edmonton, Alberta, Canada: Association for Computational Linguistics, pp. 1–9.

Markowetz, A., Y. Chen, T. Suel, X. Long, and B. Seeger (2005). "Design and Implementation of a Geographic Search Engine". In: *WebDB*, pp. 19–24.

Marrero, M., J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís (2013). "Named Entity Recognition: Fallacies, challenges and opportunities". In: *Computer Standards & Interfaces* 35.5, pp. 482–489. ISSN: 0920-5489. DOI: http://dx.doi.org/10.1016/j.csi.2012.09.004. URL: http://www.sciencedirect.com/science/article/pii/S0920548912001080.

Martins, B. (2008). "Geographically Aware Web Text Mining". PhD thesis. Faculty of Sciences, University of Lisbon.

Martins, B., J. Borbinha, G. Pedrosa, J. Gil, and N. Freire (2007a). "Geographically-aware Information Retrieval for Collections of Digitized Historical Maps". In: *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval.* GIR '07. Lisbon, Portugal: ACM, pp. 39–42. ISBN: 978-1-59593-828-2. DOI: 10.1145/1316948.1316959. URL: http://doi.acm.org/10.1145/1316948.1316959.

Martins, B. and P. Calado (2010). "Learning to Rank for Geographic Information Retrieval". In: *Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR 2010, Zurich, Switzerland, February 18-19, 2010.* Ed. by R. Purves, P. D. Clough, and C. B. Jones. ACM. ISBN: 978-1-60558-826-1. DOI: 10.1145/1722080.1722107. URL: http://doi.acm.org/10.1145/1722080.1722107.

Martins, B., N. Cardoso, M. Chaves, L. Andrade, and M. Silva (2006). "The University of Lisbon at GeoCLEF 2006." In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Martins, B., N. Cardoso, M. Chaves, L. Andrade, and M. Silva (2007b). "The University of Lisbon at GeoCLEF 2006". English. In: *Evaluation of Multilingual and Multi-modal Information Retrieval.* Ed. by C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber. Vol. 4730. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 986–994. ISBN: 978-3-540-74998-1. DOI: 10.1007/978-3-540-74999-8127. URL: http://dx.doi.org/10.1007/978-3-540-74999-8127.

Martins, B. and M. J. Silva (2005). "A Graph-Ranking Algorithm for Geo-referencing Documents". In: *Proceedings of ICDM-05, the 5Th IEEE International Conference on Data Mining.*

Martins, B., M. J. Silva, and L. Andrade (2005). "Indexing and ranking in Geo-IR systems". In: *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval.* Bremen, Germany: ACM Press, pp. 31–34. ISBN: 1-59593-165-1. DOI: http://doi.acm.org/10.1145/1096985.1096993.

Merchant, R. and M. E. Okurowski (1996). "The Multilingual Entity Task (MET) Overview". In: *Proceedings of TIPSTER Text Program (Phase II).*

Metzler, D. and W. B. Croft (2004). "Combining the Language Model and Inference Network Approaches to Retrieval". In: *Inf. Process. Manage.* 40.5, pp. 735–750. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2004.05.001. URL: http://dx.doi.org/10.1016/j.ipm.2004.05.001.

Mihalcea, R. and D. Moldovan (2000). “Semantic indexing using WordNet senses”. In: *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval.* Hong Kong: Association for Computational Linguistics, pp. 35–45.

Minock, M. (2005). “A phrasal approach to natural language interfaces over databases”. In: *International Conference on Application of Natural Language to Information Systems.* Springer, pp. 333–336.

Mishra, A., N. Mishra, and A. Agrawal (2010). “Context-Aware Restricted Geographical Domain Question Answering System”. In: *2010 International Conference on Computational Intelligence and Communication Networks*, pp. 548–553.

Moldovan, D., M. Bowden, and M. Tatu (2006). “A Temporally-Enhanced PowerAnswer in TREC 2006”. In: *”The Fifteenth Text REtrieval Conference (TREC 2006) Notebook”.* Gaithersburg, MD, USA: NIST.

Moldovan, D., S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus (1999). “LASSO: A tool for surfing the answer net”. In: *Proceedings of the Eighth Text Retrieval Conference (TREC-8).* URL: `citeseer.ist.psu.edu/moldovan99lasso.html`.

Moldovan, D., C. Clark, S. Harabagiu, and S. Maiorano (2003). “COGEX: a logic prover for question answering”. In: *NAACL ’03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* Edmonton, Canada: Association for Computational Linguistics, pp. 87–93.

Molla, D. and J. Vicedo (2005). *AAAI-05 Workshop on Question Answering in Restricted Domains.* to appear. AAAI Press.

Monz, C. (2003). “From Document Retrieval to Question Answering.” PhD thesis. University of Amsterdam.

Muñoz, J. A. V., L. T. Li, Í. C. Dourado, K. Nogueira, S. G. Fadel, O. A. B. Penatti, J. Almeida, L. A. M. Pereira, R. T. Calumby, J. A. dos Santos, and R. da Silva Torres (2016). “RECOD @ Placing Task of MediaEval 2016: A Ranking Fusion Approach for Geographic-Location Prediction of Multimedia Objects”. In: *MediaEval.* Vol. 1739. CEUR Workshop Proceedings. CEUR-WS.org.

Nadeau, D. and S. Sekine (2007). “A survey of named entity recognition and classification”. In: *Lingvistic Investigationes* 30.1, pp. 3–26. DOI: `http://dx.doi.org/10.1075/li.30.1.03nad`. URL: `http://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad`.

Ogilvie, P. and J. P. Callan (2001). “Experiments Using the Lemur Toolkit”. In: *Proceedings of the Text REtrieval Conference (TREC-10).* URL: `citeseer.ist.psu.edu/610389.html`.

Osinski, S. and D. Weiss (2005). “A Concept-Driven Algorithm for Clustering Search Results”. In: *IEEE Intelligent Systems* 20.3, pp. 48–54.

Ounis, I., G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma (2006). “Terrier: A High Performance and Scalable Information Retrieval Platform”. In: *Proceedings of ACM SIGIR’06 Workshop on Open Source Information Retrieval (OSIR 2006).* Seattle, Washington, USA.

Overell, S. (2009). “Geographic Information Retrieval: Classification, Disambiguation and Modelling”. PhD thesis. Imperial College London. URL: `http://www.numenore.co.uk/wiki`.

Overell, S. E., J. Magalhães, and S. M. Rüger (2007). "GIR Experiments with Forostar at GeoCLEF 2007". In: *Working Notes for CLEF 2007 Workshop co-located with the 11th European Conference on Digital Libraries (ECDL 2007), Budapest, Hungary, September 19-21, 2007.* Ed. by A. Nardi, C. Peters, and N. Ferro. Vol. 1173. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1173/CLEF2007wn-GeoCLEF-OverellEt2007.pdf`.

Overell, S., J. Magalhães, and S. Rüger (2006). "Place disambiguation with co-occurrence models." In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Overell, S., J. Magalhães, and S. Rüger (2008a). "GIR Experiments with Forostar". English. In: *Advances in Multilingual and Multimodal Information Retrieval.* Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, V. Petras, and D. Santos. Vol. 5152. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 856–863. ISBN: 978-3-540-85759-4. DOI: `10.1007/978-3-540-85760-0110`. URL: `http://dx.doi.org/10.1007/978-3-540-85760-0110`.

Overell, S., A. Rae, and S. Rüger (2008b). "MMIS at GeoCLEF 2008: Experiments in GIR". In: *Working Notes for the CLEF 2008 Workshop.* Aarhus, Denmark.

Padró, L. and E. Stanilovsky (2012). "FreeLing 3.0: Towards Wider Multilinguality". In: *Proceedings of the Language Resources and Evaluation Conference (LREC 2012).* ELRA. Istanbul, Turkey.

Palacio, D. (2010). "Combination of criteria with constraints for Geographic Information Retrieval". Theses. Université de Pau et des Pays de l'Adour. URL: `https://tel.archives-ouvertes.fr/tel-00551889`.

Palacio, D., G. Cabanac, C. Sallaberry, and G. Hubert (2010). "On the evaluation of Geographic Information Retrieval systems - Evaluation framework and case study". In: *Int. J. on Digital Libraries* 11.2, pp. 91–109. DOI: `10.1007/s00799-011-0070-z`. URL: `https://doi.org/10.1007/s00799-011-0070-z`.

Pasca, M. and S. M. Harabagiu (2001a). "Answer mining from on-line documents". In: *Proceedings of the workshop on ARABIC language processing.* Toulouse, France: Association for Computational Linguistics, pp. 1–8.

Pasca, M. and S. M. Harabagiu (2001b). "High Performance Question/Answering". In: *Research and Development in Information Retrieval*, pp. 366–374. URL: `citeseer.ist.psu.edu/pasca01high.html`.

Pascual, F. L. (2002). "IR-n Un Sistema de Recuperación de Información Basado en Pasajes." PhD thesis. Universidad de Alicante.

Passos, A., V. Kumar, and A. McCallum (2014). "Lexicon Infused Phrase Embeddings for Named Entity Resolution". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014.* Ed. by R. Morante and W. Yih. ACL, pp. 78–86. URL: `http://aclweb.org/anthology/W/W14/W14-1609.pdf`.

Perea-Ortega, J. M., M. A. García-Cumbreras, M. García-Vega, and A. Montejo-Ráez (2007). "GEOUJA System. University of Jaén at GeoCLEF 2007". In: *Working Notes for the CLEF 2007 Workshop.* Budapest, Hungary.

Perea-Ortega, J. M. (2010). "Recuperacion de informacion geografica basada en multiples formulaciones y motores de busqueda". PhD thesis. Universidad de Jaén.

Perea-Ortega, J., M. García-Cumbreras, L. Ureña-López, and M. García-Vega (2010). "SINAI at Placing Task of MediaEval 2010". In: *Working Notes of the MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy.* Pisa, Italy.

Perea-Ortega, J., M. García-Cumbreras, L. Ureña-López, and M. García-Vega (2011). "Geo-Textual Relevance Ranking to Improve a Text-Based Retrieval for Geographic Queries". English. In: *Natural Language Processing and Information Systems.* Ed. by R. Muñoz, A. Montoyo, and E. Métais. Vol. 6716. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 278–281. ISBN: 978-3-642-22326-6. DOI: `10.1007/978-3-642-22327-338`. URL: `http://dx.doi.org/10.1007/978-3-642-22327-338`.

Perea-Ortega, J., M.-A. García-Cumbreras, M. García-Vega, and L.-A. Ureña-López (2008a). "SINAI-GIR System. University of Jaén at GeoCLEF 2008". In: *Working Notes for the CLEF 2008 Workshop.* Aarhus, Denmark.

Perea-Ortega, J., L. Ureña-López, D. Buscaldi, and P. Rosso (2008b). "TextMESS at GeoCLEF 2008: Result Merging with Fuzzy Borda Ranking ". In: *Working Notes for the CLEF 2008 Workshop.* Aarhus, Denmark.

Peters, C., F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke, eds. (2006). *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* Vol. 4022. Lecture Notes in Computer Science. Berlin: Springer. ISBN: 978-3-540-45697-1.

Popescu, A. (2013). "CEA LIST's Participation at MediaEval 2013 Placing Task". In: *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013.* Ed. by M. A. Larson, X. Anguera, T. Reuter, G. J. F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani. Vol. 1043. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1043/mediaeval2013submission59.pdf`.

Popescu, A. and N. Ballas (2012). "CEA LIST's Participation at MediaEval 2012 Placing Task". In: *Working Notes Proceedings of the MediaEval 2012 Workshop, Santa Croce in Fossabanda, Pisa, Italy, October 4-5, 2012.* Ed. by M. A. Larson, S. Schmiedeke, P. Kelm, A. Rae, V. Mezaris, T. Piatrik, M. Soleymani, F. Metze, and G. J. F. Jones. Vol. 927. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-927/mediaeval2012submission32.pdf`.

Popescu, A. and I. Kanellos (2009). "Creating Visual Summaries for Geographic Regions". In: *ECIR Workshop on Information Retrieval over Social Networks.*

Popescu, A., P.-A. Moëllic, and I. Kanellos (2008). "Themeexplorer: Finding and Browsing Geo-Referenced Images". In: *Proceedings of the International Workshop Content-Based Multimedia Indexing.* London, United Kingdom, pp. 576–583.

Popescu, A., S. Papadopoulos, and I. Kompatsiaris (2014). "USEMP at MediaEval Placing Task 2014". In: *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.* Ed. by M. A. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. F. E. Sutcliffe, J. Choi, and G. J. F. Jones. Vol. 1263. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1263/mediaeval2014submission70.pdf`.

Porter, M. F. (1997). "An algorithm for suffix stripping". In: San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 313–316. ISBN: 1-55860-454-5.

Pouliquen, B., R. Steinberger, C. Ignat, and T. D. Groeve (2004). "Geographical information recognition and visualization in texts written in various languages". In: *SAC*

'04: Proceedings of the 2004 ACM symposium on Applied computing. Nicosia, Cyprus: ACM Press, pp. 1051–1058. ISBN: 1-58113-812-1. DOI: http://doi.acm.org/10.1145/967900.968115.

Pradel, C., O. Haemmerlé, and N. Hernandez (2011). "A Semantic Web Interface Using Patterns: The SWIP System". In: *Graph Structures for Knowledge Representation and Reasoning - Second International Workshop, GKR 2011, Barcelona, Spain, July 16, 2011. Revised Selected Papers.* Ed. by M. Croitoru, S. Rudolph, N. Wilson, J. Howse, and O. Corby. Vol. 7205. Lecture Notes in Computer Science. Springer, pp. 172–187. DOI: 10.1007/978-3-642-29449-57. URL: http://dx.doi.org/10.1007/978-3-642-29449-57.

Prager, J., E. Brown, A. Coden, and D. Radev (2000). "Question-answering by predictive annotation". In: *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.* Athens, Greece: ACM Press, pp. 184–191. ISBN: 1-58113-226-3. DOI: http://doi.acm.org/10.1145/345508.345574.

Pu, Q., D. He, and Q. Li (2008). "University of Pittsburgh at GeoCLEF 2008: Towards Effective Geographic Information Retrieval". In: *Working Notes for the CLEF 2008 Workshop.* Aarhus, Denmark.

Purves, R. S., P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang (2007). "The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet". In: *International Journal of Geographical Information Science* 21.7, pp. 717–745. DOI: 10.1080/13658810601169840. eprint: http://dx.doi.org/10.1080/13658810601169840. URL: http://dx.doi.org/10.1080/13658810601169840.

Radev, D., W. Fan, H. Qi, H. Wu, and A. Grewal (2002). "Probabilistic question answering on the web". In: *WWW '02: Proceedings of the 11th international conference on World Wide Web.* Honolulu, Hawaii, USA: ACM Press, pp. 408–419. ISBN: 1-58113-449-5. DOI: http://doi.acm.org/10.1145/511446.511500.

Rae, A., V. Murdock, P. Serdyukov, and P. Kelm (2011). "Working Notes for the Placing Task at MediaEval 2011". In: *Working Notes of the MediaEval 2011 Workshop.* Pisa, Italy.

Ratinov, L. and D. Roth (2009). "Design Challenges and Misconceptions in Named Entity Recognition". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning.* CoNLL '09. Boulder, Colorado: Association for Computational Linguistics, pp. 147–155. ISBN: 978-1-932432-29-9. URL: http://dl.acm.org/citation.cfm?id=1596374.1596399.

Rattenbury, T. and M. Naaman (2009). "Methods for Extracting Place Semantics from Flickr Tags". In: *ACM Trans. Web* 3 (1), 1:1–1:30. ISSN: 1559-1131. DOI: http://doi.acm.org/10.1145/1462148.1462149. URL: http://doi.acm.org/10.1145/1462148.1462149.

Rauch, E., M. Bukatin, and K. Baker (2003). " A confidence-based framework for disambiguating geographic terms ". In: *HLT-NAACL 2003 Workshop: Analysis of Geographic References.* Ed. by A. Kornai and B. Sundheim. Edmonton, Alberta, Canada: Association for Computational Linguistics, pp. 50–54.

Ravichandran, D. and E. H. Hovy (2002). "Learning surface text patterns for a Question Answering System." In: *ACL*, pp. 41–47.

Rees, T. (2003). “C-Squares, a new spatial indexing system and its applicability to the description of oceanographic data.” In: *Ocenography* 16.1, pp. 11–19.

Rijsbergen, C. J. V. (1979). *Information Retrieval.* 2nd. Newton, MA, USA: Butterworth-Heinemann. ISBN: 0408709294.

Ripley, B. (2005). *Spatial Statistics.* Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471725206. URL: `http://books.google.ca/books?id=eB%5CXYz3B7qQC`.

Roberts, I. and R. J. Gaizauskas (2004). “Evaluating Passage Retrieval Approaches for Question Answering.” In: *ECIR.* Ed. by S. McDonald and J. Tait. Vol. 2997. Lecture Notes in Computer Science. Springer, pp. 72–84. ISBN: 3-540-21382-1.

Roberts, K., C. A. Bejan, and S. Harabagiu (2010). “Toponym Disambiguation Using Events”. In: *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010).*

Robertson, S. E. and S. Walker (1994). “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval.” In: *SIGIR.* Ed. by W. B. Croft and C. J. van Rijsbergen. ACM/Springer, pp. 232–241. ISBN: 3-540-19889-X.

Roller, S., M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge (2012). “Supervised Text-based Geolocation Using Language Models on an Adaptive Grid”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* EMNLP-CoNLL ’12. Jeju Island, Korea: Association for Computational Linguistics, pp. 1500–1510. URL: `http://dl.acm.org/citation.cfm?id=2390948.2391120`.

Ruiz, M. E., S. Shapiro, J. Abbas, S. B. Southwick, and D. Mark (2006). “UB at GeoCLEF 2006.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Sakai, T. (2014). “Statistical Reform in Information Retrieval?” In: *SIGIR Forum* 48.1, pp. 3–12. ISSN: 0163-5840. DOI: `10.1145/2641383.2641385`. URL: `http://doi.acm.org/10.1145/2641383.2641385`.

Salton, G. and C. Buckley (1988). “Term-Weighting Approaches in Automatic Text Retrieval.” In: *Inf. Process. Manage.* 24.5, pp. 513–523.

Sanderson, M. and J. Kohler (2004a). “Analyzing Geographic Queries.” In: *Proceedings of the Geographic Information Retrieval Workshop of the SIGIR 2004.*

Sanderson, M. and J. Kohler (2004b). “Analyzing Geographic Queries”. In: *SIGIR Workshop on Geographic Information Retrieval* 2.

Sanderson, M. and J. Zobel (2005). “Information retrieval system evaluation: effort, sensitivity, and reliability”. In: *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005.* Ed. by R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait. ACM, pp. 162–169. DOI: `10.1145/1076034.1076064`. URL: `http://doi.acm.org/10.1145/1076034.1076064`.

Santos, D. and L. M. Cabral (2009). “GikiCLEF: Crosscultural Issues in an International Setting: Asking non-English-centered Questions to Wikipedia”. In: *Working Notes for CLEF 2009 Workshop co-located with the 13th European Conference on Digital Libraries (ECDL 2009) , Corfù, Greece, September 30 - October 2, 2009.* Ed. by C. Peters and N. Ferro. Vol. 1175. CEUR Workshop Proceedings. CEUR-WS.org. URL: `http://ceur-ws.org/Vol-1175/CLEF2009wn-QACLEF-SantosEt2009.pdf`.

Santos, D. and L. Cabral (2010). “GikiCLEF: Expectations and Lessons Learned”. English. In: *Multilingual Information Access Evaluation I. Text Retrieval Experiments.* Ed. by

C. Peters, G. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda. Vol. 6241. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 212–222. ISBN: 978-3-642-15753-0. DOI: 10.1007/978-3-642-15754-723. URL: http://dx.doi.org/10.1007/978-3-642-15754-723.

Santos, D., N. Cardoso, P. Carvalho, I. Dornescu, S. Hartrumpf, J. Leveling, and Y. Skalban (2008). "Getting Geographical Answers from Wikipedia: the GikiP pilot at CLEF". In: *Working notes for the CLEF 2008 Workshop*. Ed. by Francesca Borri and Alessandro Nardi and Carol Peters. Aarhus, Denmark: CLEF 2008 Organizing Committee.

Santos, J., I. Anastácio, and B. Martins (2015). "Using machine learning methods for disambiguating place references in textual documents". In: *GeoJournal* 80.3, pp. 375–392. ISSN: 1572-9893.

Sasaki, Y., H.-H. Chen, K.-h. Chen, and C.-J. Lin (2005). "Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1)." In: *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp. 54–68.

Schulz, A., A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhüser (2013). *A Multi-Indicator Approach for Geolocalization of Tweets.* URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6063/6397.

Sekine, S., K. Sudo, and C. Nobata (2002). "Extended Named Entity Hierarchy". In: *Proceedings of Thirth International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Spain.

Serdyukov, P., V. Murdock, and R. van Zwol (2009). "Placing Flickr Photos on a Map". In: *SIGIR*. Ed. by J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, pp. 484–491. ISBN: 978-1-60558-483-6.

Shen, D., J. L. Leidner, A. Merkel, and D. Klakow (2006). "The Alyssa System at TREC 2006: A Statistically-Inspired Question Answering System". In: *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Singh, S. K. and D. Rafiei (2016). "Geotagging Flickr Photos And Videos Using Language Models". In: *MediaEval*. Vol. 1739. CEUR Workshop Proceedings. CEUR-WS.org.

Skovsgaard, A., D. idlauskas, and C. S. Jensen (2014). "A Clustering Approach to the Discovery of Points of Interest from Geo-Tagged Microblog Posts". In: *2014 IEEE 15th International Conference on Mobile Data Management*. Vol. 1, pp. 178–188. DOI: 10.1109/MDM.2014.28.

Smith, D. A. and G. Crane (2001). "Disambiguating Geographic Names in a Historical Digital Library." In: *ECDL*. Ed. by P. Constantopoulos and I. Sølvberg. Vol. 2163. Lecture Notes in Computer Science. Springer, pp. 127–136. ISBN: 3-540-42537-3.

Smucker, M. D., J. Allan, and B. Carterette (2007). "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation". In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. CIKM '07. Lisbon, Portugal: ACM, pp. 623–632. ISBN: 978-1-59593-803-9. DOI: 10.1145/1321440.1321528. URL: http://doi.acm.org/10.1145/1321440.1321528.

Solorio, T., M. Pérez-Coutiño, M. Montes-y-Gómez, L. Villaseñor-Pineda, and A. López-López (2004). "A Language Independent Method for Question Classification". In: *COLING-2004*, pp. 1374–1380.

Soriano, J. M. G., M. Montes-y-Gómez, E. S. Arnal, and P. Rosso (2005). "A Passage Retrieval System for Multilingual Question Answering." In: *TSD*. Ed. by V. Matousek,

P. Mautner, and T. Pavelka. Vol. 3658. Lecture Notes in Computer Science. Springer, pp. 443–450. ISBN: 3-540-28789-2.

Soubbotin, M. M. (2001). “Patterns of Potential Answer Expressions as Clues to the Right Answers.” In: *Proceedings of the Text Retrieval Conference TREC-2001.*

Sparck-Jones, K. (1972). “A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL”. In: *Journal of Documentation* 28.1, pp. 11–21. DOI: `10.1108/eb026526`. eprint: `http://dx.doi.org/10.1108/eb026526`. URL: `http://dx.doi.org/10.1108/eb026526`.

Speriosu, M., T. Brown, T. Moon, J. Baldridge, and K. Erk (2010). “Connecting Language and Geography with Region-Topic Models”. In:

Speriosu, M. and J. Baldridge (2013). “Text-Driven Toponym Resolution using Indirect Supervision”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.* Sofia, Bulgaria: Association for Computational Linguistics, pp. 1466–1476. URL: `http://www.aclweb.org/anthology/P13-1144`.

Spitz, A., J. Geiß, and M. Gertz (2016). “So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks”. In: *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data.* ACM, p. 2.

Steinbach, M., G. Karypis, and V. Kumar (2000). “A Comparison of Document Clustering Techniques”. In: *KDD-2000 Workshop on Text Mining, August 20.* Ed. by M. Grobelnik, D. Mladenic, and N. Milic-Frayling. Boston, MA, pp. 109–111. URL: `http://www-users.cs.umn.edu/~karypis/publications/ir.html`.

Stokes, N., Y. Li, A. Moffat, and J. Rong (2008). “An Empirical Study of the Effects of NLP Components on Geographic IR performance”. In: *International Journal of Geographical Information Science* 22.3, pp. 247–264.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). “Yago: A Core of Semantic Knowledge”. In: *Proceedings of the 16th International Conference on World Wide Web.* WWW ’07. Banff, Alberta, Canada: ACM, pp. 697–706. ISBN: 978-1-59593-654-7. DOI: `10.1145/1242572.1242667`. URL: `http://doi.acm.org/10.1145/1242572.1242667`.

Sundheim, B. (1995a). “MUC6 named entity task definition, Version 2.1.” In: *Proceedings of the Sixth Message Understanding Conference (MUC6).* Columbia, MD, USA: Morgan Kaufmann.

Sundheim, B. (1995b). “Overview of results of the MUC-6 evaluation.” In: *Proceedings of the Sixth Message Understanding Conference (MUC6).* Columbia, MD, USA: Morgan Kaufmann.

Suzuki, J., H. Taira, Y. Sasaki, and E. Maeda (2003). “Question classification using HDAG kernel”. In: *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering.* Morristown, NJ, USA: Association for Computational Linguistics, pp. 61–68.

Teitler, B. E., M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling (2008). “NewsStand: A New View on News”. In: *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* GIS ’08. Irvine, California: ACM, 18:1–18:10. ISBN: 978-1-60558-323-5. DOI: `10.1145/1463434.1463458`. URL: `http://doi.acm.org/10.1145/1463434.1463458`.

Thomee, B., D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li (2015). “The New Data and New Challenges in Multimedia Research”. In: *ArXiv e-prints.* arXiv: `1503.01817 [cs.MM]`.

Tjong Kim Sang, E. F. (2002). “Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition”. In: *Proceedings of CoNLL-2002.* Taipei, Taiwan, pp. 155–158.

Tjong Kim Sang, E. and F. De Meulder (2003). “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition”. In: *Proceedings of CoNLL-2003.* Ed. by W. Daelemans and M. Osborne. Edmonton, Canada, pp. 142–147.

Toral, A., O. Ferrández, E. Noguera, Z. Kozareva, A. Montoyo, and R. Muñoz (2006). “Geographic IR Helped by Structured Geospatial Knowledge Resources.” In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Toral, A., O. Ferrández, E. Noguera, Z. Kozareva, A. Montoyo, and R. Muñoz (2007). “GIR with Geographic Query Expansion”. English. In: *Evaluation of Multilingual and Multi-modal Information Retrieval.* Ed. by C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber. Vol. 4730. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 889–892. ISBN: 978-3-540-74998-1. DOI: `10.1007/978-3-540-74999-8112`. URL: `http://dx.doi.org/10.1007/978-3-540-74999-8112`.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). “Feature-rich part-of-speech tagging with a cyclic dependency network”. In: *NAACL ’03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* Edmonton, Canada: Association for Computational Linguistics, pp. 173–180.

Trevisiol, M., J. Delhumeau, H. Jégou, and G. Gravier (2012). “How INRIA/IRISA identifies Geographic Location of a Video”. In: *Working Notes Proceedings of the MediaEval 2012 Workshop.* Italy. URL: `https://hal.archives-ouvertes.fr/hal-00757453`.

Urbano, J., M. Marrero, and D. Martín (2013). “On the measurement of test collection reliability”. In: *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR ’13, Dublin, Ireland - July 28 - August 01, 2013.* Ed. by G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai. ACM, pp. 393–402. DOI: `10.1145/2484028.2484038`. URL: `http://doi.acm.org/10.1145/2484028.2484038`.

Van Laere, O. (2013). “Georeferencing text using social media”. PhD thesis. Ghent University.

Van Laere, O., S. Schockaert, and B. Dhoedt (2010a). “Ghent University at the 2010 Placing Task”. In: *Working Notes of the MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy.* Pisa, Italy.

Van Laere, O., S. Schockaert, and B. Dhoedt (2010b). “Towards Automated Georeferencing of Flickr Photos”. In: *Proceedings of the GIR’10 Workshop.* Zurich, Switzerland.

Van Laere, O., S. Schockaert, and B. Dhoedt (2011a). “Finding Locations of Flickr Resources Using Language Models and Similarity Search”. In: *Proceedings of the 1st International Conference on Multimedia Retrieval, ICMR 2011, Trento, Italy, April 18 - 20, 2011.* Ed. by F. G. B. D. Natale, A. D. Bimbo, A. Hanjalic, B. S. Manjunath, and S. Satoh. ACM, p. 48. DOI: `10.1145/1991996.1992044`. URL: `http://doi.acm.org/10.1145/1991996.1992044`.

Van Laere, O., S. Schockaert, and B. Dhoedt (2011b). “Ghent University at the 2011 Placing Task”. In: *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011.* Ed. by M. Larson, A. Rae, C. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones. Vol. 807.

CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-807/vanLaereUGENTPlacingme11wn.pdf.

Van Laere, O., S. Schockaert, and B. Dhoedt (2013). "Georeferencing Flickr Resources Based on Textual Meta-Data". In: *Information Sciences* 238, pp. 52–74. ISSN: 0020-0255. DOI: http://dx.doi.org/10.1016/j.ins.2013.02.045. URL: http://www.sciencedirect.com/science/article/pii/S002002551300162X.

Van Laere, O., S. Schockaert, J. A. Quinn, F. C. Langbein, and B. Dhoedt (2012). "Ghent and Cardiff University at the 2012 Placing Task". In: *Working Notes Proceedings of the MediaEval 2012 Workshop, Santa Croce in Fossabanda, Pisa, Italy, October 4-5, 2012*. Ed. by M. A. Larson, S. Schmiedeke, P. Kelm, A. Rae, V. Mezaris, T. Piatrik, M. Soleymani, F. Metze, and G. J. F. Jones. Vol. 927. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-927/mediaeval2012submission35.pdf.

Van Laere, O., S. Schockaert, V. Tanasescu, B. Dhoedt, and C. Jones (2014). "Georeferencing Wikipedia Documents Using Data from Social Media Sources". In: *ACM Trans. Inf. Syst.* 32.3, 12:1–12:32. ISSN: 1046-8188. DOI: 10.1145/2629685. URL: http://doi.acm.org/10.1145/2629685.

Vicedo, J. L. (2002). "SEMQA: Un Modelo Semántico Aplicado a los Sistemas de Búsqueda de la Respuesta." PhD thesis. Universidad de Alicante.

Villatoro-Tello, E. (2010). *Ordenamiento Basado en Ejemplos para la Recuperación de Información Geográfica*. Computer Science Department, National Institute of Astrophysics, Optics and Electronics (INAOE) Puebla, México: Ph.D. thesis.

Villatoro-Tello, E., R. O. C. García, M. Montes-y-Gómez, L. V. Pineda, and L. E. Sucar (2010). "A Probabilistic Method for Ranking Refinement in Geographic Information Retrieval". In: *Procesamiento del Lenguaje Natural* 44. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/291.

Villatoro-Tello, E., M. Montes-y-Gómez, and L. Villaseñor-Pineda (2008). "INAOE at GeoCLEF 2008: A Ranking Approach based on Sample Documents". In: *Working Notes for the CLEF 2008 Workshop*. Aarhus, Denmark.

Volz, R., J. Kleb, and W. Mueller (2007). "Towards Ontology-based Disambiguation of Geographical Identifiers". In: *Proceedings of the 6th International World Wide Web Conference (WWW2007)*. Banff, Alberta,Canada.

Voorhees, E. M. (2003). "Overview of the TREC 2003 Question Answering Track." In: *Proceedings of the Text Retrieval Conference (TREC-2003)*, pp. 54–68.

Voorhees, E. M. and D. M. Tice (1999). "The TREC-8 Question Answering Track Evaluation." In: *TREC*.

Waldinger, R., P. Jarvis, and J. Dungan (2003). "Pointing to places in a deductive geospatial theory". In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*. Association for Computational Linguistics, pp. 10–17.

Wang, C., J. Wang, X. Xie, and W.-Y. Ma (2007). "Mining Geographic Knowledge Using Location Aware Topic Model". In: *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*. GIR '07. Lisbon, Portugal: ACM, pp. 65–70. ISBN: 978-1-59593-828-2. DOI: 10.1145/1316948.1316967. URL: http://doi.acm.org/10.1145/1316948.1316967.

Wang, R. and G. Neumann (2008). "Ontology-based Query Construction for GeoCLEF". In: *Working Notes for the CLEF 2008 Workshop*. Aarhus, Denmark.

Wang, R. and G. Neumann (2009). "Ontology-Based Query Construction for GeoCLEF". English. In: *Evaluating Systems for Multilingual and Multimodal Information Access.* Ed. by C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. Jones, M. Kurimo, T. Mandl, A. Peñas, and V. Petras. Vol. 5706. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 880–884. ISBN: 978-3-642-04446-5. DOI: `10.1007/978-3-642-04447-2116`. URL: `http://dx.doi.org/10.1007/978-3-642-04447-2116`.

Wang, X., Y. Zhang, M. Chen, X. Lin, H. Yu, and Y. Liu (2010). "An evidence-based approach for Toponym Disambiguation". In: *The 18th International Conference on Geoinformatics: GIScience in Change, Geoinformatics 2010, Peking University, Beijing, China, June, 18-20, 2010*, pp. 1–7. DOI: `10.1109/GEOINFORMATICS.2010.5567805`. URL: `https://doi.org/10.1109/GEOINFORMATICS.2010.5567805`.

Whittaker, E., J. Novak, P. Chatain, and S. Furui (2006). "TREC 2006 Question Answering Experiments at Tokyo Institute of Technology (Draft)". In: *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook".* Gaithersburg, MD, USA: NIST.

Wing, B. P. and J. Baldridge (2011). "Simple Supervised Document Geolocation with Geodesic Grids". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1.* HLT '11. Portland, Oregon: Association for Computational Linguistics, pp. 955–964. ISBN: 978-1-932432-87-9. URL: `http://dl.acm.org/citation.cfm?id=2002472.2002593`.

Wing, B. and J. Baldridge (2011). "Simple supervised document geolocation with geodesic grids". In: *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Vol. 1, pp. 955–964. ISBN: 9781932432879.

Wing, B. and J. Baldridge (2014). "Hierarchical Discriminative Classification for Text-Based Geolocation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pp. 336–348.

Woodruff, A. G. and C. Plaunt (1994). "GIPSY: Automated geographic indexing of text documents". In: *Journal of the American Society for Information Science* 45.9, pp. 645–655. ISSN: 1097-4571. DOI: `10.1002/(SICI)1097-4571(199410)45:9<645::AID-ASI2>3.0.CO;2-8`. URL: `http://dx.doi.org/10.1002/(SICI)1097-4571(199410)45:9%3C645::AID-ASI2%3E3.0.CO;2-8`.

Yi Li, N. S., L. Cavedon, and A. Moffat (2006). "NICTA I2D2 Group at GeoCLEF 2006." In: *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006.* Ed. by A. Nardi, C. Peters, and J. L. Vicedo. Alicante, Spain. ISBN: 2-912335-23-x.

Younis, E. M., C. B. Jones, V. Tanasescu, and A. I. Abdelmoty (2012). "Hybrid geo-spatial query methods on the Semantic Web with a spatially-enhanced index of DBpedia". In: *International Conference on Geographic Information Science.* Springer, pp. 340–353.

Zaila, Y. L. and D. Montesi (2015). "Geographic Information Extraction, Disambiguation and Ranking Techniques". In: *Proceedings of the 9th Workshop on Geographic Information Retrieval.* GIR '15. Paris, France: ACM, 11:1–11:7. ISBN: 978-1-4503-3937-7.

Zelle, J. and R. J. Mooney (1996). "Learning to Parse Database Queries Using Inductive Logic Programming." In: *AAAI/IAAI, Vol. 2*, pp. 1050–1055.

Zhang, D. and W. S. Lee (2003). "Question classification using support vector machines". In: *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* Toronto, Canada: ACM Press, pp. 26–32. ISBN: 1-58113-646-3. DOI: `http://doi.acm.org/10.1145/860435.860443`.

Zheng, Y.-T., Z.-J. Zha, and T.-S. Chua (2012). “Mining Travel Patterns from Geotagged Photos”. In: *ACM Trans. Intell. Syst. Technol.* 3.3, 56:1–56:18. ISSN: 2157-6904. DOI: 10.1145/2168752.2168770. URL: http://doi.acm.org/10.1145/2168752.2168770.