UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Ph.D. Thesis Dissertation

# Introducing Linguistic Knowledge into Statistical Machine Translation


**Author:** Adrià de Gispert Ramis

**Advisor:** Prof. José B. Mariño Acebal

TALP Research Center, Speech Processing Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya



Barcelona, October 2006

*Time flies like an arrow. Fruit flies like a banana.*
*Groucho Marx.*

# Abstract

This Ph.D. thesis dissertation addresses the use of morphosyntactic information in order to improve the performance of Statistical Machine Translation (SMT) systems, providing them with additional linguistic information beyond the surface level of words from parallel corpora.

The statistical machine translation system in this work here follows a tuple-based approach, modelling joint-probability translation models via log-linear combination of bilingual $n$-grams with additional feature functions. A detailed study of the approach is conducted. This includes its initial development from a speech-oriented Finite-State Transducer architecture implementing $X$-grams towards a large-vocabulary text-oriented $n$-grams implementation, training and decoding particularities, portability across language pairs and tasks, and main difficulties as revealed in error analyses.

The use of linguistic knowledge to improve word alignment quality is also studied. A cooccurrence-based one-to-one word alignment algorithm is extended with verb form classification with successful results. Additionally, we evaluate the impact in word alignment and translation quality of Part-Of-Speech, base form, verb form classification and stemming on state-of-art word alignment tools.

Furthermore, the thesis proposes a translation model tackling verb form generation through an additional verb instance model, reporting experiments in English→Spanish tasks. Disagreement is addressed via incorporating a target Part-Of-Speech language model. Finally, we study the impact of morphology derivation on Ngram-based SMT formulation, empirically evaluating the quality gain that is to be gained via morphology reduction.

# Resum

Aquesta tesi està dedicada a l'estudi de la utilització de informació morfosintàctica en el marc dels sistemes de traducció estocàstica, amb l'objectiu de millorar-ne la qualitat a través de la incorporació de informació lingüística més enllà del nivell simbòlic superficial de les paraules.

El sistema de traducció estocàstica utilitzat en aquest treball segueix un enfocament basat en tuples, unitats bilingües que permeten estimar un model de traducció de probabilitat conjunta per mitjà de la combinació, dins un entorn log-linial, de cadenes d'$n$-grames i funcions característiques addicionals. Es presenta un estudi detallat d'aquesta aproximació, que inclou la seva transformació des d'una implementació d'$X$-grames en autòmats d'estats finits, més orientada a la traducció de veu, cap a l'actual solució d'$n$-grames orientada a la traducció de text de gran vocabulari. La tesi estudia també les fases d'entrenament i decodificació, així com el rendiment per a diferents tasques (variant el tamany dels corpora o el parell d'idiomes) i els principals problemes reflectits en les anàlisis d'error.

La tesis també investiga la incorporació de informació lingüística específicament en aliniament per paraules. Es proposa l'extensió mitjançant classificació de formes verbals d'un algorisme d'aliniament paraula a paraula basat en co-ocurrències, amb resultats positius. Així mateix, s'avalua de forma empírica l'impacte en qualitat d'aliniament i de traducció que s'obté mitjançant l'etiquetatge morfològic, la lematització, la classificació de formes verbals i el truncament o *stemming* del text paral·lel.

Pel que fa al model de traducció, es proposa un model de tractament de les formes verbals per mitjà d'un model de instanciació addicional, i es realitzen experiments en la direcció anglès→castellà. La tesi també introdueix un model de llenguatge d'etiquetes morfològiques del destí per tal d'abordar problemes de concordança. Finalment, s'estudia l'impacte de la derivació morfològica en la formulació de la traducció estocàstica mitjançant $n$-grames, avaluant empíricament el possible guany derivat d'estratègies de reducció morfològica.

# Agraïments

Vull donar les gràcies de tot cor al Pepe, no només per haver-me ajudat a tirar endavant aquesta tesi, sinó per mostrar-se sempre com una persona exemplar dins i fora de l'entorn professional, i de la qual he après molts valors, i moltes actituds i maneres de treballar admirables. Sempre li estaré molt agraït per tot el que ha fet per mi.

Gràcies també als companys i companyes del grup de traducció estadística, i en especial al Josep Maria Crego, sense l'aportació del qual aquesta tesi no hauria anat tan lluny, i amb qui he après a fer recerca de manera seriosa.

També vull donar les gràcies a aquells i aquelles que, amb el seu treball i potser sense adonar-se'n, han estat i sempre seran un referent de compromís i motivació per a mi, com ara el Climent Nadeu, el Jaume Padrell, el Xavi Pérez (el company de feina que tothom desitjaria tenir), el Lluís Padró i el Lluís Màrquez.

Gràcies també als molts companys de tesi amb qui he compartit molt bones estones com l'Alberto, el Ramon, el Pere, l'Ignasi, la Marta, el Javi o el Joel, entre molts d'altres. Moltes gràcies també als grans amics i amigues que m'han acompanyat en aquest trajecte, no només pel que m'han aguantat quan ha fet falta, sinó per fer-me veure clarament que sempre hi tornarien a ser si calgués (Jordi, Mari, Oriol, Núria, Paqui, Ferran, Pablo, Ana, Lluís, David, etc.).

Per últim, vull dedicar aquesta tesi als meus pares i a l'Aleyda, que amb el seu amor constant em donen la força necessària per treballar i viure amb il·lusió. Aquesta tesi és, en bona part, també vostra.

Adrià

Barcelona, octubre del 2006

# Contents

# Chapter 1

# Introduction

The information society we live in is undoubtedly a globalised and multilingual one. Every day, hundreds of thousands of documents are being generated, and in many cases one or several translations for them are needed in order to cover the linguistic variety of the target population. The majority of work carried out by professional translators is related to non-literary documents (technical reports, legal and financial documents, user manuals, political debates, meeting minutes, and so on), where translation tends to be mechanical and domain-specific. However, the high translation cost in terms of money and time is a bottleneck that prevents all information from being easily spread across languages.

Apart from that, the growth and popularity rise of internet has given users access to practically any written, visual and audio material from anywhere in the world. Still, the language barrier is the only obstacle for this vast information to be fully shared by all users.

In this context, automatic or machine translation (MT) services are becoming more and more attractive. Several companies are already using and offering automatic translation software, and thousands of users are automatically translating web content on a daily basis, even though translation performance is still far from perfection. Additionally, many research efforts are being focused on speech-to-speech machine translation, and the seemingly unreachable goal of automatically translating spoken language is nearer than ever before.

To a large extent, much of the optimism being shared in the MT research community nowadays has been caused by the revival of statistical approaches to machine translation, or in other words, the birth of purely Statistical Machine Translation (SMT). In contrast to previous approaches based on linguistic knowledge representation, SMT is based on large amounts of human-translated example sentences (parallel corpora) in order to estimate a set of statistical models describing the translation process.

## 1.1 Machine Translation and the Statistical Approach

### 1.1.1 A brief history of MT

The beginnings of **statistical** machine translation (SMT) can be traced back to the early fifties, closely related to the ideas from which information theory arose [Sha49b] and inspired by works on cryptography [Sha49a, Sha51] during World War II. According to this view, machine translation was conceived as the problem of finding a sentence by decoding a given "encrypted" version of it [Wea55].

At that time, machine translation was seen as a quite simple and feasible task, basically consisting of automatically reading dictionary entries in order to translate the input sentence into a hypothetical universal language, from which the target sentence could be generated. A first public Russian–English system was presented at the University of Georgetown in 1954, and despite its very restricted domain (with a vocabulary size of around 250 words), the promising prospects of rapid improvement, and undoubtedly the cold war political context, led the United States government to make strong investments in emergent machine translation technologies.

Since then, many research projects were devoted to MT during the late 1950s. However, as the complexity of the linguistic phenomena involved in the translation process together with the computational limitations of the time were made apparent, enthusiasm faded out quickly.

Despite much research effort, by the turn of the decade results had fallen short of all expectations. To crown it all, two negative reports had a dramatic impact on MT research. On the one hand, the Bar-Hillel report [BH60] concluded that Fully Automatic High-Quality Translation was an unreachable goal and that research efforts should be focused on less-ambitious tasks, such as Computer-assisted Machine Translations tools. On the other hand, the controversial 1966 ALPAC report concluded that machine translation was of poor quality and twice as expensive as human translation, effectually causing all MT research to vanish.

During the 1970s, the focus of MT activity switched from the United States to Canada and to Europe, especially due to the growing demands for translations within their multicultural societies. Météo, a fully-automatic system translating weather forecasts had a great success in Canada, and meanwhile, the European Commission installed a French–English MT system called Systran. Other reseach projects, such as Eurotra, Ariane and Susy, broadened the scope of MT objectives and techniques, and rule-based approaches emerged as the right way towards successful MT quality. Throughout the 1980s many different types of MT systems appeared [Hut86], the most prevalent being those using an intermediate semantic language such as the Interlingua approach (more detailed in §1.1.2).

In the early 1990s, the progress made by the application of statistical methods to speech

recognition inspired the introduction by IBM reseachers of purely-statistical machine translation models. The drastic increment in computational power and the increasing availability of written translated texts allowed the development of statistical and other corpus-based MT approaches. Many academic tools turned into useful commercial translation products [Arn95], and several translation engines were quickly offered in the world wide web.

Today, while commercial MT systems are not error-free, their use is widespread and there is a growing demand for high-quality automatic translation. Regarding research, basically all the research community has moved towards corpus-based techniques, which have systematically outperformed traditional knowledge-based techniques in most performance comparisons. Every year more research groups embark on SMT experimentation, and a regained optimism as regards to future progress seems to be shared among the community.

### 1.1.2 Approaches to MT

Several criteria can be used to classify machine translation approaches [GV03], yet the most popular classification is done attending to the level of linguistic analysis (and generation) required by the system to produce translations. Usually, this can be graphically expressed by the machine translation pyramid in Figure 1.1.

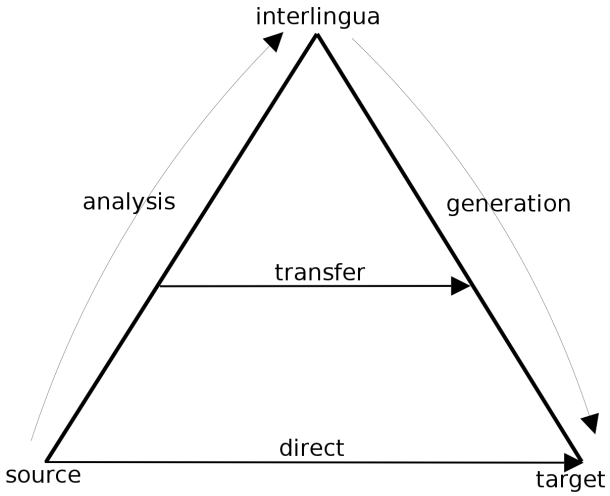



**Figure 1.1:** Machine Translation pyramid

Generally speaking, the bottom of the pyramid represents those systems which do not perform any kind of linguistic analysis of the source sentence in order to produce a target sentence. Moving upwards, the systems which carry out some analysis (usually by means of morphosyntax-based rules) are to be found. Finally, on top of the pyramid a semantic analysis of the source sentence turns the translation task into generating a target sentence according to the obtained semantic representation.

Aiming at a bird's-eye survey rather than a complete review, next each of these approaches is briefly discussed, before delving into the statistical approach to machine translation.

### Direct translation

This approach solves translation on a word-by-word basis, and it was followed by the early MT systems, which included a very shallow morphosyntactic analysis. Today, this preliminary approach has been abandoned, even in the framework of corpus-based approaches (see below).

### Transfer-based translation

The rationale behind the transfer-based approach is that, once we grammatically analyse a given sentence, we can pass this grammar on to the grammatical representation of this sentence in another language. In order to do so, rules to convert source text into some structure, rules to transfer the source structure into a target structure, and rules to generate target text from it are needed. Lexical rules need to be introduced as well.

Usually, rules are collected manually, thus involving a great deal of expert human labour and knwoledge of comparative grammar of the language pair. Apart from that, when several competing rules can be applied, it is difficult for the systems to prioritise them, as there is no natural way of weighing them.

This approach was massively followed in the 1980s, and despite much research effort, high-quality MT was only achieved for limited domains [Hut92].

### Interlingua-based translation

This approach advocates for the deepest analysis of the source sentence, reaching a language of semantic representation named Interlingua. This conceptual language, which needs to be developed, has the advantage that, once the source meaning is captured by it, *in theory* we can express it in any number of target languages, so long as a generation engine for each of them exists.

Though conceptually appealing, several drawbacks make this approach unpractical. On the one hand, the difficulty of creating a conceptual language capable of bearing the particular semantics of all languages is an enormous task, which in fact has only been achieved in very limited domains. Apart from that, the requirement that the whole input sentence needs to be *understood* before proceeding onto translating it, has proved to make these engines less robust to the grammatical *incorrectness* of informal language, or which can be produced by an automatic speech recognition system.

#### Corpus-based approaches

In contrast to the previous approaches, these systems extract the information needed to generate translations from parallel corpora that include many sentences which have already been translated by human translators. The advantage is that, once the required techniques have been developed for a given language pair, *in theory* it should be relatively simple to transpose them to another language pair, so long as sufficient parallel training data is available.

Among the many corpus-based approaches that sprung at the beginning of the 1990s, the most relevant ones are example-based (EBMT) and statistical (SMT), although the differences between them are constantly under debate. Example-based MT makes use of parallel corpora to extract a database of translation examples, which are compared to the input sentence in order to translate. By choosing and combining these examples in an appropriate way, a translation of the input sentence can be provided.

In SMT, this process is accomplished by focusing on purely statistical parameters and a set of translation and language models, among other data-driven features. Although this approach initially worked on a word-to-word basis and could therefore be classified as a direct method, nowadays several engines attempt to include a certain degree of linguistic analysis into the SMT approach, slightly climbing up the aforementioned MT pyramid.

The following section further introduces the statistical approach to machine translation.

### 1.1.3 Statistical Machine Translation

Since its revival more than a decade ago when IBM researchers presented the *Candide* SMT system [Bro90, Bro93], the statistical approach to machine translation has seen an increasing interest among both natural language and speech processing research communities. Mainly, three factors account for this increasing interest:

- There is a growing availability of parallel texts (though this applies, in general, only to major languages in terms of presence in internet), coupled with increasing computational power. This enables research on statistical models which, in spite of their huge number of parameters (or probabilities) to estimate, are *sufficiently* represented in the data.
- The statistical methods are more *robust* to speech disfluencies or grammatical faults. As no deep analysis of the source sentence is done, these systems seek the most probable translation hypothesis given a source sentence, assuming the input sentence is correct.
- And last but not least, shortly after their introduction, these methods proved at least as good or even better as rule-based approaches in various evaluation campaigns[1]. A clear

[1]See NIST annual evaluation results at http://www.nist.gov/speech/tests/mt

example is the German project VerbMobil, which concluded that preliminary statistical approaches outperformed other approaches, on which research had been focused for many years [Wah00].

At the turn of the 21st century, apart from VerbMobil[2], other projects and consortiums (C-STAR[3], LC-STAR, FAME, among others) involving many research centers have focused on SMT and its applications to text and speech translation tasks. Recently, the European project TC-STAR (Technology and Corpora for Speech to Speech Translation),[4] has among its main objectives to achieve significant performance improvements in the statistical machine translation approach.

## 1.2 Motivation

In the face of the promising horizon for Spoken Language Translation (SLT) and having a long-standing experience on Automatic Speech Recognition (ASR), the Speech Processing Group of the Universitat Politècnica de Catalunya (UPC) set among its objectives to undertake reseach on SMT in 2001.

Having to initiate this work from scratch, a first and necessary step was to develop a set of tools implementing a statistical translation model with state-of-the-art performance. The implemention with a Finite-State Transducer of the joint probability model by adapting speech recognition tools, set the foundations of this Ph.D. research work, obtaining preliminary text and speech translation results with small hand-crafted parallel training data.

Since 2003, the work of mantaining the system up-to-date with the rapid changes in the field has been shared by the author with a growing team of outstanding Ph.D. researchers and professors. Today, the initial translation engine has turned into an Ngram-based SMT system capable of dealing with large amounts of material, steadily participating in worldwide evaluation campaigns and successfully achieving state-of-the-art results.

Internationally, in this five-year period statistical machine translation has evolved from a newborn speech recognition task being looked at as unrealistic, into a prominent task bringing together the natural language processing community and the speech recognition community towards spoken language translation. Nowadays, the number of publications on SMT-related issues, as well as the number of conference sessions, workshops and shared task evaluations, is progressively augmenting.

Although Statistical Machine Translation has seen a big progress over the last years, mainly

[2] http://verbmobil.dfki.de/overview-us.html
[3] http://www.c-star.org
[4] http://www.tc-star.org

thanks to the development of new statistical models, to the ever-growing availability of parallel corpora, to the steady increase in computational power and to the widespread use of automatic evaluation measures, it still suffers from basic morphology- and syntax-related translation errors as of today.

On the other hand, every year more linguistic resources are made available to the research community. Natural Language Processing (NLP) tools such as Part-of-Speech taggers, word segmentors, lemmatisers, named-entity detectors and recognisers, chunkers, shallow and full parsers and semantic onthologies (among other tools and resources) are being made available for a growing list of languages. Despite internally taking statistical decisions in most of the cases, these tools provide a wider and deeper linguistic knowledge of a language than the one SMT is currently capturing from parallel corpora.

Therefore, having as ultimate goal the challenging balance between parallel corpora and additional linguistic tools that will lead to flawless machine translation performance, this Ph.D. thesis also investigates models and techniques which incorporate this additional morphosyntactic information into the statistical framework of SMT.

The underlying assumption is that going beyond the surface level of words can help improve the translation system performance whenever empirical evidence from parallel corpora is insufficient. This undesired situation is a highly task-dependent one. The pair of languages involved, the parallel corpus size, the domain of the task and other factors strongly affect the incorporation of morphosyntactic information, and might even condition its need at all.

Due to corpora and linguistic tools availability reasons, most of this Ph.D. research work was carried out with the English and Spanish languages, and less significanty, with the Catalan and Arabic languages.

## 1.3 Objectives of this Ph.D.

The objectives of this Ph.D. thesis are the following:

- **To define a statistical translation model and to implement a tool estimating it, in order to use it as baseline system for text and speech translation purposes.** As very few tools related with SMT training or decoding were available at the time when this research work began, the goal was to adapt UPC speech recognition tools in order to implement a joint-probability model for text and speech translation. Eventually, this fundamental goal shifted towards the next objective.
- **To achieve and mantain state-of-the-art performance with this system through the many changes in the research field.**

For all research directed towards improving SMT performance to have impact and be relevant to the community, our SMT system must have comparable results to other systems from international sites. The system has undergone many changes, the most relevant being huge-data treatment capabilities, new more efficient decoding tool, feature implementation, optimisation tools, among others. This enormous and crucial task has been jointly addressed by UPC SMT researchers as a team, as it will be duely mentioned.

- **To study the impact of incorporating morphosyntactic information into the statistical machine translation system.**
  Throughout this research work, we have been interested in trying to overcome current limitations of SMT with the inclusion of relevant morphosyntactic information. As regards to the training or decoding stage where additional morphosyntactic information can be introduced, we can divide this study into three blocks:
  - *Word Alignment*, where information is used to better perform this first step of SMT training
  - *Unit segmentation*, where information is used to better define the borders between translation units
  - *Core Machine Translation*, where information is used to estimate alternative or additional feature models

## 1.4 Thesis Organisation

The Ph.D. thesis dissertation is divided in **seven** chapters. This introductory chapter is followed by an overview on the various statistical machine translation approaches that have been and are being applied in the field, with an emphasis on related works introducing morphosyntactic information. The next four chapters are devoted to the presentation of the thesis contributions. Final chapter concludes this work.

Outline of the thesis dissertation:

**Chapter 2** presents an overview of Statistical Machine Translation, reviewing the most widely-followed approaches since its introduction in the early 1990s until our days. In particular, we trace the evolution from word-based models (§2.1) towards phrase-based (§2.2) and tuple-based models (§2.3), which are log-linearly combined with other feature functions (§2.4). The chapter also introduces the Word Alignment task in §2.5, which is associated with the training of SMT systems, as well as the published works on introducing linguistic knowledge into the SMT framework in §2.6. To conclude, the difficulties of MT evaluation are raised, presenting and discussing the most commonly-used automatic evaluation measures in §2.7.

**Chapter 3** is dedicated to a detailed study of the joint-probability translation model and its main characteristics. §3.2 and §3.3 review its evolution from an X-grams implementation by means of a Finite-State Transducer towards an standard $N$-gram smoothed model for large-vocabulary tasks. Modeling issues are discussed and empirically tested, while real-life applications of the model are also highlighted. To conclude, §3.4 focuses on the definition of the translation unit (called *tuple*) and its particular properties.

**Chapter 4** extends on the previous chapter by merging the presented bilingual model into a log-linear combination of feature functions (§4.2). The contribution of each feature model is experimentally assessed in §4.3, which includes a manual error analysis of the system output for an English↔Spanish task. Finally, §4.4 summarises the achievements of the system in international evaluation campaigns emphasising its strong and weak points.

**Chapter 5** explores the benefits of using morphosyntactic information the first training stage of the $N$gram-based SMT system, that is, word alignment. Approaches followed include developing a complete cooccurrence-based aligner including many-to-many links (§5.2), constraining GIZA alignments with verb form classification (§5.3) and extending this approach with several linguistic word classification schemes (§5.4). This section also includes a study of the impact of alignment quality variations in final translation performance.

**Chapter 6** brings together the research work done on introducing linguistic information into statistical translation modelling and MT evaluation. In particular, a verb form classification framework is presented for the $N$gram-based SMT system (§6.2), reporting English→Spanish experiments. The inclusion of target a Part-Of-Speech language model to tackle disagreement problems is developed in §6.3. Finally, §6.4 is devoted to the study of the impact of morphology derivation in the framework of Ngram-based SMT.

**Chapter 7** draws the main conclusions from the Ph.D. thesis dissertation and details possible future lines of research.

## 1.5 Research Contributions

The main contributions of this Ph.D. thesis dissertation are:

- Full description of an $N$gram-based statistical machine translation system. We trace the evolution from an initial Finite-State Transducer implementation (of the joint-probability translation model via $X$-grams) towards a currently state-of-the-art system, discussing and empirically evaluating alternative design decisions. Results from different tasks are presented, comparing the adequacy of the approach for each language pair.

- Introduction of linguistic knowledge into statistical word alignment. Given the importance of word alignment as the first step in training SMT systems, we explore the benefits of including morphosyntactic information prior to word alignment, empirically studying the impact of morphology in word alignment quality. We also undergo a correlation study between alignment improvement and translation quality improvement.

- Use of verb classification strategies for statistical machine translation models. We propose a classification approach which is coupled with standard SMT decoding and report results for English→Spanish tasks, were positive results are obtained in small-data situations. Additionally, we investigate the effects of Spanish morphology in English→Spanish Ngram-based translation modelling, showing that even though it is positive to reduce Spanish Verb morphological information in order to estimate a bilingual $n$-gram model, impact is reduced and syntax-aware techniques must be incorporated in combination with morphology reduction.

The findings presented in this Ph.D. dissertation were published in a number of publications, which will be referred to in their respective sections.

# Chapter 2

# State of the art

This chapter traces an overview of the most prominent statistical machine translation approaches being followed from initial SMT systems to the current literature, and which are more relevant to our research work.

Firstly, the mathematical foundations of word-based SMT settled by IBM in the early 1990s are reviewed in §2.1. Secondly, the emergence of statistical approaches which no longer consider single words as their translation units (but some sort of word sequence) is discussed in §2.2 and §2.3. Then, §2.4 introduces the maximum entropy approach leading to the prevailing log-linear combination of feature models, which is providing state-of-the-art results as of today.

The notion of word alignment as the first crucial step during the training of any SMT system is introduced and discussed in §2.5, reviewing its definition as a stand-alone NLP task. Later on, §2.6 is devoted to review the most relevant works on using morphosyntactic information to improve statistical machine translation performance. Alternative SMT approaches based on or strongly relying on shallow or full parsing are also commented.

To conclude, §2.7 introduces the most widely used automatic evaluation measures, together with a discussion on their main advantages and drawbacks.

## 2.1 Word-based translation models

Statistical machine translation is based on the assumption that every sentence $e$ in a target language is a possible translation of a given sentence $f$ in a source language . The main difference between two possible translations of a given sentence is a probability assigned to each, which is to be learned from a bilingual text corpus. The first SMT models applied these probabilities to words, therefore considering words to be the translation units of the process.

### 2.1.1 IBM translation and alignment models

Supposing we want to translate a source sentence $f$ into a target sentence $e$, we can follow a noisy-channel approach (regarding the translation process as a channel which distorts the target sentence and outputs the source sentence) as introduced in [Bro90], defining statistical machine translation as the optimisation problem expressed by:

$$\hat{e} = \arg\max_{e} Pr(e \mid f) \tag{2.1}$$

Typically, Bayes rule is applied, obtaining the following expression:

$$\hat{e} = \arg\max_{e} Pr(f \mid e) \cdot Pr(e) \tag{2.2}$$

This way, translating $f$ becomes the problem of detecting which $e$ (among all possible target sentences) scores best given the product of two models: $Pr(e)$, the target *language model*, and $Pr(f \mid e)$, the *translation model*. Although it may seem less appropriate to estimate two models instead of just one (considering that $Pr(e \mid f)$ and $Pr(f \mid e)$ are equally difficult to estimate), the use of such a target language model justifies the application of Bayes rule, as this model helps penalise non-grammatical target sentences during the search.

Whereas the language model, tipically implemented using Ngrams, was already being used successfully in speech processing and other fields, the translation model was first presented by introducing a hidden variable $a$ to account for the alignment relationships between words in each language, as in equation 2.3.

$$Pr(f \mid e) = \sum_{a} Pr(f, a \mid e) = Pr(J \mid e) \prod_{j=1}^{J} Pr(a_j \mid f_1^{j-1}, a_1^{j-1}, e) \cdot Pr(f_j \mid f_1^{j-1}, a_1^{j}, e) \tag{2.3}$$

where $f_j$ stands for word in position $j$ of the source sentence $f$, $J$ is the length of this sentence (in number of words), and $a_j$ stands for the alignment of word $f_j$, ie. the position in the target sentence $e$ where the word which aligns to $f_j$ is placed.

The set of model parameters, or probabilities, is to be automatically learnt from parallel data. In order to train this huge amount of parameters, in [Bro93] the EM algorithm with incresingly complex models is used. These models are widely known as the five IBM models, and are inspired by the generative process described in Figure 2.1, which interprets the model decomposition of equation 2.3.

Conceptually, this process states that for each target word, we first find how many source words will be generated (following a model denoted as *fertility*); then, we find which source words

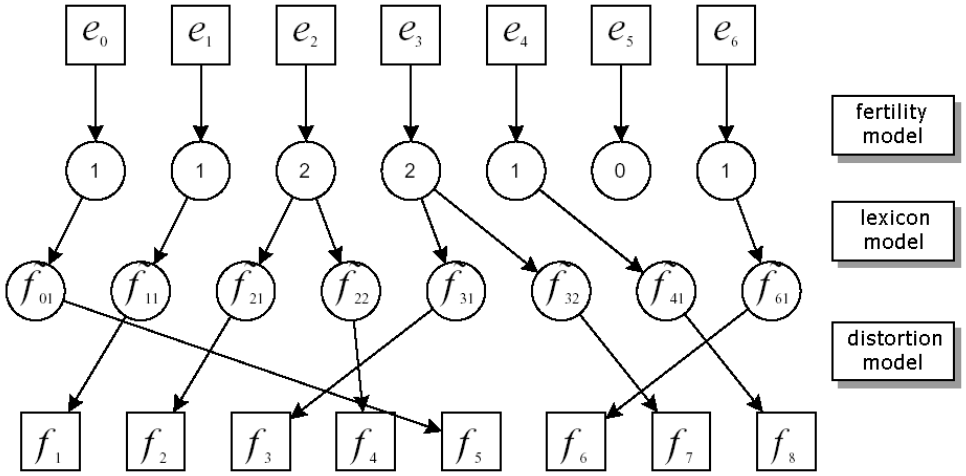



**Figure 2.1:** Illustration of the generative process underlying IBM models

are generated from each target word (lexicon or *word translation* probabilities); and finally, we reorder the source words (according to a *distortion* model) to obtain the source sentence[1].

These models are expressed as:

- $n(\phi|e)$ or Fertility model, which accounts for the probability that a target word $e_i$ generates $\phi_i$ words in the source sentence
- $t(f|e)$ or Lexicon model, representing the probability to produce a source word $f_j$ given a target word $e_i$
- $d(\pi|\tau, \phi, e)$ or Distorsion model, which models the probability of placing a source word in position $j$ given that the target word is placed in position $i$ in the target sentence (also used with inverted dependencies, and known as Alignment model)

IBM models 1 and 2 do not include fertility parameters so that the likelihood distributions are guaranteed to achieve a global maximum. Their difference is that Model 1 assigns a uniform distribution to alignment probabilities, whereas Model 2 introduces a zero-order dependency with the position in the source. [Vog96] presented a modification of Model 2 that introduced first-order dependencies in alignment probabilities, the so-called HMM alignment model, with successful results. Model 3 introduces fertility and Model 4 and 5 introduce more detailed dependencies in the alignment model to allow for jumps, so that all of them must be numerically approximated and not even a local maximum can be guaranteed.

[1]Note that the process generates from the target to the source language, due to the application of Bayes rule in equation 2.2.

A detailed description of IBM models and their estimation from a parallel corpus can be found in [Bro93]. In [Kni99] an informal yet clarifying tutorial on IBM models can be found.

#### Word Alignment

As explicitly introduced by IBM formulation as a model parameter, word alignment becomes a function from source positions $j$ to target positions $i$, so that $a(j) = i$. This definition implies that resultant alignment solutions will never contain many-to-many links, but only many-to-one[2], as only one function result is possible for a given source position $j$.

Although this limitation does not account for many real-life alignment relationships, in principle IBM models can solve this by estimating the probability of generating the source empty word, which can translate into non-empty target words.

However, as we will see in the following section, many current SMT systems do not use IBM model parameters in their training schemes, but only the *most probable* alignment (using a Viterbi search) given the estimated IBM models. Therefore, in order to obtain many-to-many word alignments, usually alignments from source-to-target and target-to-source are performed, and symmetrisation strategies have to be applied, as will be further discussed in §2.5.

### 2.1.2 Training and decoding tools

A stack decoder for IBM model 2 was presented in [Wan97], based on the A*-search algorithm. In 1999, the John Hopkins University summer workshop research team on SMT released GIZA (as part of the EGYPT toolkit), a tool implementing IBM models training from parallel corpora and best-alignment Viterbi search, as reported in [AO99], where a decoder for model 3 is also described. This was a breakthrough in that it enabled many other teams to join SMT research easily. In 2001 and 2003 improved versions of this tool were released, and named GIZA++[Och03c].

DP-based decoders both for model 2 and model 4 can be found in [Til00] and [Til03]. In [Ger01] the speed and quality of a stack-based, a greedy and an integer-programming decoder for IBM model 4 is compared. In [GV03] several decoders for IBM models are presented, ranging from greedy approaches to dynamic programming and stack solutions.

[2] By many-to-many links those relationships between more than one word in each language are refered, whereas many-to-one links associate more than one source word with a single target word. One-to-one links are defined analogously.

## 2.2 Phrase-based translation models

By the turn of the century it became clear that in many cases specifying translation models at the level of words turned out to be inappropriate, as much local context seemed to be lost during translation. Novel approaches needed to describe their models according to longer units, typically sequences of consecutive words (or *phrases*).

### 2.2.1 Alignment templates

The first approach using longer translation units was presented in [Och99b] and named Alignment Templates, which are pairs of generalised phrases that allow word classes and include an *internal* word alignment. This idea is expressed by Equation 2.4.

$$Pr(f_1^J|e_1^I) = \sum_{z_1^K, \tilde{a}_1^K} \prod_{k=1}^{K} p(\tilde{a}_k \mid \tilde{a}_{k-1}) \cdot p(z_k \mid \tilde{e}_k) \cdot p(\tilde{f}_k \mid z_k, \tilde{e}_k) \tag{2.4}$$

where $z_k$ is the $k$-th template used, the sequence of alignment templates $z_1^K$ and the alignments whithin the templates $\tilde{a}_1^K$ are hidden variables, and there are three probability distributions:

- the phrase alignment probability $p(\tilde{a}_k \mid \tilde{a}_{k-1})$, in the fashion of the word-based HMM model in[Vog96]
- the probability of applying an alignment template $p(z_k \mid \tilde{e}_k)$
- the phrase translation probability $p(\tilde{f}_k \mid z_k, \tilde{e}_k)$

Therefore, in this powerful approach the translation unit becomes a triple composed of: a source sequence of word classes, a target sequence of word classes, and a set of internal alignment links between word classes inside the borders of the template. Word classes from source and target language can be automatically estimated from monolingual or bilingual data as in [Och99a].

The generative process underlying the Alignment Template approach is drawn in Figure 2.2. As it can be seen, source words (each of them belonging to a word class) are grouped into phrases $\tilde{f}_k$, and for each phrase an alignment template is applied, originating a set of target phrases. Then, these phrases are ordered according to the phrase alignment model, and finally target words are produced.

More details on this approach can be found in [Och04b], where significant improvements over word-based approaches are reported.

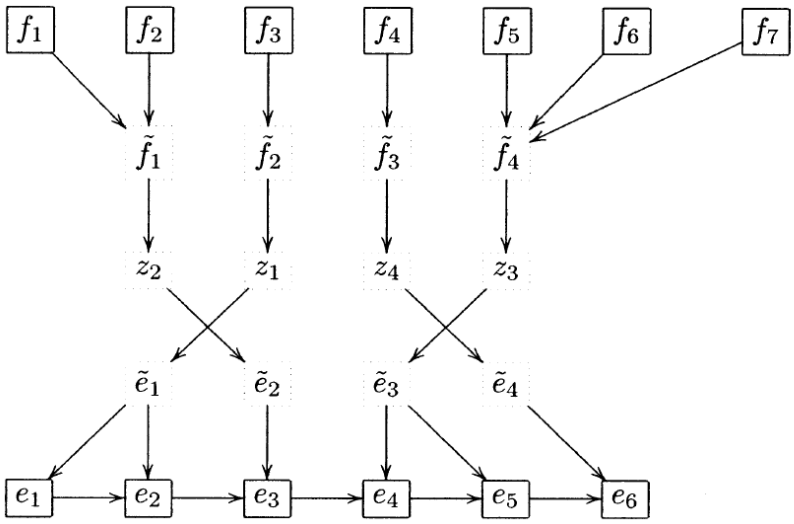



**Figure 2.2:** Illustration of the translation process of the Alignment Template approach

### 2.2.2 Phrase-based SMT

A simplified version of the previous approach is the so-called phrase-based statistical machine translation presented in [Zen02]. Under this framework, word classes are not used (but the actual words from the text instead), and the translation unit looses internal alignment information, turning into so-called *bilingual phrases*. Mathematically, former equation 2.4 gets simplified to:

$$Pr(f_1^J|e_1^I) = \alpha(e_1^I) \cdot \sum_B Pr(\tilde{f}_k \mid \tilde{e}_k) \tag{2.5}$$

where the hidden variable $B$ is the segmentation of the sentence pair in $K$ bilingual phrases $(\tilde{f}_1^K, \tilde{e}_1^K)$, and $\alpha(e_1^I)$ is assuming the same probability for all segmentations.

The phrase translation probabilities are usually estimated, over all bilingual phrases in the corpus, by relative frequency of the target sequence given the source sequence, as in:

$$Pr(\tilde{f}_k|\tilde{e}_k) = \frac{N(\tilde{f}_k, \tilde{e}_k)}{N(\tilde{e}_k)} \tag{2.6}$$

where bilingual phrases are defined as any pair of source and target phrases that have consecutive words and are consistent with the word alignment matrix. According to this criterion, any sequence of consecutive source words and consecutive target words which are aligned to each other and not aligned to any other token in the sentence, become a phrase. This is exemplified in Figure 2.3, where eight different phrases are extracted and it is worth noting that AB→WY is *not* extracted, given the definition constraint. For more details on this criterion, see [Och99b]

or [Zen02].

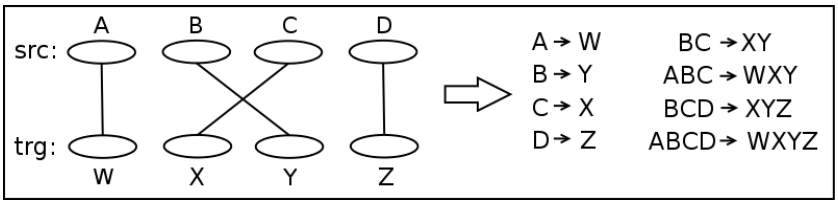



**Figure 2.3:** Phrase extraction from a certain word aligned pair of sentences.

In [Mar02] a joint-probability phrase-based model is introduced, which learns both word and phrase translation and alignment probabilities from a set of parallel sentences. However, this model is only tractable up to an equivalent of IBM model 3, due to severe computational limitations. Furthermore, when comparing this approach to the simple above-mentioned phrase generation from word alignments and a syntax-based phrase generation [Yam01], the approach from word alignment achieves best results as shown in [Koe03].

An alternative way to compute phrase translation probabilities is to use IBM model 1 lexical probabilities of the words inside the phrase pair, as presented in [Vog03]. A smoothed relative frequency is used in [Zen04].

Nowadays, many SMT systems follow a phrase-based approach, in that their translation unit is the bilingual phrase, such as [Lee06, Ber06, Mat06, Aru06, Kuh06, Kir06, Hew05], among many others. Most of these systems introduce a log-linear combination of models, as will be discussed in §2.4.

Relevantly, this phrase-based relative frequency model ignores IBM model parameters, being automatically estimated from a word-aligned parallel corpus, thus turning word alignment into a stand-alone training stage which can be done independently, as will be discussed in §2.5.

### 2.2.3 Training and decoding tools

Lately many tools are being implemented and released, so that every year it becomes easier for a beginner to get quickly introduced into phrase-based SMT, and even run preliminary experiments in one day. Without aiming at completeness, some of them are mentioned here.

Regarding phrase extraction and estimation, an open-source tool has been released in [Ort05]. As for decoding tools, in [Koe04] a freely-available beam search decoder for phrase-based translation models is described. A freely-available phrase-based and ngram-based decoder is described in [Cre05b]. A decoder based on confusion networks is presented in [Ber05], and two open-source decoders have been released in [Pat06, Olt06], programmed in C++ and Java respectively.

## 2.3 Tuple-based translation model

Without loss of generality, an alternative approach to SMT is to view translation as a stochastic process maximising the joint probability $p(f, e)$ instead of the conditional probability $p(f \mid e)$, leading to the following decomposition:

$$\hat{e}_1^I = \arg\max_{e_1^I}\{p(f_1^J, e_1^I)\} = \cdots = \tag{2.7}$$

$$\arg\max_{e_1^I}\{\prod_{n=1}^{N} p((f,e)_n|(f,e)_1, ..., (f,e)_{n-1})\} \tag{2.8}$$

where $(f, e)_n$ is the n-*th* bilingual unit, or *tuple*, of a given tuple sequence which generates monotonically both the training source and target sentences. Under this approach, the translation of a given source unit $(f, e)_n$ is conditioned by a previous bilingual context $(f, e)_1, ..., (f, e)_{n-1}$, which in practice must be limited in length.

### 2.3.1 Finite-State Transducer implementation

In [Vid97, Cas01] this joint-probability model is implemented by means of a Finite-State Transducer (FST), which can be automatically inferred from parallel data as in [Cas00]. Typically, this approach is followed in literature for speech translation. This is because, in contrast to initial word-based models and phrase-based approaches [3], the use of such a transducer allows for an elegant integration of the acoustic model $p(x|f)$ ($x$ being the input acoustic signal) into a global search, thus performing speech translation in one fell swoop [Vid97]. Mathematically, this is expressed in equation 2.9:

$$\hat{e}_1^I \simeq \arg\max_{e_1^I}\max_{f_1^J} p(f_1^J, e_1^I) \cdot p(x \mid f_1^J) \tag{2.9}$$

where $p(f, e)$ plays here the same role of the language model in speech recognition and $p(x \mid f)$ is the acoustic model.

A graphical representation of such a translation FST is shown in Figure 2.4, where each arc contains a bilingual unit comprised of one or more source words and zero, one or more target words, together with the associated probability estimated during training.

[3] Recently there is a growing and interesting research line towards integrating speech recognition and phrase-based text translation, mainly directed at outputting a word graph in recognition, and feeding it into the translation engine, which can combine its scores with some acoustic-related scores. However, falling out of the scope of this Ph.D. thesis, no further details on this issue will be given.

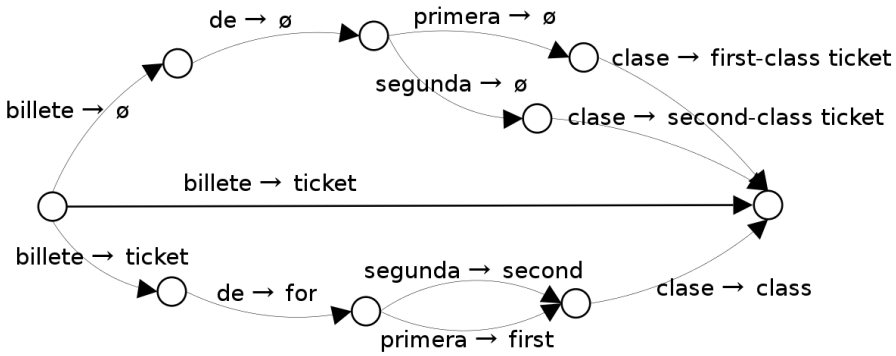



**Figure 2.4:** A translation FST from Spanish to English

Similarly to phrase-based approaches, these units are extracted in training from the Viterbi word alignment, therefore ignoring IBM model parameters. However, now the bilingual units are constrained by the sequentiality which is inherent to the N-gram decomposition of equation 2.8.

By adapting X-grams (which had been successfully used for language modelling in [Bon96] and speech recognition in [Bon98]) to this transducer architecture, this approach was followed in order to address the first main objective of this Ph.D. work, namely the implementation of a preliminary speech translation system. For this reason, all further details are addressed in detail in the following chapter.

### 2.3.2 Other implementations

Another implementation with cascaded finite-state transducers combining both statistically learnt transducers and hand-crafted rules can be found in [Vog00].

The FST-based approach is monotonous in that its model is based on the sequential order of tuples during training. Therefore, in principle it is more appropriate for pairs of languages with relatively similar word order schemes. However, an approach with reordered transducers can be found in [Ban01].

As we will see in chapter 3, due to structual inefficiency, the FST approach needs to be recast into an Ngram approach in order to deal with large quantities of data, as will be seen in detail in chapter 4. However, an efficient implementation of Weighted Finite-State Transducers for SMT is presented in [Kan04], which can incorporate word ordering capabilities as introduced in [Mat05].

## 2.4 Feature-based models combination

Another alternative to the noisy-channel approach is to directly model the posterior probability $Pr(e_1^I|f_1^J)$, a well-founded approach in the framework of maximum entropy, as shown in [Ber96]. By treating many different knowledge sources as feature functions, a log-linear combination of models can be performed, allowing an extension of a baseline translation system with the addition of new feature functions. In this case, the decision rule responds to the following expression:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \tag{2.10}$$

so that the noisy-channel approach can be obtained as a special case if we consider only two feature functions, namely the target language model $h_1(e_1^I, f_1^J) = \log p(e_1^I)$ and the translation model of the source sentence given the target $h_2(e_1^I, f_1^J) = \log p(f_1^J|e_1^I)$.

### 2.4.1 Minimum-error training

This approach, which was introduced in [Pap98] for a natural language understanding task, suggests that the training optimisation task becomes finding out the $\lambda_m$ which weight each model according to a certain criterion. In [Och02] minimum error training is introduced for statistical machine translation, stating that these weights need to be settled by directly minimising the translation error on a development set, as measured by a certain automatic measure (see §2.7).

Typically, this log-linear combination includes, apart from a translation model, other feature functions, such as:

- additional language models (word-based or class-based high-order Ngrams)
- sentence length models, also called word bonuses
- lexical models (such as IBM model 1 from source to target and from target to source)
- phrase penalties
- others (regarding information on manual lexicon entries or other grammatical features)

In order to optimise the $\lambda_m$ weights, the usual criterion is to use the maximum posterior probability $p(e|f)$ on a training corpus. Adequate algorithms for such a task are the GIS (Generalised Iterative Scaling) or the *downhill simplex* method [Nel65]. On the other hand, given a loss function based on automatic translation evaluation measures, a minimum bayes-risk

decoding scheme can also be used to tune a SMT system, as in [Kum04].

Nowadays, all SMT systems use a log-linear combination of feature models, optimised according to a certain automatic measure on the development data.

### 2.4.2 Re-ranking

In [She04] a discriminative reranking strategy is introduced for improving SMT performance (and also used in many systems, such as [Qua05]). This technique works as follows:

- First, a baseline system generates $n$-best candidate hypotheses
- Then, a set of features which can potentially discriminate between good and bad hypotheses are computed for each candidate
- Finally, these features are weighted in order to produce a new candidate ranking

The advantage is that, given the candidate sentence, features can be computed globally, enabling rapid experimentation with complex feature functions. This approach is followed in [Och03b] and [Och04a] to evaluate the benefits of a huge number of morphological and shallow-syntax feature functions to re-rank candidates from a standard phrase-based system, with little success. The introduction of linguistic information into SMT systems is later addressed in §2.6.

## 2.5 Statistical Word Alignment

Even though IBM word-based translation models include the alignment model as part of a whole translation scheme, this can also be defined as an independent Natural Language Processing task. In fact, most of current new generation translation models treat word alignment as an independent result from the translation model, as it was mentioned in §2.2.2 and §2.3.

The task of automatic word alignment focuses on detecting, given a parallel corpus, which tokens or sets of tokens from each language are connected together in a given translation context, revealing thus the relationship between these bilingual units. Among the many applications in natural language processing, such as bilingual dictionaries extraction or transfer rules learning, word alignment becomes particularly crucial in the context of statistical machine translation, where it represents an essential block in the learning process of current statistical translation models. In fact, it is reasonable to expect that a correct generation of word alignment will show a positive correlation with translation quality.

Actually, the relevance of word alignment has been corresponded by several previous works on the matter, including shared tasks in the frame of HLT-NAACL 2003 and ACL 2005 Workshops on Building and Using Parallel Texts [Mih03, Mar05]. Several competing systems were presented and evaluated against a manual reference.

### 2.5.1 Evaluating Word Alignment

In order to evaluate the quality of the word alignment task, so far Alignment Error Rate (AER) as proposed in [Och00b] is commonly used. This measure requires a manual alignment reference (also called *gold standard*), indicating which source words should be linked to which target words for each reference sentence.

Due to the ambiguity brought up by the alignment task, two link types are allowed during manual tagging, namely Sure links (which *must* be present for the alignment to be correct) and Possible links (which *may* be present for the alignment to be correct, but are not compulsory). With these definitions, one can define Recall, Precision and AER measures thus:

$$recall = \frac{|A \cap S|}{|S|}, \quad precision = \frac{|A \cap P|}{|A|}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

where $A$ is the hypothesis alignment and $S$ is the set of Sure links in the gold standard reference, and P includes the set of Possible *and* Sure links in the gold standard reference.

It has been shown that the percentage of Sure and Possible links in the gold standard reference has a strong influence in the final AER result, favouring high-precision alignments when Possible links outnumber Sure links, and favouring high-recall alignments otherwise [Lam04]. A well-founded criterion is to produce Possible links only when they allow combinations which are considered equally correct, as a reference with too many Possible links suffers from a resolution loss, causing several different alignments to be equally rated.

In [Lam05] guidelines for building a word alignment evaluation scheme are presented. Taking into account that the notion of word alignment quality depends on the application, the authors review standard scoring metrics for full text alignment and give explanations on how to use them better, and suggest a strategy to build a reference corpus particularly adapted to applications

where recall plays a significant role, like in machine translation. Specific examples for the Spanish-English European Parliament corpus are also described.

Nowadays, due to a lack of perfect correlation between Alignment Error Rate and translation evaluation scores observed in many experiments, alternative word alignment evaluation metrics are being pursued. Very recent works on the subject can be found in [Fra06a, Aya06].

### 2.5.2 Word Alignment approaches

Currently statistical word alignment based on IBM and HMM models are considered to be state-of-the-art. A systematic performance comparison in terms of AER of these models can be found in [Och03c], where the authors also advocate a positive correlation between AER improvement and translation quality, as stated in [Och00a]. Typically, the implementation by[Och03a], which is freely-available in the GIZA++ package, is used. A great majority of current approaches to statistical translation depend on the results of this alignment tool to estimate their translation models.

However, due to the model definition of alignment as a function from positions in the target sentence to positions in the source sentence, the result is strictly asymmetric, generating one-to-many word alignments. Usually, this is tackled by performing the alignment from source to target and from target to source, and symmetrising via the union of links, the intersection or other refined methods as in [Och00b].

In [GV02] contextual information is added to the IBM models in the framework of maximum entropy, with small but consistent improvements in AER.

Other alignment models have been presented based on word cooccurrences, such as the Competitive Linking Algorithm (CLA) in [Mel00], or on link probabilities, as introduced in [Che03a] with promising results (as shown in [Mih03]). However, they generally assume a one-to-one constraint that, despite generating high-precision links, does not account for many translation phenomena. A similar idea is followed in [Che05] where one-to-one CLA alignments are combined with IBM-based symmetrised aligmment to extend the phrase extraction procedure.

In 2005, following an idea already hinted in [CB04], several independent works demonstrated that discriminatively trained models can equal or surpass the alignment accuracy of the standard models, if the usual unlabeled bilingual training corpus is supplemented with human-annotated word alignments for only a small subset of the training data [Liu05, Itt05, Fra05]. Therefore, current research efforts seem to be shifting towards a log-linear combination of feature models, estimated on a small word-aligned development set [Che06, Fra06b, Moo06, Blu06].

## 2.6 Use of linguistic knowledge into SMT

Although initially SMT systems did not incorporate any linguistic analysis and worked at the surface level of word forms, an increasing number of research efforts are introducing a certain degree of linguistic knowledge into their statistical framework.

At this point, the pair of languages involved and their respective linguistic properties are crucial to justify a certain approach and explain its results. Therefore, the idea that a *good* statistical translation model for a certain pair of languages can be used for *any* other pair is faced against the view that the goodness of such a model may be, at least in part, dependent on the specific language pair. Of course, conclusions will easily hold for languages sharing many linguistic properties.

To illustrate this, consider translating from French into English. While a certain vocabulary reduction of the source language may be useful in this direction, since many French words may translate to the same English word (due to morphological derivations which are not present in English), this same technique can be useless when translating in the opposite direction. In [Tal06] this experiment is conducted via automatic model clustering, by conflating those source words with similar translation distributions.

The use of POS information for improving statistical alignment quality of the HMM-based model is described in [Tou02], where they introduce additional lexicon probability for POS tags in both languages, but actually are not going beyond full forms. In [Pop04a] words that share the same base form are considered equal in the EM training of the alignment models, resulting in a AER reduction.

Regarding translation modelling, a primary work on the subject can be found in [Nie00], where several transformations of the source string for a German→English task are proposed, leading to increased translation performance. Transformations include issues such as compound words separation, reordering of separated verb prefixes (which are placed after the object in German) or word mapping to word plus POS to distinguish articles from pronouns, among others.

In [Nie04] hierarchical lexicon models including base form and POS information for translation from German into English are introduced, among other morphology-based data transformations. The same pair of languages is used in [CO04], where the inflectional normalisation leads to improvements in the perplexity of IBM translation models and reduces alignment errors.

More recently but still for the German→English pair, a sentence reordering as preprocessing is presented in [Col05]. Similarly to [Nie00], German input strings are reodered into a more English-like sentence order, obtaining better translation quality.

Regarding Romance languages, an approach to deal with inflected forms is presented in [Uef03], tackling verbs in an English→Spanish task. The authors join personal pronouns and auxiliaries to form extended English units and do not transform the Spanish side, leading to an increased English vocabulary. Translation quality in a small-data task is improved. In the opposite translation direction, [Pop04b] also transforms a text in a more-inflected language (Catalan, Spanish and Serbian) to separate base forms and suffixes for verb forms, improving slightly the perfomance when translating into a less-inflected language (English).

Regarding translation from another highly-inflected language such as Czech into English, [AO99] and [Gol05] present a couple of techniques modifying input Czech word (substituting them for lemmas, POS tags or combinations of both) into a language more similar to English, again obtaining improvements in BLEU scores for a small-data task.

When it comes to Chinese→English translation, a post-processing approach is followed in [Och03b]. The authors explore the application of syntax-based features in reranking $N$-best list of an SMT system based on alignment templates in a large-data task. Despite the many alternative trials, no significant gains in BLEU scores are reported.

Finally, regarding Arabic→English translation, [Lee04] reports performance boost when automatically-inducing Arabic word segmentation according to a word alignment to English material. Arabic is assumed to be preprocessed with all prefixes and suffixes separated. Extending on, a thorough study of the impact of Arabic word segmentation schemes into large-vocabulary translation into English is conducted in [Hab06]. Similar works can be found in [Zol06, EI06].

### 2.6.1 Other approaches

A number of researchers have proposed other translation models where the translation process involves syntactic representations of the source and/or target languages. These models have radically different structures and parameterisations from $N$gram–based (or phrase–based) models for SMT. Without aiming at completeness, some of these relevant works are mentioned here.

With the expressiveness of a context-free grammar rather than a left-to-right finite-state transducer, the formalism of Inversion Transduction Grammars [Wu97] was applied to modeling ordering shifts between languages, balancing needed flexibility against complexity constraints. Applications include bilingual parsing and machine translation.

Adding to that, in [Als00] dependency transduction by means of head transducers is introduced and applied to machine translation. Instead of consuming the input string from left to right as standard finite-state transducers, head transducers do it middle out at positions relative to other symbols in the output string.

Inspired by these works, in [Yam01] a channel translation model from input parsed trees into

target strings is introduced. Assuming the input string is fully parsed, operations on each node of the parse tree are defined, such as reordering child nodes, inserting extra words at each node, and translating leaf words. Apart from efficiency issues, a difficulty of this approach arises when parsing algorithms performance is not robust enough to handle non-grammatical sentences (as spoken language).

[Mel04] takes a different route to SMT and generalises the notion of parsing to the multidimensional space defined by multitexts (parallel texts between an arbitrary number of languages). Under this theoretical framework, previous monolingual parsing models can be easily extended to the bilingual (or synchronous) case by adapting grammar nodes and generation rules. However, no experimental results are reported and computational problems are already mentioned.

An approach to phrasal SMT based on a parsed dependency tree representation of the source language is introduced in [Qui05]. This approach, named Treelet translation, uses a source dependency parser and projects a target dependency tree using word alignment. After this projection, tree-based phrases are extracted and a tree-based ordering model can be trained.

Related to that, hierarchical phrases [Chi05] also remove the limitation to contiguous phrases and allow phrases to include indexed placeholders, thus turning phrase-based SMT into a parallel parsing problem over a grammar with one non-terminal symbols. This improves the global reordering search.

In general terms, these syntax-aware approaches have not shown very significant gains in performance when compared to phrase-based systems[4]. Apart from that, they tend to exhibit a larger structural (and computational) complexity. For these reasons, our approach has focused on introducing linguistic information into the statistical parameterisation of $N$gram-based SMT (which can be seen as structurally equivalent to phrase-based SMT).

## 2.7 Machine Translation evaluation

It is well-known that Machine Translation is a very hard task to evaluate automatically. Usually, this task is performed by producing some kind of similarity measure between the translation hypothesis and a set of human reference translations, which represent the expected solution of the system.

The fact that there are several *correct* alternative translations for any input sentence adds complexity to this task, and whereas the higher the correlation with the human references the better quality, theoretically we cannot guarantee that incorrelation with the available set of references means bad translation quality, unless we have *all* possible correct translations available.

[4]However, recent publications suggest that this may be the case in a near future.

Therefore, in general it is accepted that all automatic metrics comparing hypotheses with a limited set of manual reference translations are pessimistic. Yet, instead of an absolute quality score, automatic measures are claimed to capture progress during system development and to statistically correlate well with human intuition.

Next the most widely-used MT evaluation measures are introduced, such as BLEU, NIST, mWER and mPER. Other measures, which have not been used during this Ph.D. research work, are just referenced.

### 2.7.1 Automatic evaluation metrics

#### 2.7.1.1 BLEU score

Arguably the most extended evaluation measure as of today, BLEU (acronym for BiLingual Evaluation Understudy) was introduced by IBM in [Pap01], and is always referred to a given n-gram order ($BLEU_n$, $n$ usually being 4).

The metric works by measuring the n-gram co-occurrence between a given translation and the set of reference translations and then taking the weighted geometric mean. BLEU is specifically designed to approximate human judgement on a corpus level and can perform badly if used to evaluate the quality of isolated sentences.

$BLEU_n$ is defined as:

$$BLEU_n = exp\left(\frac{\sum_{i=1}^{n} bleu_i}{n} + length\_penalty\right) \tag{2.11}$$

where $bleu_i$ and $length\_penalty$ are cumulative counts (updated sentence by sentence) referred to the whole evaluation corpus (test and reference sets). Even though these matching counts are computed on a sentence-by-sentence basis, the final score is *not* computed as a cumulative score, ie. it is not computed by accumulating a given sentence score.

Equations 2.12 and 2.13 show $bleu_n$ and $length\_penalty$ definitions, respectively:

$$bleu_n = log\left(\frac{Nmatched_n}{Ntest_n}\right) \tag{2.12}$$

$$length\_penalty = min\left\{0, 1 - \frac{shortest\_ref\_length}{Ntest_1}\right\} \tag{2.13}$$

Finally, $Nmatched_i$, $Ntest_i$ and $shortest\_ref\_length$ are also cumulative counts (updated sentence by sentence), defined as:

$$Nmatched_i = \sum_{n=1}^{N} \sum_{ngr \in S} min \left\{ N(test_n, ngr), \max_r \{N(ref_{n,r}, ngr)\} \right\} \tag{2.14}$$

where $S$ is the set of Ngrams of size $i$ in sentence $test_n$, $N(sent, ngr)$ is the number of occurrences of the Ngram $ngr$ in sentence $sent$, $N$ is the number of sentences to eval, $test_i$ is the $i^{th}$ sentence of the test set, $R$ is the number of different references for each test sentence and $ref_{n,r}$ is the $r^{th}$ reference of the $n^{th}$ test sentence.

$$Ntest_i = \sum_{n=1}^{N} length(test_n) - i + 1 \tag{2.15}$$

$$shortest\_ref\_length = \sum_{n=1}^{N} \min_r \{length(ref_{n,r})\} \tag{2.16}$$

From BLEU description, we can conclude:

- BLEU is a quality metric and it is defined in a range between 0 and 1, 0 meaning the worst-translation (which does not match the references in any word), and 1 the *perfect* translation.
- BLEU is mostly a measure of *precision*, as $bleu_n$ is computed by dividing by the matching n-grams by the number of n-grams in the test (*not* in the reference). In this sense, a very high BLEU could be achieved with a *short* output, so long as all its n-grams are present in a reference.
- The *recall* or *coverage* effect is weighted through the *length_penalty*. However, this is a very rough approach to recall, as it only takes lengths into account.
- Finally, the weight of each effect (precision and recall) might not be clear, being very difficult from a given BLEU score to know whether the provided translation lacks recall, precision or both.

Note that slight variations of these definitions have led to alternative versions of BLEU score, although literature considers BLEU as a unique evaluation measure and no distinction among versions is done. Very recently, an interesting discussion with counterexamples of human correlation was presented in [CB06].

#### 2.7.1.2 NIST score

NIST evaluation metric, introduced in [Dod02], is based on the BLEU matric, but with some alterations. Whereas BLEU simply calculates n-gram precision considering of equal importance each n-gram, NIST calculates how informative a particular n-gram is, and the rarer a correct n-gram is, the more weight it will be given. NIST also differs from BLEU in its calculation of the brevity penalty, and small variations in translation length do not impact the overall score as much.

Again, NIST score is always referred to a given n-gram order ($NIST_n$, usually $n$ being 4), and it is defined as:

$$NIST_n = \left( \sum_{i=1}^{n} nist_i \right) \cdot nist\_penalty \left( \frac{test_1}{\frac{ref_1}{R}} \right) \tag{2.17}$$

where $nist_n$ and $nist\_penalty(ratio)$ are cumulative counts (updated sentence by sentence) referred to the whole evaluation corpus (test and reference sets). Even though these matching counts are computed on a sentence-by-sentence basis, the final score is *not* computed as a cumulative score.

The $ratio$ value computed using $test_1$, $ref_1$ and $R$ shows the relation between the number of words of the test set ($test_1$) and the average number of words of the reference sets ($ref_1/R$). In other words, the relation between the translated number of words and the expected number of words for the whole test set.

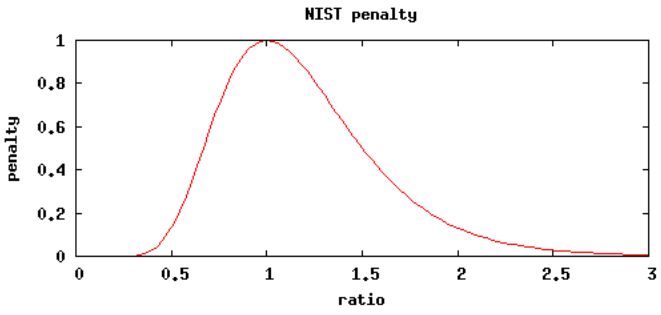



**Figure 2.5:** NIST penalty graphical representation

Equations 2.18 and 2.19 show $nist_n$ and $nist\_penalty$ definitions, respectively. This penalty function is graphically represented in Figure 2.5.

$$nist_n = \frac{Nmatch\_weight_n}{Ntest_n} \tag{2.18}$$

$$nist\_penalty(ratio) = exp\left(\frac{log(0.5)}{log(1.5)^2} \cdot log\,(ratio)^2\right) \tag{2.19}$$

Finally, $Nmatch\_weight_i$ is also a cumulative count (updated sentence by sentence), defined as:

$$Nmatch\_weight_i = \sum_{n=1}^{N} \sum_{ngr \in S} \left(\min\left\{N(test_n, ngr), \max_r \{N(ref_{n,r}, ngr)\}\right\} \cdot weight(ngr)\right) \tag{2.20}$$

where $weight(ngr)$ is used to weight every n-gram according to the identity of the words it contains, expressed as follows:

$$weight(ngr) = \begin{cases} -log_2\left(\frac{N(ngr)}{N(mgr)}\right) & \text{if mgr exists;} \\ -log_2\left(\frac{N(ngr)}{Nwords}\right) & \text{otherwise;} \end{cases} \tag{2.21}$$

where $mgr$ is the same N-gram of words contained in $ngr$ except for the last word. $N(ngram)$ is the number of occurrences of the Ngram $ngram$ in the reference sets. $Nwords$ is the total number of words of the reference sets.

The NIST score is a quality score ranging from 0 to (worst translation) to an unlimited positive value. In practice, this score ranges between 5 or 12, depending on the difficulty of the task (languages involved and test set length).

From its definition, we can conclude that NIST favours those translations that have the same length as the average reference translation. If the provided translation is perfect but 'short' (for example, it is the result of choosing the shortest reference for each sentence), the resultant NIST score is much lower than another translation with a length more similar to that of the average reference.

#### 2.7.1.3 mWER

Word Error Rate (WER) is a standard speech recognition evaluation metric, where the problem of multiple references does not exist. For translation, its multiple-reference version (mWER) is computed on a sentence-by-sentece basis, so that the final measure for a given corpus is based

on the cumulative WER for each sentence. This is expressed in 2.22:

$$mWER = \frac{\sum_{n=1}^{N} WER_n}{\sum_{n=1}^{N} Avrg\_Ref\_Length_n} \cdot 100 \tag{2.22}$$

where $N$ is the number of sentences to be evaluated. Assuming we have $R$ different references for each sentence, the *average reference length* for a given sentence $n$ is defined as:

$$Avrg\_Ref\_Length_n = \frac{\sum_{r=1}^{R} Length(Ref_{n,r})}{R} \tag{2.23}$$

Finally, the *WER* cost for a given sentence $n$ is defined as:

$$WER_n = \min_r LevDist(Test_n, Ref_{n,r}) \tag{2.24}$$

where *LevDist* is the Levenshtein Distance between the test sentence and the reference being evaluated, assigning an equal cost of 1 for deletions, insertions and substitutions. All lengths are computed in number of words.

mWER is an percentual error metric, thus defined in the range of 0 to 100, 0 meaning the *perfect* translation (matching at least one reference for each test sentence).

From mWER description, we can conclude that the score tends to slightly favour shorter translations to longer translations. This ca be explained by considering that the absolute number of errors (found as the Levenshtein distance) is being divided by the average sentence length of the references, so that a mistake of one word with respect to a long reference is being overweighted in contrast to one mistake of one word with respect to a short reference.

Suppose we have three references of length 9, 11 and 13 (avglen = 11). If we have a translation which is equal to the shortest reference, except by one mistake, we have a score of 1/11 (where, in fact, the error could be considered higher, as it is one mistake over 9 words, that is 1/9).

#### 2.7.1.4 mPER

Similar to WER, the so-called Position-Independent Error Rate (mPER) is computed on a sentence-by-sentece basis, so that the final measure for a given corpus is based on the cumulative PER for each sentence. This is expressed thus:

$$mPER = \frac{\sum_{n=1}^{N} PER_n}{\sum_{n=1}^{N} Avrg\_Ref\_Length_n} \cdot 100 \tag{2.25}$$

where $N$ is the number of sentences to be evaluated. Assuming we have $R$ different references for each sentence, the *average reference length* for a given sentence $n$ is defined as in eqnarray 2.23.

Finally, the $PER$ cost for a given sentence $n$ is defined as:

$$PER_n = \min_r \left(Pmax(Test_n, Ref_{n,r})\right) \tag{2.26}$$

where $Pmax$ is the maximum between:

- POS = num. of words in the REF that are not found in the TST sent. (*recall*)
- NEG = num. of words in the TST that are not found in the REF sent. (*precision*)

in this case, the number of words includes repetitions. This means that if a certain word appears twice in the reference but only once in the test, then POS=1.

#### 2.7.1.5 Other evaluation metrics

Apart from these, several other automatic evaluation measures comparing hypothesis translations against supplied references have been introduced, claiming good correlation with human intuition. Although not used in this Ph.D. dissertation, here we refer to some of them.

- Geometric Translation Mean, or**GTM**, measures the similarity between texts by using a unigram-based F-measure, as presented in [Tur03]
- Weighted N-gram Model, or **WNM**, introduced in [Bab04], is a variation of BLEU which assigns different value for different n-gram matches
- **METEOR** includes a word stemming process of the hypothesis and references to extend unigram matches (see [Ban05b])
- **ORANGE** ([Lin04b]) uses unigram co-occurrences and adapts techniques from automatic evaluation of text summarisation, as presented in the ROUGE score ([Lin04a])

- **mCER** is a simple multiple-reference character error rate, and is supplied by ELDA
- As a result from a 2003 John Hopkins University workshop on Confidence Estimation for Statistical MT, [Bla04] introduce evaluation metrics such as Classification Error Rate (**CER**) or the Receiving Operating Characteristic (**ROC**)
- From a more intuitive point of view, in [Sno05] Translation Error Rate, or **TER**, is presented. This measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation. Its application in real-life situation is reported in [Prz06].

Finally, rather than aiming at defining a supermetric 'XXX', in [Gim06] the IQMT framework is presented. This tool follows a 'divide and conquer' strategy, so that one can define a set of metrics and then combine them into a single measure of MT quality in a robust and elegant manner, avoiding scaling problems and metric weightings.

### 2.7.2 Human evaluation metrics

Human evaluation metrics require a certain degree of human intervention in order to obtain the quality score. This is a very costly evaluation strategy that seldom can be conducted. However, thanks to international evaluation campaigns, these measures are also used in order to compare different systems.

Usually, the tendency has been to evaluate *adequacy* and *fluency* (or other relevant aspects of translation) according to a 1 to 5 quality scale. Fluency indicates how natural the hypothesis sounds to a native speaker of the target language, usually with these possible scores: 5 for Flawless, 4 for Good, 3 for Non-native, 2 for Disfluent and 1 for Incomprehensible. On the other hand, Adequacy is assessed after the fluency judgement is done, and the evaluator is presented with a certain reference translation and has to judge how much of the information from the original translation is expressed in the translation by selecting one of the following grades: 5 for all of the information, 4 for most of the information, 3 for much of the information, 2 for little information, and 1 for none of it[5].

However, another trend is to manually post-edit the references with information from the test hypothesis translations, so that differences between translation and reference account only for errors and the final score is not influenced by the effects of synonymia. The human targeted reference is obtained by editing the output with two main constraints, namely that the resultant references preserves the meaning and is fluent.

[5]These grades are just orientative, and may vary depending on the task.

In this case, we refer to the measures as their human-targeted variants, such as HBLEU, HMETEOR or HTER as in [Sno05]. Unfortunately, this evaluation technique is also costly and cannot be used constantly to evaluate minor system improvements. Yet we are of the opinion that, in the near future, these methods will gain popularity do to the fact that, apart from providing a well-founded absolute quality score, they produce new reference translations that can serve to automatically detect and classify translation errors.

Regarding automatic error classification or analysis, some recent works on the subject suggest that it is possible to use linguistic information to automatically extract further knowledge from translation output than just a single quality score (we note the work of [Pop06a, Pop06b]).

### 2.7.3 International evaluation campaigns

Another very relevant factor contributing to the growth in SMT research are international evaluation campaigns. Organised by different institutions, consortiums, conferences or workshops, these campaigns are the perfect tool to assess the translation quality of state-of-the-art SMT systems. Furthermore, systems can be compared and knowledge is shared among researchers from several different sites.

With a large experience in automatic speech recognition benchmark tests, the National Institute of Standards and Technology (NIST), belonging to the Government of the United States, organises yearly machine translation tests since the early 2000s. Aiming at a breakthrough in translation quality, these tests are usually unlimited in terms of data for training. The target language is English, and sources include Arabic and Chinese. Further information can be accessed through `http://www.nist.gov/speech/tests/mt/index.htm`.

Since October 2004, the C-STAR[6] consortium organises the International Workshop on Spoken Language Translation (IWSLT) on a yearly basis. This workshop includes an evaluation campaign oriented towards speech translation and small data availability. Therefore, training material tends to be limited. Language pairs include Chinese, Japanese, Korean, Arabic, Italian and English (usually English being the target language). Reports of the 2004 and 2005 editions[7] are published in [Aki04] and [Eck05], respectively.

In 2005, a Workshop on *Building and Using Parallel Texts: data-driven MT and beyond*, organised at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), also included a machine translation shared task reported in [Koe05b]. In this case, translation between European languages (Spanish, Finnish, French, German and English) was the main task. Training included the European Parliament proceedings corpus ([Koe05a]).

[6] Consortium for Speech Translation Advanced Research, http://www.c-star.org

[7] Further information at http://www.slt.atr.jp/IWSLT2004 and http://www.is.cs.cmu.edu/iwslt2005 .

In 2006, a new edition of this evaluation campaign was conducted in the HLT/NAACL'06 Workshop on *Statistical Machine Translation*, as reported in [Koe06].

Additionally, the the European project TC-STAR (Technology and Corpora for Speech to Speech Translation) organised a first internal evaluation in 2005 (for members of the project, including UPC) and an open evaluation in 2006. Further details can be obtained from `http://www.elda.org/tcstar-workshop/2006eval.htm`.

# Chapter 3

# The Bilingual N-gram Translation Model

## 3.1 Introduction

This chapter is devoted to the study of the bilingual translation model implemented, which is the core model of UPC SMT system. This chapter is organised as follows:

- Firstly, the initial implementation using $X$-grams for speech translation (on a Finite-State Transducer architecture) is discussed in §3.2. The training of such a model, along with experiments and results with small corpora are reported. Main achievements, limitations and related publications are also commented.
- Then, in §3.3 we report on the evolution of this model towards an $N$-gram implementation, which is able to work on large-vocabulary tasks. Modelling issues such as tuple pruning and smoothing are also addressed, and experiments are reported on larger data sets.
- Finally, §3.4 presents a thorough study of the translation unit used in the model, ie. the tuple. Its main properties and alternative tuple definitions are evaluated and compared. These topics include a discussion on tuple structural definition in §3.2.3, the problem of embedded words in §3.4.1 and details on tuple segmentation in §3.4.2.

To conclude, in §3.5 a summary of the chapter can be found, highlighting the main conclusions extracted from it.

## 3.2 X-grams FST implementation

### 3.2.1 Reviewing X-grams for Language Modelling

A language model can be defined as the probability of a certain sequence of tokens (usually words), and it is a broadly-used model which plays a very important role in several statistical pattern matching tasks, ranging from speech recognition to machine translation or other language-related tasks defined on posterior probability frameworks.

In order to estimate the probability of a certain sequence of $M$ words $w_1, w_2...w_M$ as defined in equation 3.1, the most extended method are $n$-grams, which reduce the exponentially-increasing history space to the $n-1$ previous words, as shown in equation 3.2.

$$p(w_1 w_2 ... w_M) = \sum_{i=1}^{M} p(w_i \mid w_1 ... w_{i-1}) \tag{3.1}$$

$$p(w_1 w_2 ... w_M) \approx \sum_{i=1}^{M} p(w_i \mid w_{i-n+1} ... w_{i-1}) \tag{3.2}$$

Still, even for a moderate vocabulary size of $K = 1000$ words, the number of model parameters to estimate for *trigrams* ($n = 3$) is very large (more than $10^9$) and usually unobserved in training material. In this case, the maximum likelihood estimator is not appropriated, and smoothing techniques are required.

In [Bon96] an alternative approach to language modelling, named $X$-grams, is presented. Three main conclusions can be drawn from this work, namely:

- $n$-grams can be implemented *efficiently* by means of a Finite-State Automaton.

  Each state is represents a conditioning history $(w_{i-n+1}...w_{i-1})$ which has been seen in training material, and each arc contains the words $w_i$ which have followed them, together with the associated language model probability $p(w_i \mid w_{i-n+1}...w_{i-1})$. This way, the language model of a certain sentence can be computed by traversing the full sentence through the FSA and multiplying probabilities.

- State merging techniques can be used to smooth probabilities and achieve a language model with low perplexity.

  Two criteria to merge the state defined by history $(w_{i-n+1}...w_{i-1})$ into the smaller-history state defined by $(w_{i-n+2}...w_{i-1})$ are introduced, resulting in language models with low perplexity values. These criteria are:

- the states are merged if the longest-history state has occurred less than $k_{min}$ number of times in the training material
- the states are merged if the divergence between their output probability distributions $\mathbf{p} = p(w \mid w_{i-n+1}...w_{i-1})$ and $\mathbf{q} = q(w_i \mid w_{i-n+2}...w_{i-1})$ is smaller than a certain threshold $f$, where divergence is defined as in equation 3.3, a well-known information theory function

$$D(\mathbf{p} \parallel \mathbf{q}) = \sum_{j=1}^{J} p(i) \cdot \log \frac{p(j)}{q(j)} \tag{3.3}$$

- History size can vary depending on the words involved.

  When $n$-gram models increase their history size $n$, probabilities estimated for long histories may not be always reliable. On the other hand, even when reliable, perhaps these give no additional information with respect to probabilities conditioned on smaller histories. As the $X$-gram term denotes, the presented state merging techniques produce a model conditioned on variable-size histories, capturing long histories only whenever these bear some relevant information.

This language modelling technique was successfully incorporated into an speech recognition engine in [Bon98]. For this purpose, a very compact representation of $X$-grams using null transitions is introduced. However, this representation is not a formal automaton, as the valid transitions now depend on the way a certain state was accessed, causing a less efficient decoding search (as the valid transitions need to be *marked* at decoding time).

### 3.2.2 Bilingual X-grams for Speech Translation

Regarding machine translation, we can follow the same approach by estimating $X$-grams for a new *bilingual language*, whose words are not regular words anymore, but a composition of words from a source and a target language having a translational relationship (ie. in principle carrying the same meaning). This unit is to be called tuple, and can be defined as:

$$T_k = (s_{i_k}^{i_{k+1}-1}, t_{j_k}^{j_{k+1}-1}) \equiv (s,t)_k \tag{3.4}$$

where $s_{i_k}^{i_{k+1}-1}$ indicates the sequence of source words from position $i_k$ (where tuple $T_k$ starts) to position $i_{k+1}-1$, and $t_{j_k}^{j_{k+1}-1}$ the associated sequence of target words. As we can see, a source sentence and its translation can be decomposed into a sequence of such tuples $T_1, T_2, ..., T_k$.

Under this framework, given an input sentence, standard search algorithms (such as Viterbi) can still be used to find the best-probability path through the automaton. Only minor changes need to be done in order to quickly implement such a translator; since the vocabulary is now

bilingual and our input sentence is expressed only in the source language, the search must ignore the tuple target-language words. Once the most probable path is found, only the target-language words of the selected tuples must be output.

In doing so, the former Finite-State Automaton from §3.2.1 becomes a translation Finite-State Transducer (FST), as already introduced in [Vid97]. For $X$-grams, the same state merging techniques presented in the previous section can be equivalently applied to this bilingual model in order to smooth probability estimates.

In fact, this seems very advisable as a bigger data-sparseness problem is expected due to the tuple bilinguality; for a certain sequence of source words, a number of different tuples will contain possible translations into target words, which will always lead to a bigger tuple vocabulary than that of any monolingual text.

Additionally, the approach can easily be extended to speech translation by an elegant integration of the acoustic models into the FST. While in speech recognition each FSA arc representing a word is encoded into the HMM acoustic models of its phonetical representation, now these models encode the phonetic representation of the tuple source-language part.

Mathematically, if $x$ is the input speech signal and $s$ the source sentence associated to it, the translation search is defined as finding the target sentence $t$ which maximises equation 3.5:

$$\arg\max_t p(t \mid x) = \arg\max_t \sum_s p(t, s \mid x) \tag{3.5}$$

which can be simplified, by applying Bayes, approximating the sum over source sentences $s$ with the maximum, and assuming that the source input speech signal does not depend on the target sentence but only on the source sentence (ie. $p(x \mid s, t) \approx p(x \mid s)$), into equation 3.6:

$$\arg\max_t \max_s p(s, t) \cdot p(x \mid s) \tag{3.6}$$

where $p(s, t)$ is the $X$-grams bilingual model of tuples from equation 3.7, and $p(x \mid s)$ are the acoustic models of the source words. This is equivalent to equation 2.9 mentioned in §2.3.1.

$$p(s, t) = \prod_{k=1}^{K} p((s, t)_k | (s, t)_{k-X+1}, ..., (s, t)_{k-1}) \tag{3.7}$$

#### 3.2.2.1 Training from parallel data

Therefore, assuming we are able to extract a set of tuples from a given parallel text, we can use $X$-grams to estimate the bilingual model and, by means of a non-intrusive modification of Viterbi search algorithm[1], we can perform statistical machine translation.

The training procedure from a parallel text is graphically represented in Figure 3.1.

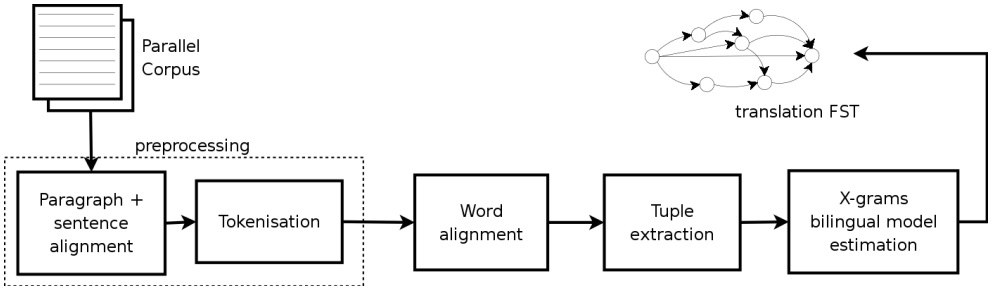



**Figure 3.1:** Training a translation FST from parallel data. Flow diagram.

The first preliminary step requires the preprocessing of the parallel data, so that it is sentence aligned and tokenised. By sentence alignment the division of the parallel text into sentences and the alignment from source sentences to target sentences is referred. This alignment is usually (though not always) monotone and, even though it produces one-to-one links (one source sentence aligned to one target sentence), it can also include one-to-many, many-to-one, and many-to-many links.

By tokenisation, we refer to separating punctuation marks, classifying numerical expressions into a single token, and in general, simple normalisation strategies tending to reduce vocabulary size without an information loss (ie. which can be reversed if required).

Then, word alignment is performed, by estimating IBM translation models (see §2.1.1) from parallel data and finding the Viterbi alignment in accordance to them. This process is carried out using the GIZA toolkit (see §2.1.2).

Finally, before estimating the bilingual $X$-grams, a tuple extraction from word-aligned data needs to be done. Given a word-aligned sentence, this process segments it into a sequence of tuples, respecting two crucial constraints:

- Monotonicity. The resultant segmentation can be traced sequentially in order to produce back *both* the source and the target sentences. This will be discussed in detail in §3.2.4.
- Minimal tuple size. For the resultant model to be less sparse, or in other words, for tuples

[1]It is non-intrusive in that is does not change the core Viterbi search function, but only read/write functions related with accessing the units that identify arcs.

to have the biggest generalising power, we are interested in the shortest tuples (in number of source and target words) which respect the previous constraint

This process is illustrated in Figure 3.2, where four tuples are extracted and it is worth noting that A→W is *not* extracted given the previous constraints, and that tuples with empty target side are allowed (source word D translates to no target word).

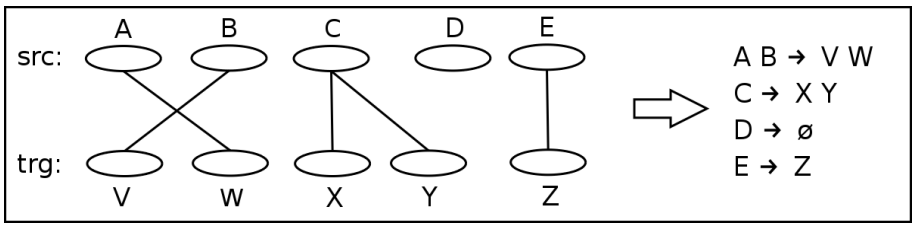



**Figure 3.2:** Tuple extraction from a certain word aligned pair of sentences.

Given a word alignment, these restrictions define a unique set of tuples except in one situation, which is whenever the resulting tuple contains no source word (a NULL-source tuple). In order to reuse these units in decoding new sentences, the search should allow for no input word to generate units, and this is not the case. Therefore, these units cannot be allowed, and a certain decision must be taken to re-segment tuples in these cases, as will be discussed in detail in §3.4.2.

#### 3.2.2.2 Preliminary experiment

In [Gis02b] the very preliminary first experiments using $X$-grams for speech translation from Spanish to English are reported. Due to the data unavailability at that time, the corpus used had to be manually collected, consisting of a small set of around 3 K very short sentences in English and Spanish. As these were collected from tourist phrase books, the domain mainly reduced to transportation, lodging, commercial and entertainment questions and answers. The statistics are shown in Table 3.1.

| Phrase books | | sent. | words | vcb. | avg.len. |
|---|---|---|---|---|---|
| train | English | 3,070 | 18,700 | 2,044 | 6.1 |
| | Spanish | | 16,500 | 2,650 | 5.4 |

**Table 3.1:** *Phrase books English-Spanish parallel corpus.*

In corpus preprocessing, six word categories were used to classify personal names, cities and countries (manually), as well as date, time and numerical expressions (automatically). Therefore, resultant tuples might contain category tags, indicating that smaller finite-state transducers will

translate each of their possible alternative values[2].

For instance, a certain tuple might translate ($s_i$ NAME$_S$ $\rightarrow$ NAME$_T$ $t_j$). Although having several such tuples with the same category but different $s_i$ and $t_j$ words might be inefficient (as the category translation does not depend on them), the layer architecture of the Xgram implementation offered a good solution to this inefficiency. Once the upper-level (bi-language units level) path is output, it searches through the lower level (word level) to look for the specific input category representation that lies under the category name. A list can offer then its translation. This way, only one finite-state transducer per category is needed, no matter the bi-language unit in which this category is found, as long as the same category appears in the source and the target sides.

For testing, 10 people were asked to produce 10 new sentences from this domain, amounting to a total of 100 sentences in Spanish for testing, for which a reference English translation was produced manually. As about speech translation experiments, the same set of 100 sentences was recorded three times on the phone by 20 speakers (15 utterances each), sampled at 8kHz and quantified using the A law at 8 bits per sample. The phonetic representation unit was the demiphone [Mn00], obtained through clustering as explained in [Mn99]. The recognition models were 750 units, trained with 25 hours of Spanish speech obtained from the SpeechDat database.

| **Spanish $\rightarrow$ English** | WER | BLEU (1-gram / 2-gram / 3-gram / 4-gram) |
|---|---|---|
| text input | 30.9 | 0.78 / 0.66 / 0.57 / 0.49 |
| speech input | 57.0 | 0.56 / 0.44 / 0.36 / 0.30 |

**Table 3.2:** *Preliminary experiment translation results from Spanish to English.*

Due to the small training data, to a certain mismatch between test and train, and to having only one reference translation (which leads to pessimistic translation error metrics), the obtained results were inconclusive as to the potential of the approach. However, for completeness and historical reasons, they are shown in Table 3.2. Recognition WER for Spanish was as high as 41.2 % by using the translation FST, whereas using monolingual $X$-grams trained on the Spanish side of the same material WER was 36.8 %.

### 3.2.3 Tuple definition: from one-to-many to many-to-many

During the training of a joint-probability bilingual translation model, and especially during tuple extraction, certain considerations need to be taken into account. On the one hand, one has to decide whether to structurally allow tuples to be *one-to-many* or *many-to-many*. If we restrict tuples to one-to-many structure, only **one** source word will be allowed for each tuple

[2]To avoid mismatch problems between source and target categories, these were forced to share the same tuple whenever statistical word alignment did not link them.

(independently of the number of target words), whereas if we allow them to have a more general many-to-many structure, they can have any number of source and target words[3].

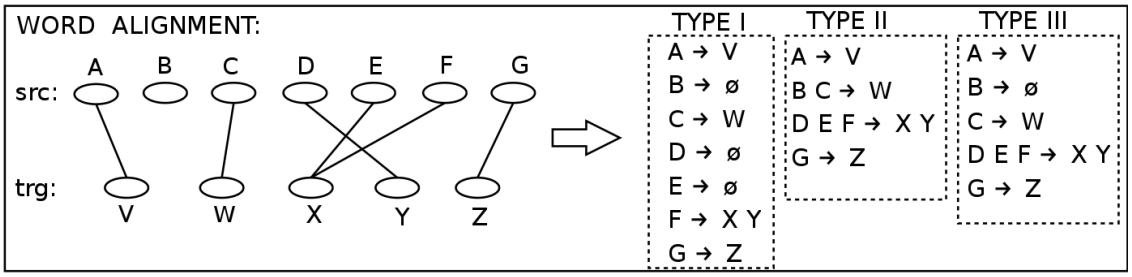



**Figure 3.3:** Three alternative tuple definitions from a given word alignment.

In [Cas00] two alternative units are presented, which are graphically represented in Figure 3.3. Firstly, all units are required to be one-to-many (**type I**), and whenever many source words are linked to a same target word or if there is a crossed dependency, only the last source word carries the full translation (whereas the previous ones are left without translation).

A second option is to force all tuples which have an empty target side to be linked to the next tuple with a non-empty target side, thus allowing the introduction of many-to-many units (**type II**). This solution is claimed to behave significantly worse in all translation experiments in [Cas00], due to the fact that the obtained finite-state transducers are much bigger, and consequently, the assigned probabilistic distributions are poorly estimated.

We experimented with two different alternative tuple definitions; on the one hand, the one-to-many structure defined as type I; on the other, we allowed for many-to-many tuples, but *only if* automatic word alignment had linked the words (be it through a many-to-one link or due to a crossed dependency). In practice, this is equivalent to type II, except for those tuples comprised of one source word without a link to any target word. These tuples are kept in the model, as shown in Figure 3.3 (**type III**).

These experiments were carried out on two Spanish→English and one English→Spanish tasks with two corpora containing transcriptions of spontaneously spoken dialogues. Therefore, sentences often lack correct syntactic structure.

On the one hand, we use a subset of VerbMobil corpus comprised of 20k sentences, and on the other, a Spanish-English parallel corpus developed in the framework of the LC-Star project. While the domain of VerbMobil dialogues is appointment scheduling and travel planning, LC-Star presents a broader domain including general tourist information, apart from appointment scheduling and travel planning.

[3]Excluding the case with no source words, as discussed below in §3.4.2.

| VerbMobil-20k | | sent. | words | vcb/OOVs | max.len. | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | English | 20,533 | 201,727 | 3,130 | 66 | 9.8 | 1 |
| | Spanish | | 191,344 | 4,850 | 64 | 9.3 | |
| test | English | 2,059 | 21,344 | 289 | 57 | 10.4 | 1 |

**Table 3.3:** *VerbMobil-20k parallel corpus statistics.*

Corpus preprocessing only included standard tokenisation and punctuation marks removal. The statistics of each corpus are shown in Tables 3.3 and 3.4, respectively. Note the big difference in number of running words and vocabulary sizes.

| LC-Star | | sent. | words | vcb/OOVs | max.len. | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | English | 30,117 | 424,047 | 6,180 | 94 | 14.1 | 1 |
| | Spanish | | 406,001 | 10,155 | 95 | 13.5 | |
| test | English | 500 | 9,220 | 67 | 84 | 18.4 | 1 |
| | Spanish | | 9,126 | 42 | 85 | 18.3 | 1 |

**Table 3.4:** *LC-Star parallel corpus statistics.*

In principle, any unit definition allowing for many-to-many tuples is prone to generate a sparser model. Therefore, it is reasonable to expect type I to outperform other tuple types. However, all our experiments systematically showed a different trend, being type III the best performing option for all tasks tested, as results show in Table 3.5.

| | **VerbMobil-20k** | | **LC-Star** | | | |
|---|---|---|---|---|---|---|
| | **Spa → Eng** | | **Spa → Eng** | | **Eng → Spa** | |
| | WER | BLEU | WER | BLEU' | WER | BLEU' |
| type I | 31.91 | 0.4915 | 44.16 | 0.3393 | 45.71 | 0.3558 |
| type III | 30.66 | 0.5042 | 42.89 | 0.3512 | 42.80 | 0.3839 |

**Table 3.5:** *Translation results with two different tuple types.*

In our opinion, this model preference for type III was further confirmed by human inspection of the translation outputs. As we observed, type I had a strong tendency to omit translations, possibly by catenating many *one-to-NULL* tuples (which have no target side). In fact, whereas the percentage of tuples with empty target side is about 20% for type I, this figure is lowered to about 10% when extracting tuple type III ([Bat04]). It seems that, at least for Eng↔Spa tasks, the benefits of minimising the risk of catenating several one-to-NULL tuples during translation (causing undesired omissions) overweights the sparsity increase of the model, achieving better translations. Clearly, in such tasks, having many one-to-NULL tuples weakens the automat's capacity to learn from past history, especially for the Eng→Spa direction, where omission errors seem to be particularly harmful, possibly due to the fact that Spanish tends to express the same meaning in more words.

Closely related to tuple structure is the word alignment used to extract tuples from. If tuples are extracted from source-to-target word alignment, many one-to-NULL tuples are generated, and the model is prone to omit translations (resembling tuple type I). On the contrary, if a symmetrised word alignment is used (the 'union' of source-to-target and target-to-source alignments), then less tuples have empty translation, and more many-to-many tuples make the model more sparse. This issue will be addressed again in §3.2.5.2.

### 3.2.4 Monotonicity vs. word reordering

Given the need for a sequence of tokens related to the $X$-grams formulation, the very basic treat of tuples is their *monotonicity*. In other words, when following them in monotone order, tuples *must* produce both the source and target sentence in the correct word order.

However, each language expresses concepts in different order, and depending on the pair of languages involved, translation will not be a monotone process. In fact, this is also captured in the definition of IBM translation and alignment models of §2.1.1, where crossed dependencies are expected to occur.

Therefore, if two languages are close in word order, alignment will tend to be monotone, and the extracted tuples from training data will be shorter (include less words) and easier to re-use during decoding. On the contrary, if these languages differ strongly in word order, alignments will have so-called long-distance links, forcing the extraction of long tuples (including many words), leading to a sparser bilingual model with a bigger tuple vocabulary. This is exemplified in Tables 3.6 and 3.7.

| | NULL | La | troballa | de | l' | objecte | el | va | dur | a | l' | indret | que | cercava |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| buscaba | . | . | . | . | . | . | . | . | . | . | . | . | . | + |
| que | . | . | . | . | . | . | . | . | . | . | . | . | + | . |
| lugar | . | . | . | . | . | . | . | . | . | . | . | + | . | . |
| al | . | . | . | . | . | . | . | . | . | + | + | . | . | . |
| llevó | . | . | . | . | . | . | . | + | + | . | . | . | . | . |
| le | . | . | . | . | . | . | + | . | . | . | . | . | . | . |
| objeto | . | . | . | . | . | + | . | . | . | . | . | . | . | . |
| el | . | . | . | . | + | . | . | . | . | . | . | . | . | . |
| Encontrar | . | + | + | . | . | . | . | . | . | . | . | . | . | . |
| NULL | . | . | . | + | . | . | . | . | . | . | . | . | . | . |

**Extracted tuples:**

Encontrar # La troballa
NULL # de
el # l'
objeto # objecte
le # el
llevó # va dur
al # a l'
lugar # indret
que # que
buscaba # cercava

**Table 3.6:** *Monotone alignments and tuple extraction. Spanish-Catalan example.*

Thus, we can conclude that, in principle, the bilingual model approach implemented by $X$-grams is more appropriate for addressing translation with closely-related languages, in terms

| | NULL | A | rather | sad | and | unacceptable | incident | happened | here | yesterday |
|---|---|---|---|---|---|---|---|---|---|---|
| ayer | . | . | . | . | . | . | . | . | . | + |
| aquí | . | . | . | . | . | . | . | . | + | . |
| inadmisible | . | . | . | . | . | + | . | . | . | . |
| e | . | . | . | . | + | . | . | . | . | . |
| triste | . | . | . | + | . | . | . | . | . | . |
| muy | . | . | + | . | . | . | . | . | . | . |
| incidente | . | . | . | . | . | . | + | . | . | . |
| un | . | + | . | . | . | . | . | . | . | . |
| produjo | . | . | . | . | . | . | . | + | . | . |
| Se | . | . | . | . | . | . | . | + | . | . |
| NULL | . | . | . | . | . | . | . | . | . | . |

**Extracted tuples:**

Se produjo un incidente muy triste e inadm. #
A rather sad and unacc. incident happened
aquí # here
ayer # yesterday

**Table 3.7:** *Non-monotone alignments and tuple extraction. Spanish-English example.*

of word order. For instance, translating between Romance languages (Catalan, French, Italian, Portuguese, Spanish, etc.).

Note that, on the other hand, tuple extraction will be more influenced from erroneous word alignments which include false long-distance links, than the phrase extraction strategy from §2.2.2, which is able to extract smaller units independently apart from the largest units.

#### 3.2.4.1 Studying English–Spanish cross patterns

Notwithstanding this monotonicity restriction, the bilingual approach was applied to two limited-domain English↔Spanish tasks in 2003, as reported in [Gis03]. These tasks were the complete VerbMobil database[4], containing around 30k sentences and featuring a Catalan parallel translation, and an extract of the European Parliament 1996-2001 proceedings, containing 30k carefully-selected sentences[5].

These corpora statistics are shown in Tables 3.8 and 3.9, respectively. Statistics include number of sentences, words, vocabulary size, maximum sentence length, average sentence length, and number of human reference translations available. Note that, for test sets, vocabulary size is not shown but the number of Out-Of-Vocabulary (OOVs) words (ie. words appearing in the test which are not present in the training data) instead.

[4]Similar to the corpus used in §3.2.3, but now repeated training pairs are included, as well as punctuation marks.

[5]Due to computational problems at that time, all sentences were scored according to a geometric mean, computed by assigning each of its words a value, equivalent to its position within a frequency-of-occurrence ranking (extracted from the whole corpus). This way, the most 'vocabulary-consistent' sentences were selected.

| VerbMobil | | sent. | words | vcb/OOVs | max.len. | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | Catalan | 27,995 | 253,241 | 4,909 | 75 | 9.1 | 1 |
| | English | | 262,681 | 3,174 | 74 | 9.4 | |
| | Spanish | | 253,394 | 5,620 | 74 | 9.1 | |
| test | Catalan | 2,059 | 25,425 | 234 | 71 | 11.8 | 1 |
| | English | | 25,454 | 137 | 68 | 12.3 | |
| | Spanish | | 25,421 | 296 | 69 | 11.8 | |

**Table 3.8:** *VerbMobil parallel corpus statistics.*

| EuParl-30k | | sent. | words | vcb/OOVs | max.len. | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | English | 30,000 | 444,224 | 25,244 | 181 | 14.8 | 1 |
| | Spanish | | 417,435 | 39,998 | 177 | 14.0 | |
| test | English | 500 | 5,830 | 217 | 79 | 11.7 | 1 |
| | Spanish | | 5,639 | 390 | 78 | 11.3 | |

**Table 3.9:** *EuParl-30k parallel corpus statistics.*

In order to tackle the non-monotonicity problem of the English↔Spanish pair, first a study of the most common non-monotonic alignments patterns (or *cross patterns*) was performed. For this, we defined each cross as a set of links $(x_1, y_1), (x_2, y_2), ...(x_N, y_N)$ where each $(x_i, y_i)$ describes a link between position $x_i$ and $y_i$ in the source and target sub-sequences composed of the words appearing in the cross.

By extracting all word-aligned training examples with an alignment cross, we obtain the statistics from Table 3.10, which present the ten most frequent cross patterns when aligning from Spanish to English and from English to Spanish with the VerbMobil data. These patterns account for more than 60 % of all cases with a non-monotonic alignment, and reflect linguistic relationships between both languages, whereas the patterns appearing only once generally link many words and either are produced by infrequent long links or by a poor automatic word alignment.

The columns entitled 'EuParl' show how representative are these patterns in the EuParl corpus. As it can be seen, although there are some slight differences in the distributions, the trend tends to be the same. It is worth noting the much higher percentage of patterns appearing only once, which might be outlining a bad performance in the automatic alignment due to data sparseness.

**Spanish→English patterns**

The most frequent pattern (1,2)(2,1) usually takes the form of 'Noun + Adjective # Adjective + Noun', as in '`semana siguiente # next week`'. Less frequently, we found the form

| Spa→Eng pattern | Vmobil | EuParl | Eng→Spa pattern | Vmobil | EuParl |
|---|---|---|---|---|---|
| (1,2) (2,1) | 24.6 % | 32.1 % | (1,2) (2,1) | 36.7 % | 36.4 % |
| (1,2) (2,0) (3,1) | 15.0 % | 3.6 % | (1,3) (2,1) | 8.7 % | 2.1 % |
| (1,3) (2,1) (2,2) | 5.4 % | 1.2 % | (1,2) (2,3) (3,1) | 4.8 % | 4.7 % |
| (1,3) (2,0) (3,1) (4,2) | 3.9 % | 0.2 % | (1,3) (2,4) (3,1) | 3.4 % | 0.1 % |
| (1,3) (2,1) (3,2) | 3.4 % | 1.5 % | (1,2) (2,0) (3,1) | 2.5 % | 0.6 % |
| (1,2) (2,0) (3,0) (4,1) | 2.9 % | 0.4 % | (1,3) (2,1) (3,2) | 2.3 % | 0.2 % |
| (1,2) (2,3) (3,1) | 1.8 % | 0.2 % | (1,2) (1,3) (2,1) | 1.9 % | 0.5 % |
| (1,3) (2,0) (3,0) (4,1) | 1.4 % | 0.3 % | (1,2) (2,0) (3,3) (4,1) | 1.8 % | 0 % |
| (1,3) (2,0) (3,0) (4,1) (5,0) (6,2) | 1.3 % | 0.1 % | (1,4) (2,0) (3,1) | 1.3 % | 0.2 % |
| (1,4) (2,0) (3,1) (4,2) (4,3) | 0.9 % | 0.1 % | (1,3) (2,1) (2,2) | 1.2 % | 0.1 % |
| 10 Vmobil most freq. patterns | 60.5 % | 39.7 % | | 64.6 % | 44.9 % |
| Patterns occurring only once | 19.1 % | 59.1 % | | 15.3 % | 54.5 % |

**Table 3.10:** *Most-frequent cross patterns for Spa→Eng and Eng→Spa alignments*

'Adverb1+ Adverb2 # Adverb2 + Adverb1', as in '`bastante bien # good enough`. Interestingly, some cases are clear omissions, as in '`hecho es # is actually`', where the preceding Spanish word '`de`' was omitted due to the asymmetry of the alignment models (discussed in §2.1.1). This can be addressed with information of the crosses when aligning in the other direction.

The second most frequent pattern (1,2)(2,0)(3,1) nearly always takes the form of a Noun followed by a Spanish prepositional clause (Noun + Preposition + Noun), and Adjective + Noun in English, as in '`viaje de negocios # business trip`, or comparative adjectives as in '`hotel más barato # cheaper hotel`. The third pattern (1,3)(2,1)(2,2) reveals the relationship 'Adverb + Verb = Pronoun + Verb + Adverb' existing in cases such as '`siempre podemos # we can always`, a relationship impossible to detect when aligning from English to Spanish.

However, even in this case some crosses are wrongly detected, as in '`teoría puedo # I can conceivably`, where Spanish preceding word '`en`' is left out of the cross.

#### English→Spanish patterns

Interestingly, in this direction the great majority of patterns are concentrated into the most frequent one (1,2)(2,1), because it includes many cases from the first and third most-frequent patterns from the opposite alignment direction. It is clear that the Spa→Eng alignment is better than the Eng→Spa for these cases. However, in other situations, we found the conclusion to be the contrary. For example, whenever Spa→Eng alignment detects the (1,2)(2,1) pattern (such as '`vez entonces # then again`) and the Eng→Spa detects the (1,3)(2,1)(2,2) pattern ('`otra vez entonces # then again`), the latter is more reliable.

This suggests the existence of some complementary information between both views and

thus a combination strategy to better detect non-monotonic alignments might be possible.

#### 3.2.4.2 An initial reordering strategy

In order to improve the generation of bilingual tuples during the FST training, the introduction, for the most frequent patterns, of indexes referring to the relative positions of the target-language words given a cross was presented. For example, for pattern (1,2)(2,1), we swapped the order of the target words, introducing an index to the last target word to preserve the original link. This is expressed by:

$$\left\{\begin{array}{lcl} s_i & \rightarrow & t_{j+1} \\ s_{i+1} & \rightarrow & t_j \end{array}\right. \Longrightarrow \left\{\begin{array}{lcl} s_i & \rightarrow & t_j \\ s_{i+1} & \rightarrow & t_{j+1}(-1) \end{array}\right.$$

where $s_i$ is the source word in position $i$ and $t_j$ is the target word in position $j$. The same applies for the second most frequent pattern (1,2)(2,0)(3,1), as the middle source word is aligned to no target word. For the third pattern (1,3)(2,1)(2,2), the same can be done after joining the two last words of the target sentence into a single token.

$$\left\{\begin{array}{lcl} s_i & \rightarrow & t_{j+2} \\ s_{i+1} & \rightarrow & t_j \\ s_{i+1} & \rightarrow & t_{j+1} \end{array}\right. \Longrightarrow \left\{\begin{array}{lcl} s_i & \rightarrow & t_j \\ s_{i+1} & \rightarrow & t_{j+1}t_{j+2}(-1) \end{array}\right.$$

Equivalently, the same procedure was carried out for the five most frequent VMobil cross patterns. Results are shown in the upper rows of Table 3.11, where **cross5** results compare favourably to the baseline for all tasks reported.

As complementary information regarding non-monotonic alignments was found when aligning from one direction to the other, another translation experiment was performed by adding as valid all those crosses which were detected in the opposite direction and belonged to the three most frequent patterns, but were not found in the straight direction. For instance, in Spa→Eng, we found 582 Eng→Spa crosses (489 for (1,2)(2,1), 39 for (1,3)(2,1) and 54 for (1,2)(2,3)(3,1)) which did not map into any cross in Spa→Eng alignment, and were thus added and subsequently coded. Results, shown in row **crossX** show a very slight further improvement.

All in all, results showed a weak tendency to improvement, but definitely not a marked one. In fact, the best result was achieved through a much simpler categorisation strategy (described in the following section). The main reason to account for this seems to be that basically, the effect of introducing a relative-position index in some words when coding tuples seems to generalise but only in very few test cases. On the other hand, making the crossed relationship only dependent on relative positions and words might not help in some cases. For example, when dealing with the typical Spa→Eng pattern $Noun_S + Adj_S = Adj_E + Noun_E$, the new coding will be:

$Noun_S/Noun_E$ $Adj_S/Adj_E$-1

| | **VerbMobil** | | | | **EuParl-30k** | |
|---|---|---|---|---|---|---|
| | **Spa → Eng** | | **Cat → Eng** | | **Spa → Eng** | |
| | WER | BLEU | WER | BLEU | WER | BLEU |
| baseline | 31.15 | 0.564 | 33.18 | 0.541 | 48.25 | 0.463 |
| +categ | 29.67 | 0.583 | 31.40 | 0.564 | | |
| +cross5 | 30.80 | 0.568 | | | 47.70 | 0.466 |
| +crossX | 30.70 | 0.569 | 32.68 | 0.542 | 47.29 | 0.470 |
| +baseV | 30.80 | 0.570 | 32.75 | 0.542 | | |
| +baseA | 30.90 | 0.566 | | | | |
| +baseVA | 30.91 | 0.569 | | | | |
| +categ +crossX | 29.37 | 0.584 | | | | |
| +categ +baseV | 29.57 | 0.583 | | | | |
| +categ +crossX +baseV | 29.30 | 0.584 | 31.16 | 0.565 | | |

**Table 3.11:** *Translation results when reordering by indexing cross information ('cross' rows) and classifying words to base form for alignment ('base' words), for VerbMobil and EuParl-30k tasks.*

where the noun is generalised (whenever the same noun appears in a different context, it will be correctly translated, because its probability is boosted thanks to the coding), but the adjective is not. When this adjective appears in a different unseen context, the transducer will have to choose between a monotone (without index) and non-monotone (with index) translation, deciding for the most frequent in the training.

In addition to that, the upper bound of this reordering strategy was not measured (ie. by means of an oracle score), being hard to conclude whether the new FST morphology (including relative-position indexes) was insufficient to achieve correct order, or whether simple the new estimated probabilities were not good enough (or both).

#### 3.2.4.3 Morphology-reduced word alignment

In addition, some preliminary results were presented on introducing POS-tagging and lemmatisation, as well as some preprocessing such as categorisation, to help improving the training of the system. In row **categ** from Table 3.11, date and time expressions were automatically categorised (substituted by a unified tag), achieving a more generalised translation FST[6].

Several experiments have been conducted substituting the original forms of Spanish verbs, adjectives, nouns and determiners, by their base form before aligning. The objective was to reduce variability and enforce links between, for example, a Spanish adjective, which might be declined in gender or number, and its English counterpart, which is invariant. For this, Spanish and Catalan data were tagged using **maco+** and **relax**, the tagger and morphological disambiguator for unrestricted text developed at TALP Research Centre [Ats98].

[6] 2746 instances of time expressions and 897 of date expressions were substituted in the training corpus

As seen in the lower part of Table 3.11, the most relevant results were found by classifying Verbs or Adjectives (see rows **baseV** and **baseA**), but improvements were again very meager. By far the major impact on evaluation metrics was achieved by categorising date and time expressions, or combinations of this with the previous presented strategies. Finally, given the structural similarity between Spanish and Catalan, the results for Catalan→English show the same trend.

### 3.2.5 The TALP X-grams translation system

The bilingual X-grams translation system implemented with an FST architecture and presented above served as UPC–TALP Research Centre Statistical Machine Translation system until the end of 2004, before undergoing severe changes (which will be discussed in detail in chapter 4).

The presentation of the system, as well as the preliminary morphology and reordering experiments, were published in the following contributions:

- [Gis02b] A. de Gispert and J.B. Mariño, "Using X-grams for Speech-to-Speech Translation," in *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP'02*, pps. 1885–1888, September 2002.
- [Gis02a] A. de Gispert and J.B. Mariño, "Análisis de las relaciones cruzadas en el alineado estadístico para la traducción automática," in *II Jornadas en Tecnología del Habla*, December 2002.
- [Gis03] A. de Gispert and J.B. Mariño, "Experiments in word-ordering and morphological preprocessing for transducer-based statistical machine translation," in *IEEE Automatic Speech and Understanding Workshop, ASRU'03*, pps.634–639, November 2003.

In July 2004, the system was showcased in a public demonstration for the FAME project closure, and by August the same year it participated in the 1st IWSLT Workshop, competing against other systems in a Chinese→English task.

#### 3.2.5.1 FAME project public demonstration

Funded by European Union, the FAME[7] project (acronym of Facilitating Agents for Multicultural Exchange) joined efforts from Interactive Systems Labs at Universität Karlsruhe, IRST (Trento), Univ. Politècnica de Catalunya (Barcelona), Laboratoire GRAVIR (Grenoble), Laboratoire CLIPS (Grenoble), SONY Germany and ATLAS (Barcelona), in order to advance technologies towards a more human-like communication and information access environment.

Among other research fields such as robust speech recognition or visual recognition and

[7] EU contract IST-2000-28323. http://islold.ira.uka.de/fame/

tracking of human action, improving limited-domain Catalan↔English and Spanish↔English speech translation was one of the main goals. Two research lines were conducted, namely an interlingua-based approach ([Arr04]) and an statistical approach (X-gram-based SMT as presented above), whose quality was compared in [Arr05].

The city council of the city of Barcelona pledged support for the FAME project by providing financial and organisational support for a demonstration of the system at the "Forum of Cultures", which was held in Barcelona in July 2004 (during the celebration of the 42nd Annual Meeting of the Association for Computational Linguistics, ACL'2004).

The demo, followed by the Mayor of Barcelona and all the major media, featured both speech translation systems working on-line by showing their speech recognition and translation outputs on a panel screen. Even though the domain was restricted to travel-related topics (as users were asked to pretend they were in need of travel agency services), it was seen as a success. From a technical point of view, the statistical translation component was of the same quality as that of the system just discussed in §3.2.4.2.

#### 3.2.5.2 IWSLT'04 participation

In October 2004, the C-STAR[8] consortium organised the 1st International Workshop on Spoken Language Translation (IWSLT), with the objective of providing a framework for the applicability validation of existing machine translation evaluation methodologies to evaluate speech translation technologies. It included an evaluation campaign whose details can be found in [Aki04].

The task was to translate from Japanese and Chinese to English, and 14 institutions took part in it. Not only statistical MT systems but also example-based MT systems, one rule-based system and other hybrid approaches were presented. UPC-TALP participated with the $X$-grams FST translation system in the Chinese→English task.

| IWSLT04 | | sent. | words | vcb/OOVs | max.len. | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | Chinese | 20,000 | 182,904 | 7,643 | 69 | 9.1 | 1 |
| | English | | 188,935 | 8,191 | 75 | 9.4 | |
| develop. | Chinese | 506 | 3,515 | 160 | 24 | 6.9 | 16 |
| test | Chinese | 500 | 3,794 | 104 | 62 | 7.5 | 16 |

**Table 3.12:** *IWSLT'04 Chinese→English parallel corpus statistics.*

Table 3.12 shows the details of the parallel corpus used for this task. During the system development work, the effect of state merging techniques on bilingual modelling (see §3.2.1), yielding a slight improvement by merging states whose divergence was smaller than 0.2 (even

[8] Consortium for Speech Translation Advanced Research, http://www.c-star.org

for such a reduced corpus size).

Two runs were submitted, basically comparing the effect of choosing a symmetrised word alignment to extract tuples from (by performing the union of Chi→Eng and Eng→Chi links) or the straight Chi→Eng word alignment (**s2t**, meaning source-to-target). Table 3.13 presents the main results achieved by the system for the development and test sets (the latter including a manual evaluation of fluency and adequacy).

| set | run | align | BLEU | NIST | WER | PER | fluency | adequacy |
|---|---|---|---|---|---|---|---|---|
| dev | A | union | 0.255 | **5.210** | **60.3** | **51.8** | | |
| | B | s2t | **0.314** | 3.678 | 60.7 | 54.8 | | |
| test | A | union | 0.279 | **6.778** | 55.6 | **46.5** | 2.792 | 3.022 |
| | B | s2t | **0.331** | 5.391 | **55.0** | 49.0 | | |

**Table 3.13:** *Evaluation results for IWSLT'04 Chinese→English task*

Despite contradictory evaluation metrics, a qualitative inspection of the results suggested the use of the union instead of the source-to-target alignment to be more adequate (even though it produced a lower BLEU score).

The main reason was output length. Given the difficulty of the task in terms of word ordering, many long-distance links are found in automatic word alignment and, as explained in §3.2.4, this enlarges extracted tuples and makes the bilingual model more sparse. Theoretically, this situation seems to be aggravated when extracting tuples from the union alignment, since more links are present and it is more likely that long links occur. However, when extracting from the source-to-target alignment (which is a one-to-many alignment), many more tuples are linked to NULL, ie. they do not have any token in their target side. For this *hard* task, in which due to sparseness the bilingual model is often forced to go without context, this has a very negative effect on output. In fact, whereas the union translation generated around 3k words, the s2t had only 2,5k.

Other research efforts were directed towards automatically extracting a translation dictionary for those words which were present within tuples along with other words and never seen as a single tuple (*embedded* words, which will be discussed in detail in §3.4.1), but with little impact on scores.

All in all, when compared to other participants, results were not positive. The TALP SMT system landed on the 8th position of a 9-row ranking (see [Aki04] for the complete table), confirming the inadequacy of its monotonicity for a task in need of reordering strategies. It goes without saying that the lessons learned from this experience were crucial to further develop the system, as will be clearly seen in chapter 4.

Further details on these experiments are reported in the following publication:

- [Gis04a] A. de Gispert and J.B. Mariño, "TALP: Xgram-based Spoken Language Translation System," in *Proceedings of the 1st International Workshop on Spoken Language Translation, IWSLT'04*, pps. 85–90, October 2004.

## 3.3 N-gram implementation

The $X$-grams implementation of the joint-probability bilingual model evolved into a standard $N$-gram implementation by the end of 2004. The motivation behind this change was double:

- Software incapacity to deal with large amounts of data.

  The $X$-grams implementation by means of a Finite-State Transducer was achieved by adapting a speech recognition toolkit developed at UPC during the late 1990s, and which was not prepared to work with vocabulary sizes beyond the hundred thousand words. Therefore, this posed a strong limitation to scale the translation system to large-vocabulary tasks, unless a substitute $X$-grams translation software was fully reprogrammed.

- Availability of tools implementing large-vocabulary language modelling which included the most advanced smoothing strategies

  At the same time, SRI International released a freely-available language-modelling toolkit (SRILM), presented in [Sto02]. This collection of C++ libraries, executable programs, and helper scripts was designed to allow both production of and experimentation with statistical language models for speech recognition and other applications, supporting creation and evaluation of a variety of language model types based on N-gram statistics (including many smoothing strategies), among other related tasks.

However, another implication of such a change was the need for a decoder. In order to study the $N$-gram translation model, and given that the FST architecture had been discarded for hash tables, the Viterbi search implemented onto the FST architecture was not useful anymore. An ngram-based SMT decoder was developed by Josep M. Crego (member of the TALP Research Centre at UPC) and presented in [Cre05b].

Next the most relevant modelling aspects of this $N$-gram bilingual model are studied in detail.

### 3.3.1 Modelling issues

In this section we focus on the $N$-gram implementation of the tuples bilingual model, conducting a thorough study of its main modelling aspects, and their effects on translation quality and model size. These aspects include:

- a study on model history length $N$
- a study on pruning strategies

- a study on smoothing strategies

The following study was carried out on a large-data Spanish↔English task, defined by a corpus containing the European Parliament Proceedings from 1996 to September 2004, whose main statistics are shown in Table 3.14. Statistics include number of sentences, running words, vocabulary size, out-of-vocabulary words (for development and test sets), average sentence length and number of human references available.

| EuParl ver1 | | sent. | words | vcb | OOVs | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | English | 1.22 M | 33.38 M | 105.0 k | - | 27.3 | 1 |
| | Spanish | | 34.79 M | 168.7 k | - | 28.4 | |
| develop | English | 504 | 15.3 k | 2.30 k | 15 | 30.4 | 3 |
| | Spanish | | 15.4 k | 2.75 k | 25 | 30.5 | 3 |
| test | English | 1094 | 26.88 k | 4.0 k | 112 | 24.6 | 2 |
| | Spanish | 840 | 22.73 k | 4.1 k | 46 | 27.1 | 2 |

**Table 3.14:** *European Parliament English-Spanish corpus version 1 statistics.*

Due to corpus improvements and updates during the last two years, up to three slightly different versions of this corpus will be used in the experiments reported throughout this dissertation[9]. For this reason, and for clarification purposes, this corpus will be referred to as version 1 (ie. '**EuParl ver1**').

#### 3.3.1.1 History length

In Table 3.15 the impact of changing the history length (ie. the order of the bilingual $N$-gram model) on translation quality is shown. Regarding model size, the right-most column shows the number of $n$-grams included in the model (in Millions).

| history length | | BLEU | mWER | NIST | PER | 1grams | 2grams | 3grams | 4grams | 5grams |
|---|---|---|---|---|---|---|---|---|---|---|
| Eng→Spa | n=2 | 0.4008 | 46.77 | 8.847 | 36.39 | 2.45 | 7.14 | – | – | – |
| | n=3 | 0.4212 | 45.40 | 9.043 | 35.29 | 2.45 | 7.14 | 11.81 | – | – |
| | n=4 | 0.4236 | 45.27 | 9.079 | 35.19 | 2.45 | 7.14 | 11.81 | 13.74 | – |
| | n=5 | 0.4220 | 45.40 | 9.052 | 35.46 | 2.45 | 7.14 | 11.81 | 13.74 | 14.00 |
| Spa→Eng | n=2 | 0.4449 | 42.13 | 9.519 | 31.99 | 2.46 | 7.02 | – | – | – |
| | n=3 | 0.4649 | 40.64 | 9.677 | 31.14 | 2.46 | 7.02 | 11.78 | – | – |
| | n=4 | 0.4674 | 40.40 | 9.694 | 31.09 | 2.46 | 7.02 | 11.78 | 13.96 | – |
| | n=5 | 0.4650 | 40.48 | 9.694 | 31.18 | 2.46 | 7.02 | 11.78 | 13.96 | 14.40 |

**Table 3.15:** *Effect of translation model history length n on translation quality and model size.*

As it can be seen, in both directions the trigram model ($n$=3) is always significantly better than the bigram model and suffices to obtain a high performance. In fact, increasing the model

[9]However slight, the differences between versions will be commented on when each new version is presented.

above $n$=3 does not produce any further significant boost in performance, especially in the Eng→Spa direction.

Obviously, regarding model size, increasing the history length has a direct correlation with storage and computational costs, as not only have these new probabilities to be estimated and stored at training time, but also they need to be searched at decoding time. History length is associated with the size of the data structures needed by an $N$gram-based decoder in order to produce translations. Therefore, trigrams seem to be the best option, at least for this task.

#### 3.3.1.2 Pruning strategies

As we have seen in the previous section, the bilingual translation model becomes huge when augmenting the history length, or simply when increasing training material. Pruning strategies are therefore needed.

Pruning can be defined as any technique taking a decision on discarding certain training material as useless. This decision must be a hard one, ie. taken before having any knowledge on the test set. Usually, pruning reveals the classical trade-off between efficiency and performance. Whereas a strong pruning produces small-sized models which can be used in a much more efficient way, performance usually falls off. In addition to that, a balanced degree of pruning can sometimes make a more efficient model at no performance cost (or even at an improvement in performance).

Several pruning strategies can be devised, but here we divided them in two types (which can, of course, be combined as well):

- $N$-gram pruning. A classic pruning strategy in language modelling is to perform the $N$-gram modelling, but according to restrictions affecting the $N$-gram counts extracted from training data, so that for instance a certain tuple may participate in a certain trigram while not participating in any bigram.

- Tuple pruning. This approach refers to any technique which takes a hard decision on tuple vocabulary, taking an *a priori* decision on which tuples are allowed to belong to the tuples vocabulary, and which are not. According to this, Tuple pruning is a special case of $N$-gram pruning in which, for a given discarded tuple, all $N$-gram counts in which this tuple participates are set to 0 (for all $N$).

##### $N$-gram pruning

During language model estimation, one can assume that all $N$-grams occurring less than a

certain number of times are discounted to zero. This strategy is very often used for large $N$ values, and has as a consequence smaller models with improved performance.

The idea behind this is to consider long $N$-grams occurring just once in the training material as important as all those $N$-grams which do not occur at all (and which are taken into account via smoothing the model probabilities, as discussed in §3.3.1.3).

If we define a threshold $t_n$ for each history length $n$, translation performance and model size obtained are shown in Table 3.16 for different history lengths and threshold values. Note that rows with all $t_n$ set to 1 (coloured in grey) correspond to no pruning at all (from Table 3.15).

| hist. | $t_2$ | $t_3$ | $t_4$ | $t_5$ | BLEU | mWER | NIST | PER | 1grams | 2grams | 3grams | 4grams | 5grams |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Eng→Spa | | | | | | | | | | | | | |
| n=3 | 1 | 1 | - | - | 0.4212 | 45.40 | 9.043 | 35.29 | 2.45 | 7.14 | 11.81 | – | – |
| | 1 | 2 | - | - | 0.4300 | 44.20 | 9.203 | 34.50 | 2.45 | 7.14 | 1.72 | – | – |
| | 2 | 2 | - | - | 0.4283 | 43.78 | 9.221 | 34.52 | 2.46 | 1.29 | 1.72 | – | – |
| | 2 | 3 | - | - | 0.4255 | 44.03 | 9.172 | 34.73 | 2.46 | 0.79 | 0.51 | – | – |
| n=4 | 1 | 1 | 1 | - | 0.4236 | 45.27 | 9.079 | 35.19 | 2.45 | 7.14 | 11.81 | 13.74 | – |
| | 1 | 2 | 2 | - | 0.4309 | 44.23 | 9.212 | 34.68 | 2.45 | 7.14 | 1.57 | 1.47 | – |
| | 2 | 2 | 2 | - | 0.4272 | 43.80 | 9.204 | 34.71 | 2.45 | 1.29 | 1.57 | 1.47 | – |
| n=5 | 1 | 1 | 1 | 1 | 0.4220 | 45.40 | 9.052 | 35.46 | 2.45 | 7.14 | 11.81 | 13.74 | 14.00 |
| | 1 | 2 | 2 | 2 | 0.4302 | 44.46 | 9.197 | 34.88 | 2.45 | 7.14 | 1.57 | 1.35 | 1.22 |
| | 2 | 2 | 2 | 2 | 0.4257 | 44.03 | 9.183 | 34.91 | 2.45 | 1.29 | 1.57 | 1.35 | 1.22 |
| Spa→Eng | | | | | | | | | | | | | |
| n=3 | 1 | 1 | - | - | 0.4649 | 40.64 | 9.677 | 31.14 | 2.46 | 7.02 | 11.78 | – | – |
| | 1 | 2 | - | - | 0.4732 | 40.07 | 9.812 | 30.69 | 2.46 | 7.02 | 1.77 | – | – |
| | 2 | 2 | - | - | 0.4795 | 39.00 | 9.940 | 30.09 | 2.46 | 1.26 | 1.77 | – | – |
| | 2 | 3 | - | - | 0.4791 | 39.02 | 9.952 | 29.93 | 2.46 | 0.77 | 0.56 | – | – |
| n=4 | 1 | 1 | 1 | - | 0.4674 | 40.40 | 9.694 | 31.09 | 2.46 | 7.02 | 11.78 | 13.96 | – |
| | 1 | 2 | 2 | - | 0.4790 | 39.51 | 9.881 | 30.24 | 2.46 | 7.02 | 1.61 | 1.56 | – |
| | 2 | 2 | 2 | - | 0.4802 | 38.89 | 9.953 | 29.96 | 2.46 | 1.26 | 1.61 | 1.56 | – |
| n=5 | 1 | 1 | 1 | 1 | 0.4650 | 40.48 | 9.694 | 31.18 | 2.46 | 7.02 | 11.78 | 13.96 | 14.40 |
| | 1 | 2 | 2 | 2 | 0.4774 | 39.57 | 9.866 | 30.35 | 2.46 | 7.02 | 1.61 | 1.44 | 1.30 |
| | 2 | 2 | 2 | 2 | 0.4793 | 38.96 | 9.943 | 30.09 | 2.46 | 1.26 | 1.61 | 1.44 | 1.30 |

**Table 3.16:** *Effect of $N$-gram pruning strategies on translation quality and model size.*

As seen in the table, setting thresholds $t_n$ to 2 produces a very important model size reduction, whereas translation performance keeps stable or even goes up. This is a very relevant result, since it enables efficient data processing times with good quality scores.

Hereto-after this pruning will be assumed in reported experiments, particularly setting the thresholds to $t_2 = 1$ and $t_{3+} = 2$ (3+ refers to trigram and higher $n$-grams, in case the model includes them).

#### Tuple pruning

In our implementation, in addition to the previous pruning, the tuple vocabulary is pruned by using histogram counts. This pruning, called *tuple n-best*, is performed by keeping the *tnb* most frequent tuples for each tuple source side. This way only the *tnb* most-frequent translations of each tuple source part are allowed (so long as more than *tnb* different target parts exist for the source part).

Typically, the optimal value of this pruning parameter will depend on data and should be adjusted empirically for each considered translation task. For our studied Spanish↔English task, the impact of this pruning on translation quality and model size is shown in Table 3.17 for a history length of 3 (number of unigrams, bigrams and trigrams are expressed in Millions).

| | pruning | BLEU | mWER | NIST | PER | unigrams | bigrams | trigrams |
|---|---|---|---|---|---|---|---|---|
| Eng→Spa | none | 0.4300 | 44.20 | 9.203 | 34.50 | 2.45 | 7.14 | 1.72 |
| | tnb=30 | 0.4276 | 44.46 | 9.175 | 34.71 | 2.02 | 6.09 | 1.75 |
| | tnb=20 | 0.4273 | 44.30 | 9.181 | 34.68 | 1.96 | 5.84 | 1.73 |
| | tnb=10 | 0.4225 | 44.63 | 9.128 | 34.95 | 1.84 | 5.34 | 1.68 |
| Spa→Eng | none | 0.4732 | 40.07 | 9.812 | 30.69 | 2.46 | 7.02 | 1.77 |
| | tnb=30 | 0.4747 | 39.66 | 9.847 | 30.46 | 2.11 | 6.23 | 1.81 |
| | tnb=20 | 0.4745 | 39.71 | 9.839 | 30.51 | 2.04 | 6.01 | 1.80 |
| | tnb=10 | 0.4773 | 39.40 | 9.889 | 30.20 | 1.92 | 5.57 | 1.76 |

**Table 3.17:** *Effect of tuple n-best pruning on translation model quality and size.*

Notice that such a pruning, since performed before computing tuple n-gram probabilities, has a direct incidence on the translation model probabilities, and then on the overall system performance.

However, as will be discussed in §4.3.4, this pruning parameter is not independent from the presence of feature functions which complement the translation model. In the case of the European Parliament data under consideration, pruning parameter values of $tnb = 20$ and $tnb = 30$ for Spanish-to-English and English-to-Spanish, respectively, prove to be the most adequate.

### 3.3.1.3 Smoothing the bilingual model

It has been shown in ASR research field that solely relying on a maximum likelihood estimate may not be the best option when performing language modelling, especially when it comes to unfrequent events. Consider a trigram which occurs just once in the training material. Taking into account that the number of possible trigrams is very huge and that only a few will occur in the training data, it may be unreasonable to perform maximum likelihood estimation for these

events, as all unseen events will have a zero probability.

Smoothing refers to all techniques which redistribute probability from seen events to unseen events. This is usually done by removing a certain amount of probability mass from seen events and reserve it for the wide range of unseen combinations of words.

Several alternative smoothing techniques have been applied for language modelling. A very thorough comparative study of these techniques in the context of language modelling was published in [Che96, Che98]. In this work the authors review each smoothing technique and contrast their performance, for different training corpus size, in terms of perplexity. This measure is related to the probability that the model assigns to a test data, and is defined as in equation 3.8:

$$PP_p(T) = 2^{H_p(T)} \tag{3.8}$$

where $H_p(T)$ is the cross-entropy of the language model on data $T$, defined as:

$$H_p(T) = -\frac{1}{W_T} \log_2 p(T) \tag{3.9}$$

and where $W_T$ is the number of words in the test data $T$.

Smoothing techniques investigated in [Che98] include Additive smoothing, the Good-Turing estimate, Jelinek-Mercer smoothing, Katz smoothing, Witten-Bell smoothing, Absolute discounting and Kneser-Ney smoothing. Furthermore, a slight modification of Kneser-Ney smoothing is also presented, rendering the best performance. Being out of the scope of this Ph.D. research work, all details regarding each smoothing formulation and implementation are omitted here. However, they can be found in [Che98].

Still, due to the bilingual nature of our translation model, a comparative experiment was conducted in order to assess which smoothing was most suited for the bilingual $N$-gram modelling. Instead of defining a perplexity score for the translation task, we opted for evaluating translation performance directly.

Results are shown in Table 3.18 for the EuParl version 2 task (see Table 3.24 for statistics, which are basically equivalent to version 1). History length is set to 3-grams, $N$-gram pruning is assumed to be $(t_2, t_3)$=(1,2) and tuple pruning parameter *tnb* is 20 (Spa→Eng) and 30 (Eng→Spa).

As it can be seen, the modified Kneser-Ney presented in [Che98] achieves the best translation scores in both translation directions. It must be noted that all internal experiments across various tasks have followed the same tendency. All in all, this correlates with the results of [Che98], thus

| | Spa→Eng | | Eng→Spa | |
|---|---|---|---|---|
| smoothing technique | BLEU | mWER | BLEU | mWER |
| Good-Turing | 0.4629 | 39.05 | 0.4164 | 44.27 |
| Modif. Kneser-Ney | 0.4645 | 39.11 | 0.4171 | 44.39 |
| Modif. Kneser-Ney +interpolation | **0.4698** | **38.73** | **0.4221** | **43.60** |
| Natural Discounting | 0.4494 | 40.31 | 0.4026 | 45.66 |
| Kneser-Ney | 0.4633 | 39.26 | 0.4147 | 44.71 |
| Witten-Bell | 0.4544 | 39.93 | 0.4055 | 45.33 |
| Witten-Bell +interpolation | 0.4556 | 39.89 | 0.4109 | 44.85 |

**Table 3.18:** *Obtained scores estimating the bilingual model with different smoothing techniques (EuParl ver 2 task Eng→Spa).*

leading to the conclusion that the bilingual model behaves similarly to a standard (monolingual) language model when it comes to smoothing techniques.

Given that modified Kneser-Ney smoothing consistently performs better, it was assumed in previous experiments on history length and pruning issues, reported in §3.3.1.1 and §3.3.1.2 respectively.

### 3.3.2 Case study: the Catalan-Spanish task

To conclude this section devoted to the $N$-gram implementation of the tuple bilingual model, here we review a case study of its application to a Catalan↔Spanish large-vocabulary task ([Abe06]).

Belonging to the same family of languages, being very much influenced and sharing in many cases the same speakers, Spanish and Catalan languages exhibit a morphological and grammatical similarity (including strong monotonicity in word order) favouring the deployment of the presented bilingual $N$-gram model, thus serving as test bank to assess the validity of the approach when initial conditions are quasi-optimal.

| Cat News | | sent. | words | vcb/OOVs | max.len. | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | Catalan | 2.18 M | 43.28 M | 390.2 k | 100 | 19.9 | 1 |
| | Spanish | | 41.51 M | 397.4 k | 100 | 19.1 | |
| test | Catalan | 2 k | 48,3 k | 217 | 98 | 24.2 | 1 |
| | Spanish | | 46,4 k | 200 | 98 | 23.2 | 1 |

**Table 3.19:** *Catalan–Spanish newspaper parallel corpus statistics.*

The corpus used for this experimentation, named **CatNews**, is a collection of web-mined news from a general bilingual newspaper published in Catalonia in the period of 1998 to 2003. Data collection and preprocessing is reported in [Fri03, FM05, Abe06]. Table 3.19 shows the main statistics of the corpus. As observed from the high vocabulary figures, this is largely an

open-domain task.

Results obtained by the SMT system are shown in Table 3.20. As it can be seen from the obtained evaluation scores, the translation quality improvement with respect to all other translation tasks is noteworthy. Even with a single manual reference, error is much smaller.

| | BLEU | mWER | NIST | PER |
|---|---|---|---|---|
| Spa→Cat | 0.8421 | 9.88 | 13.93 | 8.74 |
| Cat→Spa | 0.8334 | 10.08 | 13.73 | 8.86 |

**Table 3.20:** *Translation results for Catalan–Spanish newspaper task.*

Besides, and interestingly, whereas the $N$-gram translation model takes advantage of additional feature models in most translation tasks (as will be studied in chapter 4), this behaviour is not observed in the Catalan-Spanish tasks, where simply the ngram translation model suffices to generate high-quality translations. These facts reflect the grammatical similarity between Spanish and Catalan, which allows for a well-estimated model (nearly free from sparseness problems) even with large vocabulary sizes.

A thorough human error analysis of this translation task was carried out in [Abe06], concluding that around 80% of the test sentences were entirely correct, ie. had no translation error at all. In most cases, synonyms accounted for differences between translated test and reference translations. Among errors, unseen words, agreement issues and omitted words were the most frequent types.

### A web demonstrator

With the purpose of disseminating research activity results to a broader public, and given the positive results achieved by the system in this task, a Catalan↔Spanish web demonstrator of this system was implemented at UPC[10].

The result was **N-II**, which is accessible at http://www.n-ii.org , and allows users to obtain high-quality translations online at any time. A screenshot of this demo with a translation example is shown in Figure 3.4.

### Bridging to Catalan–English

Another application of this high-quality Catalan↔Spanish system is related to language portability, by building a Catalan↔English SMT system without parallel corpus.

Regarding the parallel corpora necessary to build statistical machine translation systems, it

[10]By Antonio Abellán, Patrik Lambert, Josep Maria Crego, José B. Mariño and Adrià de Gispert

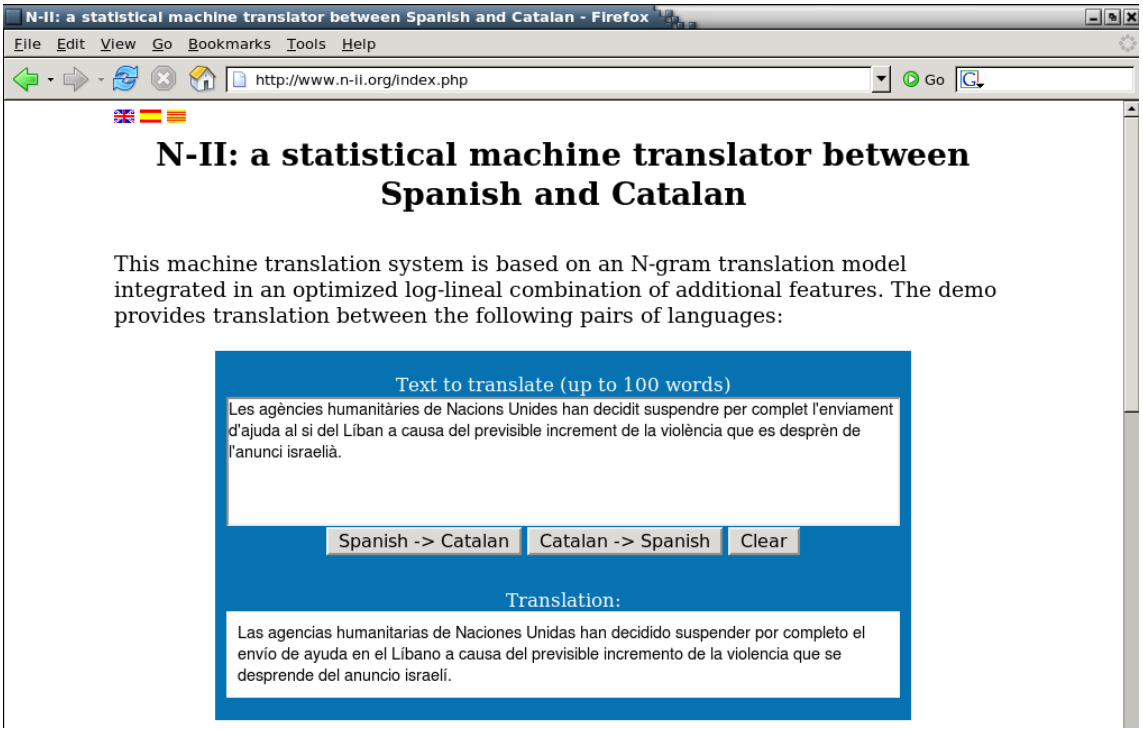



**Figure 3.4:** Screenshot of www.n-ii.org, the Catalan↔Spanish online demostrator.

becomes nearly impossible to find freely-available large-vocabulary data in Catalan and other languages (except for Spanish). Since many more parallel corpora in Spanish are available, one can reasonably think of using Spanish as a bridge towards statistical machine translation from Catalan to other languages, and vice versa.

Further details on these experiments are reported in the following publication:

- [Gis06b] A. de Gispert and J.B. Mariño, "Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish," in *Proceedings of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages, SALTMIL'06*, pps. 65–68, May 2006.

To conclude, [Gis06b] also presents a result when translating from Spanish to Catalan in the European Parliament task (version 2), as shown in Table 3.21. In this case, the manual reference available is adapted to the translated output (ie. it is human-targeted, a concept introduced in §2.7.2) and does not suffer from synonymy, rendering a quality score of the system which is not pessimistic.

| | BLEU | mWER | NIST | PER |
|---|---|---|---|---|
| Spa→Cat | 0.9345 | 3.79 | 13.73 | 3.49 |

**Table 3.21:** *Translation results for Spanish→Catalan EuParl ver2 task.*

## 3.4 The Tuple as Translation Unit

Apart from this structural definition, two additional issues regarding the $N$-gram translation model must be considered (*embedded words* and *NULL-to-one tuples*), which are discussed next.

### 3.4.1 Embedded words

A first issue is connected with the fact that an (possibly) important amount of single-word translation probabilities are often left out of the model. This happens for all those words that appear always embedded into tuples containing two or more source words. For example, consider source word 'D' from Figure 3.3, which is extracted together with 'E' and 'F' into the tuple 'DEF→XY' (tuple type III).

When examining the word-aligned training material, if all occurrences of word 'D' present a similar pattern, ie. the word is forced to share a tuple with other source words, no translation probability for this single word will ever be estimated by the model. In this case, we say that the word is *embedded*. Clearly, this problem is aggravated if the pair of languages involved has a non-monotone alignment, leading to very long tuples (and a higher number of embedded words).

To address this problem, we build up a dictionary of translations for embedded words from the most accurate word alignment available (usually, the intersection of source-to-target and target-to-source alignments). For a certain embedded word $f_j$ and a given word alignment, we look for the target words $e_i...e_{i+K}$ that are most frequently aligned to $f_j$ with these two conditions:

1. Target words $e_i...e_{i+K}$ are consecutive in the target sentence.
2. Target words are aligned *only* to $f_j$ or to null.

This way, we build up a statistical dictionary independently of the non-monotonicity of the word alignment. The entries of the dictionary are included as unigrams in the bilingual model.

Under a Spanish↔English framework, the embedded words problem is practically irrelevant (affecting a list of singleton words[11]), due to the relative monotonicity of the pair, basically turning into a question of completeness, so that the system is able to produce *a* translation at

[11] Words occurring only once in the training parallel corpus.

all. However, this dictionary solution, which was presented in [Gis04a], forces the model to fall back to an incontextual word-based translation for embedded words, which can be negative for language pairs with strong reordering needs (such as Chinese and English).

### 3.4.2 Tuple segmentation

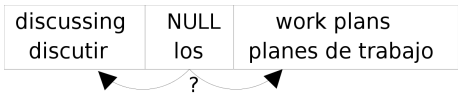
As we saw in §3.2.2.1, tuple extraction is the process by which tuples are generated from word alignment. From a conceptual point of view, the final goal is to obtain a set of tuples which have the highest 're-usability' capacity, ie. that they can be recycled in order to produce valid translations in certain unseen situations, the more the better. For example, if we segment the bilingual sentence '`through the vote on Thursday # mediante la votación del jueves`' into the following tuples:

| | | |
|---|---|---|
| `through the`<br>`mediante` | `vote`<br>`la` | `on Thursday`<br>`votación del jueves` |

it is intuitive that these tuples will only be useful to translate exactly the same sentence (no re-use), as we can expect that in very few situations `votación del jueves` will be a valid translation of `on Thursday`. The intuition tells us that a more 'usable' segmentation (ie. leading to a less entropic model) would be something like:

| | | | | |
|---|---|---|---|---|
| `through`<br>`mediante` | `the`<br>`la` | `vote`<br>`votación` | `on`<br>`del` | `Thursday`<br>`jueves` |

On the other hand, the tuple definition (according to the constraints laid down in §3.2.2.1) defines a unique set of tuples except in one situation, which is whenever the resulting tuple contains no source word (NULL-source tuple). In order to re-use these units in decoding new sentences, the search should allow for no input word to generate units, and this is not the case. Therefore, these units cannot be allowed, and a certain hard decision must be taken as for tuple segmentation in these cases, as in the following example:




#### 3.4.2.1 Segmentation strategies

According to the aforementioned conceptual framework of $N$-gram translation model, it seems clear that the ideal tuple segmentation strategy should take a global decision based on the segmentation for all other NULL-source cases, attempting to obtain that set of tuples and $N$-grams which better represented the unseen universe of events, meaning the one with less entropy. However, no feasible algorithm can perform that calculation in a reasonable time given current

computational power, as this would involve a whole model re-estimation for each particular segmentation alternative.

#### Deterministic always NEXT

A very pragmatic and simple approach to take this decision is to always join the target words involved in NULL links (NULL-linked words) to the following tuple, if there is any (otherwise, to the previous one). This approach, first introduced in [Gis04a], was used in all $X$-gram experiments reported in §3.2. Apart from simplicity and extreme efficiency, we do not observe any other advantage of this approach, which on the other hand does not follow any linguistic or statistical criterion.

#### IBM model 1 weight

Being independent of word position, IBM model 1 probabilities provide a probabilistic lexicon between pairs of word of each language (see §2.1.1 for details on these models). This information can be used to weight and compare the resulting tuples from two competing segmentations, as introduced in [Cre05a]. The weight for each tuple is defined as:

$$\frac{1}{I}\prod_{j=1}^{J}\sum_{i=0}^{I} p_{IBM1}(t^i|s^j)p_{IBM1'}(s^j|t^i) \tag{3.10}$$

where $s$ and $t$ represent source and target sides of a tuple, $I$ and $J$ their respective number of words and $IBM1'$ stands for the IBM model 1 estimated in the opposite direction.

While this approach is appealing in that it takes bilingual information into account, observation of these situations leads to a different conclusion. Many NULL-linked words represent articles, prepositions, conjunctions and other particles whose main function is to ensure the grammatical correctness of a sentence, complementing other more informative words. Therefore, their probabilities to translate to another word are not very meaningful.

#### Entropy of the POS distribution

Alternatively, from a linguistic point of view, one can regard the tuple segmentation problem around source NULLs as a monolingual decision related to whether a given target word is more connected to its preceding or following word.

Intuitively, we can expect that a good criterion to perform tuple segmentation lays in preserving grammatically-connected phrases (such as, for instance, articles together with the noun they precede) in the same tuple, as this may probably lead to a simplification of the translation

task. On the contrary, splitting linguistic units into separate tuples will probably lead to a tuple vocabulary increase and a higher sparseness, producing a worse (and more entropic) $N$-gram translation model.

In this direction, we proposed to take the segmentation decision according to the entropy of the forward and backward Part-Of-Speech (POS) distributions, which was defined conditioned to context. In detail, given the following 3-tuple sequence:

$$\begin{array}{ccc} < ...s_j > & \text{NULL} & < s_{j+1}... > \\ | & | & | \\ < ...t_{i-1} > & t_i & < t_{i+1}... > \end{array}$$

where $s_j$ means word in position $j$ in source sentence, and equivalently $t_i$ means word in position $i$ in target sentence, one can define a 'forward' entropy of the POS distribution in position $i+1$ given $(t_{i-1}, t_i)$ as in equation 3.11:

$$H^f_{POS} = -\sum_{POS} p^f_{POS} \log p^f_{POS} \tag{3.11}$$

where

$$p^f_{POS} = \frac{N(t_{i-1}, t_i, POS_{i+1})}{\sum_{POS'} N(t_{i-1}, t_i, POS'_{i+1})} \tag{3.12}$$

is the probability of observing a certain Part-Of-Speech *following* the sequence of words defined by $t_i$ and $t_{i+1}$, estimated by relative frequency.

Equivalently, one can define a 'backward' entropy of the POS distribution in position $i-1$ given $(t_i, t_{i+1})$ as in equation 3.13:

$$H^b_{POS} = -\sum_{POS} p^b_{POS} \log p^b_{POS} \tag{3.13}$$

where

$$p^b_{POS} = \frac{N(POS_{i-1}, t_i, t_{i+1})}{\sum_{POS'} N(POS'_{i-1}, t_i, t_{i+1})} \tag{3.14}$$

is the probability of observing a certain Part-Of-Speech *preceding* the sequence of words defined by $t_{i-1}$ and $t_i$.

Then, we can take a tuple segmentation decision favouring the most POS-entropic case. The rationale behind this is that, if $H^f_{POS} > H^b_{POS}$, we have observed the first sequence of

words comprised of $(t_{i-1}, t_i)$ in more grammatically different situations than the latter sequence comprised of $(t_i, t_{i+1})$. Therefore, we can induce that $t_{i-1}$ and $t_i$ tend to be more often connected than $t_i$ and $t_{i+1}$, and should belong to the same translation tuple. Analogously, one can conclude the contrary if $H^f_{POS} < H^b_{POS}$.

| hay | NULL | ninguna |
|---|---|---|
| \| | \| | \| |
| there | are | no |
| $H^f_{POS}(there, are, ***) = 0.83$ | | |
| $H^b_{POS}(***, are, no) = 0.62$ | | |

**Table 3.22:** *Example of $H^f_{POS}$ and $H^b_{POS}$ entropies.*

To illustrate this idea, consider the example in Table 3.22, where 'there are' proves to be more connected than 'are no', thus being linked in the same translation unit.

| Source | Target | alwaysNEXT | IBM1weight | POSEntropy |
|---|---|---|---|---|
| We | Nos | We — Nos | | |
| are pleased | alegramos | are pleased — alegramos | | |
| at | NULL | at—NULL | at—de | at—de |
| **NULL** | **de** | this—de esta | | |
| this | esta | | this—esta | this—esta |
| visit | visita | visit — visita | | |
| which | que | which—que | which—que se | which—que |
| **NULL** | **se** | reflects—se enmarca | | |
| reflects | enmarca | | reflects—enmarca | reflects—se enmarca en |
| **NULL** | **en** | the—en la | the—en la | |
| the | la | | | the—la |
| cooperation | cooperación | cooperation — cooperación | | |
| between | entre | between — entre | | |
| parliaments | parlamentos | parliaments — parlamentos | | |
| in | NULL | in—NULL | in—NULL | in—de |
| **NULL** | **de** | the—de la | the—de la | |
| the | la | | | the—la |
| Union | Unión | Union — Unión | | |

**Table 3.23:** *Example of segmentation decisions around NULL-linked words taken by each criterion in a real-life English-to-Spanish sentence.*

While this is a monolingual decision on the target language morphology, it proves very consistent with human intuition. To sum up, Table 3.23 shows an English–Spanish sentence example, where original links (from word alignment) are shown in the first two columns, and the segmentation decisions taken by each strategy are shown in the right-most three columns. As it can be seen, being linguistically-guided, the POS entropy approach is much more correlated

with human intuition.

#### 3.4.2.2 Comparative results

In order to compare each segmentation strategy and evaluate its impact on translation quality, experiments were carried out using two parallel corpora, differing in language pair and corpus size. On the one hand, we used a Spanish–English large-vocabulary corpus (European Parliament Proceedings Corpus version 2[12]), and on the other hand, an Arabic-to-English small-vocabulary corpus, containing a small part of the Basic Travel Expressions Corpus.

| EuParl ver2 | | sent. | words | vcb | OOVs | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | English | 1.22 M | 33.37 M | 104.8 k | - | 27.3 | 1 |
| | Spanish | | 34.96 M | 151.4 k | - | 28.6 | |
| develop | English | 504 | 15.3 k | 2.30 k | 20 | 30.4 | 3 |
| | Spanish | | 15.4 k | 2.74 k | 22 | 30.6 | 3 |
| test | English | 1094 | 26.88 k | 4.0 k | 113 | 24.6 | 2 |
| | Spanish | 840 | 22.75 k | 4.1 k | 44 | 27.1 | 2 |

**Table 3.24:** *European Parliament English-Spanish corpus version 2 statistics.*

The main statistics of each corpus, including number of sentences, running words, vocabulary size, out-of-vocabulary words (for development and test sets), average sentence length and number of human references, are shown in Tables 3.24 and 3.25, respectively.

| BTEC-20k | | sent. | words | vcb | OOVs | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | Arabic | 20 k | 180.5 k | 16.0 k | - | 9.0 | 1 |
| | English | | 189.2 k | 7.2 k | - | 9.5 | |
| develop | Arabic | 506 | 3.63 k | 1.18 k | 196 | 7.2 | 16 |
| test | Arabic | 1006 | 7.22 k | 1.9 k | 356 | 7.2 | 16 |

**Table 3.25:** *Travel Expressions Arabic-English 20k-corpus statistics.*

In both bases, English was tagged using *TnT*[13] tagger ([Bra00]), and Spanish using *FreeLing*[14] analysis toolkit ([Car04]).

For the train set, Table 3.26 also shows the number of running tuples extracted from the word alignment used[15], together with the percentage of tuples with NULL in one of its sides. As expected, this percentage is higher in English side (14.5%), given that Spanish contains more running words that have no direct correspondence in English.

[12] Version 2 is nearly identical to version 1, whose statistics are shown in Table 3.14, except for Spanish preprocessing, which includes an improved tokenisation tool

[13] Available at www.coli.uni-saarland.de/~thorsten/tnt

[14] Available at http://garraf.epsevg.upc.es/freeling

[15] Union of src→trg and trg→src alignments, obtained with GIZA++ tool.

| | | EuParl ver2 | | BTEC-20k | |
|---|---|---|---|---|---|
| | | Spanish | English | Arabic | English |
| train | running tuples | 20,032 k | | 122,176 | |
| | one-to-NULL tuples | 11.7% | 14.5% | 7.0% | 7.2% |

**Table 3.26:** Tuple statistics for each two parallel corpora used.

As the $N$-gram model does not allow for tuples with empty source language, in Eng→Spa a segmentation decision on these 14.5% tuples must be taken, whereas in the opposite direction, only 11.7% of the tuples must be re-segmented. Therefore, we can expect a bigger impact of segmentation strategies in the Eng→Spa direction.

| | | BLEU | mWER | NIST |
|---|---|---|---|---|
| Eng→Spa | alwaysNEXT | 0.4215 | 43.98 | 9.22 |
| | IBM1weight | 0.4221 | 43.60 | 9.19 |
| | POSentropy | **0.4325** | **43.48** | **9.30** |
| | trgNULL | 0.4249 | 44.47 | 9.21 |
| | trgNULLpos | 0.4313 | 43.75 | 9.29 |
| Spa→Eng | alwaysNEXT | 0.4661 | 39.37 | 9.86 |
| | IBM1weight | 0.4698 | 38.73 | 9.91 |
| | POSentropy | **0.4756** | **38.64** | **9.95** |
| | trgNULL | 0.4728 | 39.23 | 9.91 |
| | trgNULLpos | 0.4733 | 38.78 | 9.93 |
| Ara→Eng | alwaysNEXT | 0.3684 | 41.80 | 7.16 |
| | IBM1weight | 0.3656 | 41.94 | 7.14 |
| | POSentropy | **0.3691** | 41.91 | **7.17** |

**Table 3.27:** *Translation model performance for each segmentation strategy.*

A comparison of the translation model performance for each task is shown in rows named 'alwaysNEXT', 'IBM1weight' and 'POSentropy' in Table 3.27, referring to each segmentation strategy discussed in §3.4.2.1.

Regarding the large-vocabulary tasks, the proposed linguistically-guided segmentation outperforms all other strategies significantly, especially in the Eng→Spa direction. This result is consistent with the fact that Spanish is a more word-generative language than English, and therefore, more NULLs are found in the English side of extracted tuples.

Even though the impact of changing the segmentation criterion when translating into English is smaller, the improvement of the POSentropy approach is significant. In the small-vocabulary Ara→Eng task differences are less significant, in correlation with the fact that *only* 7% of tuples contain NULLs in Arabic side, compared to the 14% of Eng→Spa task, as shown in Table 3.26.

Remarkably, whereas IBM1weight provides better results in large-vocabulary tasks than the

alwaysNEXT criterion, the result is opposite in the small-vocabulary Ara→Eng task. On the other hand, the POSentropy approach proves to be more general and robust to a task change, achieving best performance in all tasks.

#### 3.4.2.3 Removing NULLs in target

Whereas the segmentation decision is required when a target word is unlinked (or linked to NULL), this is not so when the unlinked word is in the source target, in which case these units are allowed in the tuple vocabulary for Ngram estimation (in a clear difference to the phrase-based approach [Zen02], where no NULL tokens exist and segmentation decisions are put off to decoding time as all possible units are extracted in training as unigrams).

However, one can think of applying the same criterion to remove NULLs in the target side of tuples, possibly addressing omission errors in translation. Aiming at evaluating the impact of this decision, we have also applied the POS entropy strategy to segment tuples with unlinked source words.

Table 3.27 presents results when applying this segmentation criterion (POSentropy) to avoid NULLs in the tuples *target* side, as shown in rows named 'trgNULL' and 'trgNULLpos' for large-vocabulary Spa→Eng task. The first refers to applying the criterion to all tuples, whereas the latter to only applying it when the tuple contains a POS of a Noun, Adjective or Verb. The objective of this is to minimise omission errors by preventing tuples with content words in source side and NULL in target to belong to the model dictionary.

However, results show that none of these techniques is beneficial for translation quality. It seems clear that, in contrast to NULLs in the tuple source side, NULLs in the target side are a useful mechanism for the Ngram model to find good contexts and significantly increase performance regardless of the translation direction. This conclusion holds even when we allow tuples with content words in source side and NULL in target. These results correlate with the fact that tuple type II is not adequate for a tuple $N$-gram model, as discussed in §3.2.3.

#### 3.4.2.4 Translation Ngrams study

To better understand these results, Table 3.28 shows the tuple vocabulary obtained in training for each segmentation (tup vcb), and relevant statistics of translated output, namely the percentage of test tuples *seen* as 1-grams, 2-grams and 3-grams during training, the average tuple length obtained (measuring source and target sides separately) and the number of tuples with NULL in target (in the translated output).

Regarding tuple vocabulary size, the alwaysNEXT criterion produces the biggest vocabulary in training when compared to POSentropy and IBM1weight, which produces the smallest. When

| | | tup vcb | % 1–2–3grams | tup len | NULLs |
|---|---|---|---|---|---|
| Eng→Spa | alwaysNEXT | 2110085 | 17.6 – 44.4 – 38.0 | 1.157-1.096 | 3119 |
| | IBM1weight | 2035523 | 18.0 – 44.7 – 37.3 | 1.157-1.090 | 2466 |
| | POSentropy | 2084640 | 17.8 – 44.3 – 37.9 | 1.156-1.106 | 2282 |
| | trgNULL | 2347743 | 23.2 – 45.1 – 31.7 | 1.253-1.190 | 0 |
| | trgNULLpos | 2178470 | 19.0 – 44.5 – 36.5 | 1.180-1.139 | 1625 |
| Spa→Eeng | alwaysNEXT | 2149595 | 14.1 – 41.5 – 44.4 | 1.135-1.064 | 2761 |
| | IBM1weight | 2080171 | 14.2 – 41.4 – 44.4 | 1.131-1.054 | 2318 |
| | POSentropy | 2109351 | 14.2 – 41.5 – 44.3 | 1.134-1.064 | 2194 |
| | trgNULL | 2421446 | 19.9 – 44.1 – 36.0 | 1.260-1.224 | 0 |
| | trgNULLpos | 2164076 | 14.7 – 41.6 – 43.7 | 1.143-1.075 | 1977 |

**Table 3.28:** *Tuple vocabulary and N-gram translation statistics for each segmentation strategy.*

removing the target NULLs, the vocabulary size is significantly increased.

In Eng→Spa, we observe that translation with alwaysNEXT and POSentropy segmentation criteria tend to use more 3grams than IBM1weight, which can be explained by their consistency in taking segmentation decisions (they invariably take the same decision given the target words involved), whereas IBM1weight depends on source and target words and is more variable.

However, using more 3grams is not directly correlated with translation scores, and the number of tuples to NULL needs to be taken into account. The high number of tuples to NULL for the alwaysNEXT criterion is outstanding, and tells us that translation is indeed achieving many 3grams by catenating sequences to NULL, which do not necessarily achieve better performance. In the case of IBM1weight and especially of POSentropy, the number of tuples with NULL in target is strongly reduced.

Whereas this appears to be positive for translation performance, when completely or partially removing NULLs in target (trgNULL and trgNULLpos), average tuple length increases, not only in the source side but also in the target side, and the model looses tuple context and falls much more often to 1gram. Apparently, this has a negative effect in translation quality.

Therefore, we can conclude that the best relationship between high-order tuple context and small amount of tuples to NULL is achieved by the proposed POSentropy segmentation criterion.

Differences are much smaller in the Spa→Eng direction, although the same tendency in number of tuples with target NULL is to be found, and conclusions are analogous.

#### 3.4.2.5 Absolute impact

Finally, aiming at finding out which is the absolute impact of taking these segmentation decisions, we define the worst case as taking decisions at random and evaluate translation results, as shown

in Table 3.29, where the median of 5 experiments has been selected.

| | | BLEU | mWER | NIST |
|---|---|---|---|---|
| Eng→Spa | random | 0.4202 | 43.80 | 9.17 |
| Spa→Eng | random | 0.4707 | 38.60 | 9.92 |
| Ara→Eng | random | 0.2758 | 50.74 | 5.78 |

**Table 3.29:** *Worst-case (random segmentation) translation results.*

Surprisingly, alwaysNEXT and IBM1weight strategies perform similarly to the random case, and even worse in the Spa→Eng case. Despite the low statistical significance of just 5 random experiments, the qualitative conclusion of this is that none of these strategies is significantly better than random in this task.

However, in the Ara→Eng case, probably due to corpus size, random strategies generate very sparse data, providing a very bad translation result.

This research was published in the following two papers:

- [Gis06e] A. de Gispert and J.B. Mariño, "Segmentación lingüística de tuplas para el modelado de la traducción estocástica mediante n-gramas," in *Sociedad Española del Procesamiento del Lenguaje Natural, SEPLN'06*, pps. 241–248, September 2006.
- [Gis06d] A. de Gispert and J.B. Mariño, "Linguistic tuple segmentation in ngram-based statistical machine translation," in *Proceedings of the 9th International Conference on Spoken Language Processing, ICSLP'06*, pps. 1149–1152, September 2006.

## 3.5 Chapter Summary and Conclusions

With a marked historical perspective, this chapter traced the development of an initial statistical machine translation model for small-data speech-to-text tasks and its evolution towards a more scalable text-to-text translation system.

By adapting $X$-gram language modelling tools (previously used in speech recognition), we reported the development of an $X$-grams Finite-State Transducer implementation, paying special attention to training details. Along with a record of the first text and speech translation experiments using SMT at UPC, a study of bilingual unit definition (ie. tuple) and a discussion on the monotonicity constraint were also done.

The participation of this preliminary SMT system in an international campaign was reviewed, serving as testbed for the identification of limitations, therefore pointing future work towards improvement (which is extended in the following chapter).

Apart from that, the chapter also presented how scalability obliged to abandon the use of FST implementation, thus reformulating $X$-grams into an standard $N$-gram model over the bilingual language defined by tuples. For this new framework, evaluation of history length, pruning strategies and smoothing techniques in Spanish↔English translation is conducted. As a result from experimental work, we can conclude that it is indeed feasible to build such a model efficiently, as standard language model pruning techniques can be applied here successfully. Additionally, a real-life application of this model was presented for a very monotone task as Spanish↔Catalan translation.

To conclude, the chapter delved into details on tuple definition (embedded words, resolution of source NULLs), and their implications on translation modelling and performance. Various alternative tuple segmentation strategies were compared for various tasks, and the proposed technique based on Part-Of-Speech entropy information performed best in all tasks.

# Chapter 4

# Ngram-based SMT

## 4.1 Introduction

This chapter is devoted to the study of a state-of-the-art statistical machine translation system based on $N$-grams. Built upon the $N$-gram bilingual translation model studied in detail in the previous chapter, this system incorporates a set of additional feature functions into a log-linear combination. Therefore, the core translation model is extended with complementary information, achieving better performance in most of the tasks.

This chapter is organised as follows:

- The mathematical framework underlying the log-linear combination of models is presented in §4.2. Each additional feature model is also described, along with some relevant training and optimisation issues. A comment on the decoding and optimisation tools is also done.
- §4.3 reports on the experiments conducted in order to evaluate the impact of each feature function in translation quality. It also includes an study of examples and manual error analysis for an Eng↔Spa task.
- Finally, §4.4 offers an overview of the results achieved by the $N$gram-based SMT system in various evaluation campaigns during 2005 and 2006, which serve as comparative study by facing the system against alternative SMT systems and in different tasks. A discussion on how reordering strategies have been introduced into the $N$-gram translation approach is also done.

To conclude, in §4.5 a summary of the chapter can be found, highlighting the main conclusions extracted from it.

## 4.2 Feature-based log-linear combination

As already mentioned in §2.4, recent translation systems have replaced the original noisy channel approach by a more general approach, which is founded on the principles of maximum entropy applied to Natural Language Processing tasks ([Ber96]). Under this framework, given a source sentence $s$, the translation task is defined as finding that target sentence $t$ which maximises a log-linear combination of multiple feature functions $h_i(s,t)$, as described by the following equation (equivalent to equation 2.10):

$$\hat{t} = \underset{t}{argmax} \sum_{m} \lambda_m h_m(s,t) \tag{4.1}$$

where $\lambda_m$ represents the coefficient of the $m^{th}$ feature function $h_m(s,t)$, which corresponds to a log-scaled version of $m^{th}$-model probabilities. Optimal values for the coefficients $\lambda_m$s are estimated via an optimisation procedure on a certain development data set.

Next the feature functions used in the $N$-gram SMT system are presented. Then, we discuss on the global training process of the system and $N$gram-based decoding. This section concludes with a discussion on the optimisation procedure.

### 4.2.1 Feature functions

In addition to the tuple $N$-gram translation model, the $N$-gram based SMT system implements four feature functions which provide complementary views of the translation process, namely a target-language model, a word-bonus model and two lexicon models. These features are described next.

#### 4.2.1.1 Target-language model

This feature provides information about the target language structure and fluency, by favouring those partial-translation hypotheses which are more likely to constitute correctly structured target sentences over those which are not. The model implements a standard word $n$-gram model of the target language, which is computed according to the following expression:

$$h_{LM}(s,t) = h_{TL}(t) = log \prod_{k=1}^{K} p(w_k | w_{k-1}, w_{k-2}, \ldots, w_{k-n+1}) \tag{4.2}$$

where $w_k$ refers to $k^{th}$ word in the considered partial-translation hypothesis. Notice that this model only depends on the target side of the data, and can actually be trained by including additional information from other available monolingual corpora.

From a theoretical point of view, the translation bilingual model already constitutes a source *and* target language model. Therefore, one could be led into believing this target language model to be redundant and unnecessary. On the other hand, the bilingual model is more liable to suffer from sparseness than any monolingual model, which can turn this model helpful whenever tuple $n$-grams are not well estimated. Indeed, as will be seen in §4.3.2, this model does provide useful information to improve translation quality.

#### 4.2.1.2 Word-bonus model

The use of any language model probabilities for decoding a new test sentence is associated with a length comparison problem. In other words, when two hypotheses compete in the search for the most probable path, the one using less number of elements (be it words or tuples) will be favoured against the one using more. This is just because a smaller amount of probabilities will be multiplied to obtain the accumulated partial score. This problem results from the fact that the number of target words used for translating a test set is not fixed.

In order to compensate this preference for short translations over large ones, we introduce a word bonus which depends on the partial-translation hypothesis length. This simple model is implemented through a bonus factor which directly depends on the total number of words contained in the partial-translation hypothesis, and it is computed as follows:

$$h_{WB}(s,t) = h_{WB}(t) = K \tag{4.3}$$

where $K$ is the number of words contained in the partial-translation hypothesis.

An alternative approach could be to compute a certain mean of the language model (arithmetic or geometric), so that scores for partial hypotheses with a different number of target words can be comparable.

#### 4.2.1.3 Source-to-target lexicon model

We can defined as a lexicon model any model assigning each translation unit a fixed score. This score is supposed to account for the translational equivalence between the source and target words within the translation unit. For example, we expect such a model to provide tuple '`por favor # please`' with a much better score than tuple '`por favor # will you`.'

Several statistical approaches can be followed to automatically obtain these scores from parallel corpora, mainly focused on cooccurrence or associative measures. However, the most widely-used approach is based on IBM model 1 probabilities (already discussed in §2.1.1).

In $N$gram-based SMT, for each tuple $(s,t)_n$ containing $I$ source words and $J$ target words,

we implement a source-to-target lexicon feature computed for each tuple $(t,s)_n$ according to equation 4.4:

$$h_{LEX}(s,t) = log\frac{1}{(I+1)^J}\prod_{j=1}^{J}\sum_{i=0}^{I}p_{IBM1}(t_j^n|s_i^n) \tag{4.4}$$

where $s_i^n$ and $t_j^n$ are the $i^{th}$ and $j^{th}$ words in the source and target sides of tuple and $p_{IBM1}(.)$ refers to IBM-1 lexical parameters estimated from alignments computed in the source-to-target direction.

Note that this lexicon feature favours tuples with a smaller number of *target* words against longer tuples (for a fixed number of source words), as the formulation from equation 4.4 lacks a square $\sqrt[J]{}$ to compensate for the product over the number of target words (so it becomes a geometric mean).

Alternative implementations of this model can be defined if, for example, IBM model 1 probabilities are substituted by IBM model 2 or HMM model parameters (with a corresponding tuple internal alignment function), or even by a link relative frequency on the word-aligned training data.

#### 4.2.1.4 Target-to-source lexicon model

Given the asymmetry of IBM models discussed in §2.1.1, a complementary model can be computed by making use of the target-to-source lexicon probabilities, estimated from alignments computed in the target-to-source direction. This results in an equivalent yet additional lexicon model.

Therefore, for the same tuple $(s,t)_n$ containing $I$ source words and $J$ target words, this model computes the following score:

$$h_{LEXinv}(s,t) = log\frac{1}{(J+1)^I}\prod_{i=1}^{I}\sum_{j=0}^{J}p_{IBM1}(s_i^n|t_j^n) \tag{4.5}$$

where $p_{IBM1}(.)$ now refers to the inverse IBM-1 lexical parameters, ie. estimated from alignments computed in the source-to-target direction.

Note that this lexicon feature favours tuples with a smaller number of *source* words against longer tuples (for a fixed number of target words), as the formulation from equation 4.4 lacks a square $\sqrt[I]{}$ to compensate for the product over the number of source words (so it becomes a geometric mean).

### 4.2.2 Global training scheme

Training an *N*gram-based SMT system as described above can be graphically represented as in Figure 4.1. When it comes to the bilingual *N*-gram translation model, its training scheme is analogous to that of the *X*-gram translation model from §3.2.2.1, save for the final estimation stage. In fact, the same issues regarding which original alignment to use, which tuple segmentation strategy to follow or which pruning and smoothing techniques to apply are to be found here. Therefore, they will be skipped.

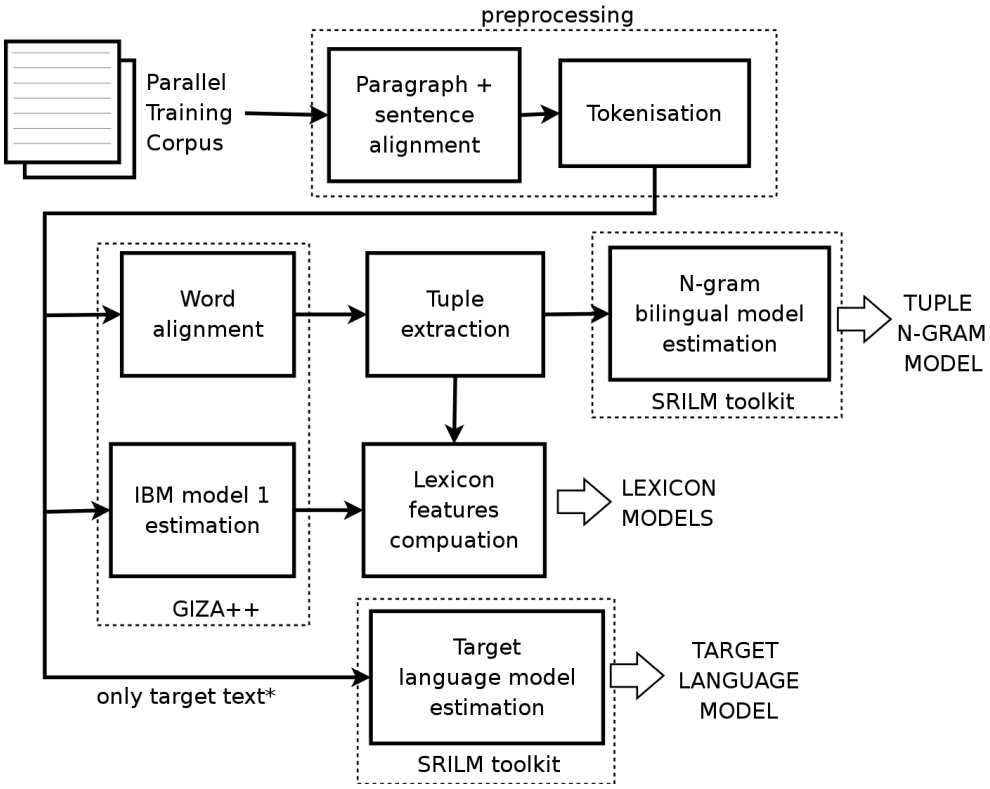



**Figure 4.1:** Training of the models included in an Ngram-based SMT system. Flow diagram.

The additional training blocks include estimating a monolingual language model with the target language material only (which could be extended with monolingual data, if available) and computing the two aforementioned lexicon models from IBM model 1 probabilities.

### 4.2.3 Decoding

As decoding when an *N*gram-based translation model is present is slightly different from phrase-based decoding, in all our experiments we use MARIE, an n-gram based searching engine developed at UPC[1], which was presented in [Cre05b].

[1] By Josep Maria Crego.

MARIE implements a beam-search strategy based on dynamic programming. The decoding is performed monotonically and is guided by the source. During decoding, partial-translation hypothesis are arranged into different stacks according to the total amount of source words they cover. In this way, a given hypothesis only competes with those hypotheses that provide the same source-word coverage. At every translation step, stacks are pruned in order to keep decoding tractable. MARIE allows for two different pruning methods:

- Threshold pruning. All partial-translation hypotheses scoring below a predetermined threshold value are eliminated.
- Histogram pruning. The maximum number of partial-translation hypotheses to be considered is limited the $b$-best ranked ones.

Both these parameters are crucial since they balance a trade-off between translation quality and translation efficiency (decoding times). This research, however, is out of the scope of this Ph.D. thesis. Therefore, no threshold pruning is used in any result, whereas histogram pruning is always set to 50.

MARIE also considers the additional feature functions during decoding. All these models are taken into account simultaneously, along with the $n$-gram translation model. Even though this decoding tool also allows for the generation of $N$-best lists or word graphs as output (necessary for passing through to re-ranking modules), we did not use it in this research work. Therefore, all models are combined in search and a single best hypothesis is output.

Tackling word ordering translation problems, MARIE also includes non-monotone search strategies, as presented in [Cre05c].

### 4.2.4 Optimisation procedure

Minimum-error training states that we can directly train our models according the an error-minimisation function on a certain development data, as discussed in §2.4.1. In our $N$gram-based SMT system, this process is related to the $\lambda_m$ weights which have to be assigned to each feature function.

In order to find out these weights for the log-linear combination, the development set is translated several times with different model weights, and each time evaluated with an automatic score. Typically, BLEU is used (sometimes also mWER, or even combinations of BLEU and NIST as in [Che05]) and computed against a limited number of references. This process in graphically illustrated in Figure 4.2.

Theoretically, this combination approach is justified by the Maximum Entropy framework, which allows for adding any kind of feature model to the final decision. However, the BLEU-

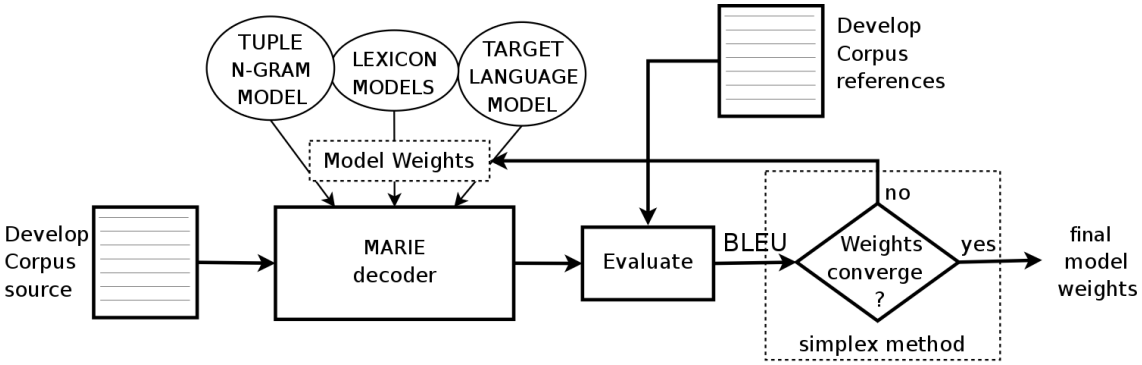



**Figure 4.2:** Optimisation procedure. Flow diagram.

optimisation is a simplification of what should be an entropy maximisation in the development set, which is too complex computationally. The adequacy of this optimisation is founded on the following assumptions:

- There exists a set (or sets) of weights maximising the score in the development set, and it can be found
- The weights maximising the score on the development set will maximise the score on the test set (which should happen unless our development strategy suffers from over-fitting to the development)
- Maximising the score produces better translations (which is related to the correlation between automatic and manual evaluation metrics)

In our implementation, we use an optimisation tool, which is based on the downhill simplex method presented in [Nel65]. BLEU score was used as optimisation function in all reported experiments unless explicitly stated.

## 4.3 Experiments, examples and error analysis

In this section Spanish↔English large-vocabulary translation experiments with the *N*gram-based SMT system are reported. First, an initial baseline is settled by making use of the bilingual $n$-gram model from chapter 3.

Then, three extensions of this baseline are evaluated; firstly, the target language model and word bonus features are added to the log-linear combination; secondly, we complement the baseline with the two lexicon models; and lastly, all five models are log-linearly combined to obtain a *full* system.

All experiments are carried out in the European Parliament corpus (version 1), whose data sets are described in detail in Table 3.14 from §3.3.1.

### 4.3.1 Translation model (alone)

Results of the baseline system (only taking into account the $n$-gram bilingual model) are shown in Table 4.1, where three different word alignment sets are used to extract tuples from, namely the source-to-target alignment, the symmetrised union and the symmetrised refined technique presented in [Och03c].

| | align. | BLEU | mWER | NIST | PER | unigrams | bigrams | trigrams |
|---|---|---|---|---|---|---|---|---|
| Eng→Spa | src-to-trg | 0.4152 | 44.61 | 8.762 | 36.54 | 1.81 | 6.26 | 2.27 |
| | union | **0.4276** | **44.46** | **9.175** | **34.71** | 2.02 | 6.09 | 1.75 |
| | refined | 0.4193 | 44.39 | 8.952 | 35.81 | 2.08 | 6.92 | 2.32 |
| Spa→Eng | src-to-trg | 0.4424 | 40.94 | 9.366 | 32.86 | 1.92 | 6.43 | 2.35 |
| | union | **0.4745** | **39.71** | **9.839** | **30.51** | 2.04 | 6.01 | 1.80 |
| | refined | 0.4594 | 40.24 | 9.618 | 31.80 | 2.11 | 6.85 | 2.40 |

**Table 4.1:** *Effect of alignment set on n-gram translation model quality and size.*

As in the previous chapter, the union of source-to-target and target-to-source alignments always produces the best tuple translation model. These results are marked in bold face. Possibly this is explained by the smaller model sizes (in terms of number of bigrams and trigrams) of this option, leading to a less sparse model.

### 4.3.2 Target language model and word bonus

For the same task, Table 4.2 shows the results of including the target language model and word bonus in the log-linear combination (rows entitled '+trgx +WB'). The difference between '+trg3' and '+trg4' lies in the order of the target language model (trigram or fourgram, respectively). The right-most column shows the number of words of the output generated translation.

As it can be seen, including these models yields a small improvement in terms of BLEU and NIST scores for both translation directions. However, mWER and PER do not show improvements (except for the minor change in mWER from Spanish to English). In the Spa→Eng direction, note that the contribution of these models produces more output words, whereas in the opposite direction the effect is contrary.

Increasing the order of the target language model from 3-grams to 4-grams appears to have a negative effect on evaluation scores for both translation directions (except for NIST and PER scores in Eng→Spa). However, differences are not very significant.

### 4.3.3 Lexicon (IBM) models

Table 4.2 also shows the results of including the two lexicon models described in §4.2.1 in the log-linear combination (rows entitled '+lexicon').

| | config. | BLEU | mWER | NIST | PER | trg words |
|---|---|---|---|---|---|---|
| Eng→Spa | baseline | 0.4276 | 44.46 | 9.175 | 34.71 | 25,523 |
| | +trg3 +WB | 0.4367 | 44.67 | 9.196 | 35.50 | 25,226 |
| | +trg4 +WB | 0.4358 | 45.21 | 9.258 | 35.11 | 26,114 |
| | +lexicon | 0.4482 | 41.69 | 9.541 | 32.43 | 25,391 |
| | full (trg3) | 0.4688 | **40.34** | 9.715 | **32.02** | 25,062 |
| | full (trg4) | **0.4714** | 40.55 | **9.740** | 32.22 | 25,094 |
| Spa→Eng | baseline | 0.4745 | 39.71 | 9.839 | 30.51 | 21,313 |
| | +trg3 +WB | 0.4856 | 39.51 | 9.865 | 30.63 | 21,490 |
| | +trg4 +WB | 0.4826 | 40.53 | 9.774 | 30.99 | 22,447 |
| | +lexicon | 0.5356 | 35.77 | 10.540 | 26.64 | 22,030 |
| | full (trg3) | 0.5434 | 34.94 | 10.620 | 26.60 | 21,565 |
| | full (trg4) | **0.5483** | **34.66** | **10.670** | **26.35** | 21,689 |

**Table 4.2:** *Contribution of feature models in translation quality (EuParl version 1 task). Average reference length is 20,409 (for English) and 24,705 (for Spanish).*

In this case, the impact on translation quality is much more remarkable, as *all* automatic evaluation scores improve significantly. To understand this important contribution, let us consider the lexicon model formulation from equation 4.4. According to it, this model has two main characteristics which seem to contribute positively to the performance boost:

- Tuple ranking. While the bilingual translation models makes each current translation dependant on previous translated parts via a tuple $N$-gram dependency, the lexicon feature assigns an *a priori* fixed cost to each tuple regardless of the context it is used in. Apparently, this has a noise-filtering effect coupled with a better capacity of the system to distinguish correct translations when the bilingual model falls to low-order $n$-grams.

- Short tuple preference. As discussed in §4.2.1.3, the direct and inverse lexicon models preference for shorter tuples (in number of target words and in number of source words, respectively) seems to promote the use of high-order bilingual $n$-grams (by catenating short tuples) rather than using long unfrequent tuples which tend to cause a back-off fall to unigram.

#### 4.3.3.1 Alternative lexicon features

In order to study the validity of the previous explanation on the remarkable contribution of IBM-based lexicon features, we conducted the following experiment, aiming at separating these two effects (tuple ranking and short tuple preference).

Let us define an alternative source-to-target lexicon model, as in the following equation:

$$h_{nLEX}(s,t) = log \sqrt[J]{\frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p_{IBM1}(t_j^n|s_i^n)} \tag{4.6}$$

where $s_i^n$ and $t_j^n$ are the $i^{th}$ and $j^{th}$ words in the source and target sides of tuple and $p_{IBM1}(.)$ refers to IBM-1 lexical parameters estimated from alignments computed in the source-to-target direction. This definition is analogous to equation 4.4 except for the included square function, which converts the feature into a geometric mean.

Therefore, this new 'normalised' lexicon feature does not include an explicit dependency on tuple length. Let us define the target-to-source 'normalised' lexicon feature as:

$$h_{nLEXinv}(s,t) = log \sqrt[I]{\frac{1}{(J+1)^I} \prod_{i=1}^{I} \sum_{j=0}^{J} p_{IBM1}(s_i^n|t_j^n)} \tag{4.7}$$

where $p_{IBM1}(.)$ now refers to the inverse IBM-1 lexical parameters, ie. estimated from alignments computed in the source-to-target direction.

Let us now define a very simple tuple bonus feature model depending on the partial-translation hypothesis length in *number of tuples*. This can be mathematically expressed as:

$$h_{TB}(s,t) = h_{TB}(t) = T \tag{4.8}$$

where $T$ is the number of tuples contained in the partial-translation hypothesis. This aim of this feature is to supply the log-linear combination with a mechanism to prioritise the use of more tuples (ie. catenating shorter tuples), somehow emulating the length preference effect from original lexicon models.

Having defined these new features, we can evaluate their separate and aggregated impact on translation quality, as well as compare it to the contribution from original lexicon models.

Due to data availability reasons, this experiment was carried out on EuParl version 3 task[2]. The European Parliament Proceedings Corpus version 3 differs from version 2 in that it adds data from December 2004 to May 2005 as training material. The main statistics of this corpus are shown in Table 4.3, where similarity with version 2 in Table 3.24 is noticeable.

| EuParl ver3 | | sent. | words | vcb | OOVs | avg.len. | refs. |
|---|---|---|---|---|---|---|---|
| train | English | 1.28 M | 34.92 M | 106.5 k | - | 27.2 | 1 |
| | Spanish | | 36.58 M | 153.1 k | - | 28.5 | |
| develop | English | 504 | 15.4 k | 2.27 k | 10 | 30.5 | 3 |
| | Spanish | | 15.4 k | 2.73 k | 19 | 30.6 | 3 |
| test | English | 1094 | 26.92 k | 4.0 k | 56 | 24.6 | 2 |
| | Spanish | 840 | 22.77 k | 4.1 k | 28 | 27.1 | 2 |

**Table 4.3:** *European Parliament English-Spanish corpus version 3 statistics.*

Table 4.4 shows the results of the baseline system (only using bilingual translation model) and its extensions with the original lexicon model ('+lexicon'), the normalised lexicon models ('+lexNORM'), the tuple bonus model ('+TB'), and the combination of normalised lexicon and tuple bonus ('+lexNORM +TB'). Additionally, the table shows the number of output words, the number of tuples used in the translation, and the percentage of these which were used as 3grams, 2grams and 1grams by the bilingual model.

| | config. | BLEU | mWER | words | tuples | 3gram 2gram 1gram |
|---|---|---|---|---|---|---|
| Eng→Spa | baseline | 0.4270 | 43.60 | 25,533 | 23,217 | 38.0% 43.7% 18.3% |
| | +lexicon | **0.4568** | 41.24 | 26,076 | 23,892 | 38.9% 43.5% 17.7% |
| | +lexNORM | 0.4390 | 42.52 | 25,692 | 22,890 | 37.0% 43.7% 19.3% |
| | +TB | 0.4308 | 43.33 | 25,489 | 23,864 | 39.3% 43.8% 16.9% |
| | +lexNORM +TB | **0.4526** | 41.55 | 25,971 | 23,986 | 39.3% 43.5% 17.3% |
| Spa→Eng | baseline | 0.4735 | 38.51 | 21,075 | 20,026 | 45.2% 40.6% 14.2% |
| | +lexicon | **0.5398** | 34.65 | 22,137 | 20,962 | 46.0% 40.0% 14.0% |
| | +lexNORM | 0.5056 | 36.75 | 21,554 | 19,687 | 43.7% 40.7% 15.6% |
| | +TB | 0.4881 | 37.77 | 21,387 | 20,842 | 47.3% 40.0% 12.7% |
| | +lexNORM +TB | **0.5335** | 35.30 | 22,253 | 20,990 | 46.6% 39.8% 13.6% |

**Table 4.4:** *Contrasted contribution of lexicon, normalised lexicon and tuple bonus features (European Parliament version 3 task).*

As it can be seen, the contribution of the normalised lexicon models (rows '+lexNORM') is significantly worse than the original lexicon model (up to 0.034 and 0.018 BLEU absolute points

[2]However, qualitative results correlation among all EuParl tasks has been observed systematically, possibly due to the slight differences between them

from Spa→Eng and Eng→Spa, when compared to results from row 'lexicon').

In addition to that, and despite forcing the generation of a greater number of output words, the additional gain provided by the simple tuple bonus model is also meagre, especially in the Eng→Spa direction.

As expected, the combination of normalised lexicon models and tuple bonus *does* achieve a significant improvement over the baseline, as well as over their separate contributions. This confirms our suggestion that the two effects of the original lexicon model formulation (from equations 4.4 and 4.5) strengthen each other in combination for a significant translation quality improvement.

However, obtained results with this combination are still a bit below the lexicon models result, as observed comparing the figures in bold face. This is possibly explained by the fact that lexicon models depend on tuple source length and tuple target independently[3], while the simple tuple bonus is just an approximation of this effect. A further comparative experiment could be conducted by substituting this tuple bonus by two word bonuses depending on source and target length independently.

### 4.3.4 Full system

Back to Table 4.2, the results of the full system (ie. incorporating all four features in the combination with the bilingual model) can be seen and compared to the previous extensions. As shown in rows named 'full (trgx)', the obtained results overcome all previous system extensions. Clearly, there exists a certain degree of complementary information between lexicon models and the target language model (plus word bonus).

Interestingly, in this case increasing the order of the target language model to 4-grams (shown in row 'full (trg4)') is positive in both translation directions, although again differences with the 3-grams case (shown in row 'full (trg3)') are minor.

All in all, in Spa→Eng features provide an increase of 0.074 BLEU absolute points and a reduction of 5.05 mWER absolute points, a very significant improvement. In the Eng→Spa direction, BLEU increase amounts to 0.044 absolute points while mWER reduction is of 3.9 absolute points. Again, the higher difficulty of this latter translation direction, mainly due to the richer Spanish morphology and bigger vocabulary size, is made evident by the smaller contribution of feature functions.

In addition to that, and in spite of the presence of a word bonus model, the *N*gram-based SMT system seems to be short of output words. In fact, in all translation experiments the

[3]The direct lexicon preferring tuples with less number of target words, and the inverse lexicon preferring those with less number of source words.

produced translation has less number of words than the input sentence.

Note from Table 3.14 that Spanish input test set contains 22,730 words, while output translations range between 21,313 (only bilingual model) and 22,447 (including target language model and word bonus), as shown in the right-most column of Table 4.2. This input/output decreasing ratio seems to be adequate considering that Spanish training data contains 34.79 M words and English contains 33.38 M words.

However, in the Eng→Spa direction, English input test contains 26,883 words, while output translations range between 25,523 (only bilingual model) and 26,114 (including target language model and word bonus). Given that we expect Spanish to have more words than English, this input/output decreasing ratio is clearly inadequate for this translation direction.

To conclude this section, the impact of tuple pruning and tuple segmentation on the full *N*gram-based SMT system performance is assessed.

#### 4.3.4.1 Effect of tuple pruning

In this section we review the experiment of tuple pruning from §3.3.1.2. Since this pruning has a strong influence in the amount and variety of translation hypothesis the decoder will be able to generate, we re-evaluated the impact of the *tnb* parameter (maximum number of target translations for each source tuple part) on the English↔Spanish EuParl version 1 task, but now including all feature functions, ie. using the full system.

| | | BLEU | mWER | NIST | PER | unigrams | bigrams | trigrams |
|---|---|---|---|---|---|---|---|---|
| Eng→Spa | tnb=30 | 0.4688 | 40.34 | 9.715 | 32.02 | 2.02 | 6.09 | 1.75 |
| | tnb=20 | 0.4671 | 41.29 | 9.731 | 32.27 | 1.96 | 5.84 | 1.73 |
| | tnb=10 | 0.4595 | 41.81 | 9.658 | 32.73 | 1.84 | 5.34 | 1.68 |
| Spa→Eng | tnb=30 | 0.5440 | 34.89 | 10.661 | 26.01 | 2.11 | 6.23 | 1.81 |
| | tnb=20 | 0.5434 | 34.94 | 10.620 | 26.60 | 2.04 | 6.01 | 1.80 |
| | tnb=10 | 0.5399 | 35.05 | 10.632 | 26.28 | 1.92 | 5.57 | 1.76 |

**Table 4.5:** *Effect of tuple n-best pruning on full system translation quality and on translation model size.*

Results are shown in Table 4.5. In the case of Spanish→English, values of $tnb = 20$ and $tnb = 10$, while providing a tuple vocabulary reduction of 3.27% and 8.91% with respect to $tnb = 30$, respectively, produce a translation BLEU reduction of 0.11% and 0.75%. On the other hand, in the case of English→Spanish, values of $tnb = 20$ and $tnb = 10$ provide a tuple vocabulary reduction of 3.31% and 8.89% and a translation BLEU reduction of 0.36% and 1.98% with respect to $tnb = 30$, respectively.

According to these results, a similar tuple vocabulary reduction seems to produce a more negative effect on the English→Spanish direction than in the opposite direction, possibly due to the bigger Spanish vocabulary which demands for a wider range of translation possibilities when translation to Spanish. For this reason, $tnb = 20$ for Spanish→English and $tnb = 30$ English→Spanish are chosen as the most appropriate values.

#### 4.3.4.2 Effect of tuple segmentation

In order to further investigate the impact of tuple segmentation strategies discussed in detail in §3.4.2, here we extend the comparative results presented in §3.4.2.2 and re-evaluate the English↔Spanish EuParl version 2 and Arabic→English small BTEC tasks, but including all feature functions, ie. using the full system.

Table 4.6 shows the translation results for the two best segmentations for each task. As it can be seen, the improvement of better segmenting tuples with NULLs in source is practically compensated by the contribution of additional features, especially in the Spa→Eng task.

| | | BLEU | mWER | NIST |
|---|---|---|---|---|
| Eng→Spa | IBM1weight | 0.4714 | 40.22 | 9.83 |
| | POSentropy | **0.4744** | 40.56 | **9.85** |
| Spa→Eng | IBM1weight | 0.5470 | 34.41 | 10.74 |
| | POSentropy | 0.5466 | 34.44 | 10.72 |
| Ara→Eng | alwaysNEXT | 0.3974 | 40.16 | 7.23 |
| | POSentropy | **0.4024** | **40.05** | **7.39** |

**Table 4.6:** *Translation model performance with additional features for each segmentation strategy.*

Apparently, in the large-vocabulary Spanish-English tasks the target language and lexicon models add robustness to the whole system by penalising tuples with 'wrong' segmentations, or at least their catenation to build up the translation output. Yet the proposed segmentation based on POS-entropy distributions achieves slightly better results in the Eng→Spa direction.

However, small-vocabulary tasks seem more sensitive to segmentation even when combining the core translation model with additional features, as improvements are more significant and higher than using only the translation model.

### 4.3.5 Study of examples

This section is devoted to the study of a few translation examples (including errors and well-translated sentences), in order to offer a real-life view of how each feature function contributes to translation outputs. In other words, we perform an informal subjective evaluation in order

to explain in detail why the model is choosing a certain erroneous (or correct) translation, regardless of the manual references used in automatic scoring.

#### 4.3.5.1 Translation model (alone)

Here some examples when using *only* the translation $N$-gram model (also referred to BM, meaning bilingual model) are presented.

Example 1

Table 4.7 shows the translation provided by the system for a given sentence, where the input sentence is shown above, the output sentence below, and vertical bars mark the segmentation in tuples chosen by the system. Besides, the figure between each tuple indicates whether the following tuple was used as unigram, bigram or trigram in decoding. For example, first tuple was used as 'bigram', because the decoder found a probability from 'sentence beginning' to '`Approval # La aprobación`' which is *directly* estimated in the $N$-gram model (without need to back-off).

| 2 | Approval<br>La aprobación | 3 | of<br>de | 2 | Minutes<br>actas | 2 | of<br>de | 1 | previous sitting<br>sesión | 1 | :<br>: | 2 | see<br>NULL | 1 | Minutes<br>Acta |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

**Table 4.7:** *EuParl version 2 Eng→Spa, development sentence num. 2.*

As we can see from this example, the model works 'fine' when the given context was seen in training. '`La aprobación de actas de`' may not be the most optimal translation of '`Approval of Minutes of`', but it is certainly acceptable. Probably, we'd rather begin with '`Aprobación`', but the corpus (and the model) is clearly pointing towards '`La aprobación`', as seen in Table 4.8 (left-hand side).

| prev: $< s >$ (sent. begin) | | | prev: : # : | | |
|---|---|---|---|---|---|
| Approval # La aprobación | 9 | 5.191538 | see # NULL | 1 | 4.732253 |
| Approval # La concesión | 1 | 6.923906 | see # véanse las | 1 | pruned |
| Approval # Con su decisión | 1 | 6.923906 | see # véase la | 1 | pruned |
| Approval # Aprobación | 1 | 6.923906 | | | |
| Approval # La aceptación | 1 | 6.923906 | | | |
| Approval # Con la aprobación | 1 | 6.923906 | | | |
| Approval # Deben aprobarse | 1 | 6.923906 | | | |

**Table 4.8:** *Possible translations of the word 'Approval' after sentence beginning (EuParl version 2 Eng→Spa), and of word 'see' after tuple ': # :'. For each case, the number of times that the given bigram was observed in training and its BM cost are shown.*

On the other hand, we might prefer '`del Acta`' or '`de las actas`', but we found only one sentence in the whole corpus containing the English bigram '`of Minutes`', and the translation in that case is '`de actas`'.

More problematic is the situation with 'previous sitting', which is worse translated into 'sesión' (since we loose part of the information). This tuple is used as unigram, and in this case, we have a complete tie in BM costs between all hypotheses with this source tuple. As seen in Table 4.9, the $N$-gram model is not reflecting the larger number of occurrences of the more correct target tuple 'sesión anterior' in training material.

| | | |
|---|---|---|
| previous sitting # sesión anterior | 10 | 6.778642 |
| previous sitting # sesión | 2 | 6.778642 |
| previous sitting # periodo de sesiones anterior , | 2 | 6.778642 |
| previous sitting # sesión de ayer | 1 | 6.778642 |
| previous sitting # Pleno anterior | 1 | 6.778642 |
| previous sitting # periodo de sesiones anterior | 1 | 6.778642 |

**Table 4.9:** *Possible incontextual translations of the source tuple 'previous sitting' (EuParl version 2). For each case, the number of occurrences of the tuple in the training data and its BM cost are shown.*

Finally, an omission is produced for the English word 'see' by using tuple 'see # NULL' as bigram. As it can be seen in Table 4.8 (right-hand side), only 3 tuples translating 'see' occur in the training following tuple ': # :', and all of them just once. However, two hypotheses are pruned out (due to the *tnb* unigram-based pruning discussed in §3.3.1.2 and §4.3.4.1) and only the translation to 'NULL' is considered at decoding, just because this is an usual translation of 'see' as unigram.

Given the tie in number of times, it probably would be fairer to consider all hypotheses; or even, given that all three of them occur just once, they might be considered not significant and pruned, forcing back-off. In that case, the incontextual translations of 'see' would be evaluated (see Table 4.10).

| | | |
|---|---|---|
| see # ver | 2008 | 4.487052 |
| see # NULL | 1375 | 4.447047 |
| see # que | 322 | 5.291282 |
| see # comprobar | 307 | 4.989584 |
| see # veo | 195 | 5.305566 |
| see # ve | 124 | 5.214225 |
| see # ... | | |

**Table 4.10:** *Possible incontextual translations of the word 'see' (EuParl version 2 Eng→Spa). For each case, the number of times that the tuple was seen in the training data and its BM cost are shown.*

In that case, the target tuple 'ver' should be favoured in front of 'NULL', according to the number of occurrences. However, note that smoothing ends up preferring the translation to 'NULL', possibly due to the fact that the set of different tuples preceding 'see # NULL' is larger to the one preceding 'see # ver'.

Example 2

Table 4.11 shows another translation provided by the system.

| 2 | That is something I believe<br>Creo que esto | 1 | we owe<br>debemos | 2 | our<br>a nuestros | 2 | taxpayers<br>contribuyentes | 2 | .<br>. |
|---|---|---|---|---|---|---|---|---|---|

**Table 4.11:** *EuParl version 2 Eng→Spa, development sentence num. 139.*

At first glance, translation is again quite 'fine', carrying much of the original sentence meaning. However, Spanish sentence is neither natural nor acceptable, since a '`lo`' pronoun before '`debemos`' is lacking, exactly where *N*gram model back-offs to unigram.

This pronoun is forced by the previous translation; '`Creo que esto`' is a very acceptable translation for '`That is something I believe`', although it demands for a '`lo`' posterior pronoun which seems difficult to produce. This difficulty arises because, even though for source tuple '`we owe`' the most common incontextual target tuple is '`lo debemos`' (see Table 4.12, left-hand side), the bilingual model is not capturing this fact; and in addition to that and most importantly, no tuples with source '`our`' were seen after '`we owe # lo debemos`', whereas they were seen after '`we owe # debemos`' (see Table 4.12, right-hand side).

| we owe # lo debemos | 17 | 5.831678 | prev: we owe # lo debemos | | |
|---|---|---|---|---|---|
| we owe # debemos | 14 | 5.785458 | our # (none) | - | |
| we owe # le debemos | 10 | 5.883411 | | | |
| we owe # les debemos | 4 | 6.318173 | prev: we owe # debemos | | |
| we owe # que debemos | 3 | 6.318173 | our # a nuestros | 1 | 2.069577 |
| we owe # < se lo debemos | 2 | 6.778642 | our # nuestra | 1 | 2.057362 |
| we owe # que debemos dar | 2 | 6.476917 | | | |
| we owe # que le debemos | 1 | 6.778642 | | | |

**Table 4.12:** *Possible incontextual translations of 'we owe' (EuParl version 2 Eng→Spa), and of word 'our' after tuples 'we owe # debemos' and 'we owe # lo debemos'. For each case, the number of times that the given Ngram was seen in training data and its BM cost are shown.*

A more literal translation could avoid the need for the Spanish pronoun '`lo`', for example translating into '`Eso es algo que creo que`'. However, that hypothesis cost is a bit higher as that of the chosen translation, mainly due to the selection of a tuple with 5 source words from the start (less costs to be added). Careful study reveals that the English text '`That is something I believe`' appears 4 times in training corpus beginning a sentence. These four cases (and the tuple segmentation in each case) are shown in Table 4.13.

As it can be observed, in two cases we have the same translation in Spanish ('`Eso es algo que creo que`') but their segmentation in tuples differs.

Finally, *many* instances of bigrams connecting '`That is`' or '`is something`' or '`I believe`' are

| | |
|---|---|
| < *s* ><br>That is something I believe # Creo que esto<br>is # es una<br>demanded # exigencia<br>... | < *s* ><br>That # Eso<br>is # es<br>something # algo<br>I believe # que creo que<br>the present rapporteur can remember # recuerda el actual ponente<br>... |
| < *s* ><br>That # Esto<br>is # NULL<br>something # , en<br>I # mi<br>believe # opinión<br>would be # , sería<br>... | < *s* ><br>That # Eso<br>is # es<br>something # algo<br>I believe # que creo que el<br>Mr # Sr.<br>... |

**Table 4.13:** *All occurrences of 'That is something I believe' starting a sentence in training. Tuple segmentation is shown (EuParl version 2 Eng→Spa).*

to be found in the corpus. One may wonder what is the correct way for the translation model to weight a *rare* unfrequent long tuple in front a frequent sequence of shorter tuples. Other sources of information (such as a target language model) may help providing additional info.

Example 3

Another translation example is to be found in Table 4.14, where we observe a very bad translation while the system back-offs several times to unigram. In this case, there is a clear reordering *and* generalisation (or classification) problem, as the structure '`EUR` $X$ `million #` $X$ `millones de euros`' is very well represented in the corpus for several numerical values of $X$. This situation will not be discussed here.

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | We<br>En<br>2.72 | 1 | spend more than<br>asignan más fondos que<br>7.083 | 1 | EUR<br>euros a<br>6.834 | 1 | 950<br>950<br>6.532 | 2 | million<br>millones de<br>1.226 | 1 | on<br>en<br>... | 2 | subsidies<br>subvenciones | 2 | to<br>a los | 3 | tobacco farmers<br>cultivadores de tabaco | | |

**Table 4.14:** *EuParl version 2 Eng→Spa, development sentence num. 221. For each tuple used, BM cost is shown.*

Before that, a very bad translation of '`We spend more than`' is produced. Consider the 3 alternative partial translations presented in Table 4.15, all of them having a worse cost than that of the partial hypothesis above (24.395).

It is worth mentioning that the more correct path through '`We spend # Gastamos`' and '`more than # más de`' is penalised in front of a *very strange* path through '`We # En`' and '`spend more than # asignan más fondos que`'. In both cases, the number of times the model back-offs to unigram is very high to expect a *nice* translation. In these cases, it is probably necessary to introduce additional linguistic information to improve performance.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | We spend<br>Gastamos<br>6.096 | 1 | more than<br>más de<br>4.338 | 1 | EUR<br>euros<br>7.18 | 1 | 950<br>950<br>6.532 | 2 | million<br>millones de<br>1.226 | TOTAL<br>25.372 |
| 2 | We spend<br>Gastamos<br>6.096 | 1 | more than<br>más<br>4.593 | 1 | EUR<br>euros<br>7.539 | 1 | 950<br>950<br>6.532 | 2 | million<br>millones de<br>1.226 | TOTAL<br>25.987 |
| 2 | We spend<br>Gastamos<br>6.096 | 1 | more than<br>más<br>4.593 | 1 | EUR<br>euros a<br>7.539 | 1 | 950<br>950<br>6.532 | 2 | million<br>millones de<br>1.226 | TOTAL<br>25.987 |

**Table 4.15:** *Alternative translations (EuParl version 2 Eng→Spa, dev. sentence num. 221).*

However, one last comment can be done on frequency of occurrence. Whereas source tuple '`spend more than # asignan más fondos que`' occurs only once in the whole corpus (as unigram), tuples '`more than # más de`' and '`more than # más`' occur 880 and 393 times, respectively (as unigrams). On the contrary, '`We spend # Gastamos`' occurs only twice in the corpus (as bigram after $< s >$), while '`We # En`' occurs more than 2,000 times. As these two tuples compete against each other as bigrams (regardless of the fact of their source part being different), the latter gets a better score[4].

Example 4

Another problematic situation can be found in Table 4.16. Basically, the problems arise when a back-off to unigram is forced by the inexistence of samples covering '`five and a half times as much`' in the corpus. Luckily, the expression '`and a half times`' seems to have a *nice* unigram probability leading to an adequate translation here, but the following expression '`as much as`' is badly translated, as well as the last part ('`the entire environment budget`').

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | five<br>cinco | 1 | and a half times<br>veces y media | 1 | as much as<br>como | 2 | the<br>a | 2 | entire<br>todo el | 1 | environment<br>medio ambiente | 1 | budget<br>del presupuesto | 2 | .<br>. |

**Table 4.16:** *EuParl version 2 Eng→Spa, development sentence num. 221 (last part).*

This example is illustrative of a very challenging input test sentence. One of the reasons is the need for numeric classification in order to identify a history in the expression '$N$ `times as much as`', as this is a represented history in the training, but unfortunately not with $N$='`five and a half`'.

Additionally, '`environmental budget`' poses another challenge to translate. Only four examples of this English bigram are found in the corpus, but unfortunately always inside longer tuples (containing complex non-literal structures). Therefore, a generalisation / reordering strat-

[4]Apart from that, no linguistically-guided preference is given to segmenting tuples that do not breaking a verb form, as the option 'We spend'.

egy is needed to reuse other structures like '$X$ `budget # presupuesto de` $X$' with other values for $X$.

Example 5

Finally, another interesting example is shown in Table 4.17. Again translation is unacceptable but this time, in contrast with all other previous examples, there is no back-off to unigram at all.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | We<br>Nos | 3 | are<br>NULL | 2 | actually<br>NULL | 2 | very<br>muy | 2 | thrifty<br>pocas | 1 | .<br>. |

**Table 4.17:** *EuParl version 2 Eng→Spa, development sentence num. 215.*

Two general comments can be made on this sentence. Regarding tuple segmentation, the intuition that segmenting a full verb form such as '`We are`' into two separate tuples may not be appropriate seems to be confirmed here, as we end up with a very poor translation into '`Nos`'. Clearly, expressions as '`We are actually` $X$' with $X$='`glad`', '`open`', '`delighted`', etc. translating into, for example, '`Nos complace`', have been badly segmented during training.

On the other hand, the word '`actually`' adds a degree of complexity to the sentence, since no example of '`We are actually very`' is to be found in the training, whereas hundreds of examples of '`We are very`' are represented. The bilingual translation model lacks the capacity to realise the presence of this optional verb-modifying adverb, and sequence is broken.

Finally, the unfrequent word '`thrifty`' is also very hard to translate. Not only it is an unfrequent word in the data, but also it received several translations in the corpus (see Table 4.18 for unigram tuples[5]), with different contexts.

| | | |
|---|---|---|
| thrifty # restrictivo | 2 | 6.476917 |
| thrifty # económico | 2 | 6.476917 |
| thrifty # pocas | 1 | 6.778642 |
| thrifty # económicos | 1 | 6.778642 |
| thrifty # cuidadosa | 1 | 6.778642 |
| thrifty # austeridad | 1 | 6.778642 |
| thrifty # ahorrativo | 1 | 6.778642 |
| thrifty # ahorrar | 1 | 6.778642 |
| thrifty # ahorradores | 1 | 6.778642 |
| thrifty # ahorrador | 1 | 6.778642 |

**Table 4.18:** *Possible incontextual translations of the source tuple 'thrifty' (EuParl version 2). For each case, the number of occurrences of the tuple in training data and its BM cost are shown*

Even more interesting is to analyse the 24 tuples where this word is found. Table 4.19

[5]Note that, for once, the model correctly assigns double probability to tuples occurring double the times.

| | |
|---|---|
| has # ha<br>been # sido<br>thrifty # ahorrativo<br>and # y<br>you # ustedes lo | favours # favor<br>a # de un<br>thrifty budget # presupuesto de ahorro<br>. # . |
| favour # también estamos a favor<br>a # de la<br>thrifty # austeridad<br>budget # presupuestaria | I am # soy<br>incredibly # muy<br>thrifty # cuidadosa<br>with # con el<br>taxpayers ' money # dinero de los contribuyentes |
| we must # debemos<br>be # ser<br>thrifty # ahorradores<br>, # NULL | It is # Es<br>thrifty # restrictivo<br>and # y<br>rigorous # riguroso |
| it # NULL<br>is # es<br>thrifty # económico<br>, # , en<br>contrary # contra | we are # somos<br>extremely # sumamente<br>thrifty # económicos<br>, # NULL<br>and # y |
| in # NULL<br>other words # es decir que<br>, we are very thrifty , as they say # , como se dice , somos muy ahorradores<br>. # . | |

**Table 4.19:** *Some of the 24 tuples containing English word 'thrifty' (EuParl version 2 Eng→Spa).*

shows some of them, together with a bit of context. Basically, the word occurs in two *linguistic contexts*, namely as adjective complementing a noun (in most cases the word '`budget`'), or as adjective which is predicate in a TO BE sentence.

Translation is not clear and fluctuates between '`ahorrativo`', '`ahorrador`', '`restrictivo`', '`cuidadosa`' or '`económico`' (and their gender- and number-declined Spanish variants).

Interestingly, three examples of '`we + TO BE + thrifty`' are found. However, the first one has the same tense but is modified by a modal verb ('`we must be`'); the second one (with same person and tense) has a different adverbial modifier ('`extremely`'); and unfortunately, the third one (with same person, tense and even including the following word '`very`') lacks the modifier '`actually`' and besides, it is *hidden* inside a long tuple (with a reordered expression), which could be considered an embedded $N$-gram.

To conclude, the standard $N$-gram model is practically unable to build an acceptable output in this example. The reasons might be the need to generalise / classify verb forms to detect seen Ngrams which do not exactly match the same words, but do match the same verb and even same tense and person ('`we + TO BE + thrifty # SER + ahorradores`').

An alternative to that is to better segment the training into tuples and aim at building the output with a tuple sequence such as '`we are # somos`', '`actually # NULL`', '`very # muy`', '`thrifty # ahorradores`', but this case probably demands an agreement resolution between '`somos`' and

'ahorradores'. Therefore, additional linguistic knowledge seems to be almost necessary.

#### 4.3.5.2 Comparison to full system combination

Next a few examples using all feature functions (full log-linear combination) are presented.

Example 1

Table 4.20 shows the translation provided by the full system for exactly the same sentence discussed in Example 1 of section 4.3.5.1 (see Table 4.7).

| 2 | Approval<br>La aprobación | 3 | of<br>del | 1 | Minutes<br>Acta | 2 | of<br>de la | 1 | previous sitting<br>sesión anterior | 1 | :<br>: | 2 | see<br>NULL | 1 | Minutes<br>Acta |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

**Table 4.20:** *EuParl version 2 Eng→Spa, dev. sentence num. 2 (full system).*

Additionally, Table 4.21 compares the selected new hypothesis (on the right) to the best hypothesis provided by only the Bilingual Model (on the left). For each tuple, the costs of all models are shown (already multiplied by their corresponding weights).

| TUPLE | BM | IBM | IBMi | TM | WB | TUPLE | BM | IBM | IBMi | TM | WB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Approval # La_aprobación | 5.19 | 1.33 | 1.00 | 2.32 | -1.12 | Approval # La_aprobación | 5.19 | 1.33 | 1.00 | 2.32 | -1.12 |
| of # de | 0.27 | 0.35 | 0.17 | 0.12 | -0.56 | of # del | 0.70 | 0.89 | 0.18 | 0.42 | -0.56 |
| Minutes # actas | 6.06 | 0.95 | 0.11 | 3.78 | -0.56 | Minutes # Acta | 6.42 | 0.07 | 0.04 | 0.79 | -0.56 |
| of # de | 1.23 | 0.35 | 0.17 | 0.39 | -0.56 | of # de_la | 1.94 | 0.48 | 0.46 | 0.53 | -1.12 |
| previous_sitting # sesión | 7.39 | 2.60 | 0.12 | 2.76 | -0.56 | previous_sitting # sesión_anterior | 7.44 | 1.04 | 0.33 | 2.60 | -1.12 |
| : # : | 2.88 | 0.08 | 0.05 | 1.89 | -0.56 | : # : | 2.88 | 0.08 | 0.05 | 1.49 | -0.56 |
| see # NULL | 4.73 | 1.16 | 0.36 | 0 | 0 | see # NULL | 4.73 | 1.16 | 0.36 | 0 | 0 |
| Minutes # Acta | 5.69 | 0.07 | 0.04 | 3.59 | -0.56 | Minutes # Acta | 5.69 | 0.07 | 0.04 | 3.66 | -0.56 |
| </s> | 2.77 | | | 2.09 | | </s> | 2.77 | | | 2.09 | |
| COST/model | 36.22 | 6.90 | 2.04 | 16.95 | -4.48 | COST/model | 37.77 | 5.12 | 2.46 | 13.91 | -5.6 |
| **FINAL COST** | 57.63 | | | | | **FINAL COST** | 53.66 | | | | |

**Table 4.21:** *EuParl version 2 Eng→Spa, dev. sentence num. 2 (full system).*

As it can be seen, the Bilingual Model alone assigns a smaller cost to the left-most hypothesis. Some comments on this can be made:

- The trigram '<s>' + '`Minutes # actas`' + '`of # de`' has cost 0.27 in front of 0.70 for '`of # del`'. This is *more or less* consistent with the data, the first occurring 6 times and the latter just 3 times.
- Bigram cost of '`of # de`' + '`Minutes # actas`' (occurs just once in training) is better than a back-off fall to unigram.
- As bigrams, the costs of tuple sequences '`Minutes # actas`' + '`of # de`' and '`Minutes # Acta`' + '`of # de la`' respond to the model estimations shown in Table 4.22.

- Finally, all tuples translating '`previous sitting`' have equal cost as unigram (remember Table 4.9). Their cost difference from 7.39 to 7.44 is only due to the back-off cost of the previous tuple.

| prev: Minutes # actas | | | prev: Minutes # Acta | | |
|---|---|---|---|---|---|
| of # del | 4 | 2.067005 | of # de | 70 | 1.344116 |
| of # de | 2 | 1.230284 | of # de la | 21 | 1.938414 |
| | | | of # del | 13 | 1.735061 |
| | | | of # NULL | 1 | 2.521171 |

**Table 4.22:** *Possible translations of the word of' after tuples 'Minutes # actas' and 'Minutes # Acta' (EuParl version 2 Eng→Spa). For each case, the number of times that the given bigram was seen in training data and its BM cost are shown.*

However, all other models favour the right-most alternative (except for IBM inverse, which has a very low weight). Specially relevant is the very strong preference of the target language model for the new hypothesis, even though this has two more words. This is reinforced by the word bonus model.

Regarding IBM model, note that for some tuples it favours one hypothesis and for some others it favours the opposite, but preference is stronger when differences in tuple length are big. In this sentence, the preference on the '`previous sitting`' tuple plays a decisive role to decide the final best hypothesis.

Example 2

This example refers to development sentence number 18, which is greatly improved by the log-linear model combination. First of all, we observe a positive gender change as shown in Table 4.23.

| TUPLE | BM | IBM | IBMi | TM | WB | TUPLE | BM | IBM | IBMi | TM | WB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| always # siempre | 4.22 | 0.06 | 0.04 | 1.85 | -0.56 | always # siempre | 4.22 | 0.06 | 0.04 | 1.85 | -0.56 |
| has # tiene | 2.13 | 0.88 | 0.19 | 1.38 | -0.56 | has # tiene | 2.13 | 0.88 | 0.19 | 1.38 | -0.56 |
| a # una | 1.27 | 0.508 | 0.10 | 0.98 | -0.56 | a # un | 1.21 | 0.48 | 0.09 | 0.81 | -0.56 |
| very # NULL | 1.88 | 1.10 | 0.34 | 0 | 0 | very # NULL | 2.20 | 1.10 | 0.34 | 0 | 0 |
| difficult # difícil | 2.49 | 0.14 | 0.02 | 2.42 | -0.56 | difficult # difícil | 2.51 | 0.14 | 0.02 | 1.69 | -0.56 |
| role_to_play # papel | 6.39 | 2.06 | 0.12 | 2.62 | -0.56 | role_to_play # papel | 6.39 | 2.06 | 0.12 | 2.58 | -0.56 |
| COST/model | 18.38 | 4.75 | 0.77 | 9.24 | .2.8 | COST/model | 18.67 | 4.72 | 0.76 | 8.30 | -2.8 |
| **FINAL COST** | 30.34 | | | | | **FINAL COST** | 29.65 | | | | |

**Table 4.23:** *EuParl version 2 Eng→Spa, dev. sentence num. 18 (full system).*

As expected, this change is basically motivated by the target language model, which prefers the right-most hypothesis for article-noun agreement reasons (the remaining models being mostly unchanged). The preference of the bilingual model for the left-most hypothesis (centred on the trigram completed by tuple '`very # NULL`') is worth studying.

Note the occurrences in training of these trigrams, as shown in Table 4.24. As it can be seen, neither case has an alternative translation for the word '`very`' but the empty word, according to training material.

| `prev: has # tiene + a # una` | | | `prev: has # tiene + a # un` | | |
|---|---|---|---|---|---|
| `very # NULL` | `6` | `1.878118` | `very # NULL` | `3` | `2.202106` |

**Table 4.24:** *Possible translations of the word 'very' after two different bigrams (EuParl version 2 Eng→Spa). For each case, the number of times that the given bigram was seen in training and its BM trigram cost are shown.*

Later on in the same sentence, another clear translation improvement can be found, as shown in Table 4.25 (again, new hypothesis is shown on the right-hand side of the table).

| TUPLE | BM | IBM | IBMi | TM | WB | TUPLE | BM | IBM | IBMi | TM | WB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| which # que | 2.59 | 0.44 | 0.30 | 1.25 | -0.56 | which # que | 2.59 | 0.44 | 0.30 | 1.25 | -0.56 |
| have_no # no_tienen | 3.81 | 1.78 | 0.55 | 2.20 | -1.12 | have_no # no_tienen | 3.81 | 1.78 | 0.55 | 2.20 | -1.12 |
| choice # otra_elección | 3.04 | 0.41 | 0.58 | 1.89 | -1.12 | choice # otra_elección | 3.04 | 0.41 | 0.58 | 1.89 | -1.12 |
| but # NULL | 0.50 | 1.65 | 0.51 | 0 | 0 | but # NULL | 0.50 | 1.65 | 0.51 | 0 | 0 |
| to # NULL | 1.86 | 0.86 | 0.27 | 0 | 0 | to # que | 0.65 | 0.81 | 0.26 | 0.29 | -0.56 |
| reach_agreement # un_acuerdo | 4.59 | 1.60 | 0.59 | 3.20 | -1.12 | reach # alcanzar_un | 5.13 | 0.70 | 0.64 | 2.96 | -1.12 |
| , # , | 1.89 | 0.31 | 0.11 | 0.82 | -0.56 | agreement # acuerdo | 0.19 | 0.07 | 0.07 | 0.22 | -0.56 |
| although # aunque | 2.57 | 0.21 | 0.12 | 1.11 | -0.56 | , # , | 1.23 | 0.31 | 0.11 | 0.82 | -0.56 |
| on # en | 2.83 | 0.86 | 0.35 | 0.91 | -0.56 | although # aunque | 2.80 | 0.21 | 0.12 | 1.11 | -0.56 |
| occasions # ocasiones | 3.36 | 0.11 | 0.09 | 1.63 | -0.56 | on # en | 2.83 | 0.86 | 0.35 | 0.91 | -0.56 |
| COST/model | 27.03 | 8.24 | 3.46 | 13.01 | -6.16 | occasions # ocasiones | 3.36 | 0.11 | 0.09 | 1.63 | -0.56 |
| **FINAL COST** | 45.58 | | | | | COST/model | 26.12 | 7.35 | 3.58 | 13.27 | -7.28 |
| | | | | | | **FINAL COST** | 43.04 | | | | |

**Table 4.25:** *EuParl version 2 Eng→Spa, dev. sentence num. 18 (full system).*

In this case, the target language model favours the original alternative (on the left). However, as two new words are introduced, the word bonus shows a strong preference for the right-most hypothesis. Regarding IBM models, whereas IBM inverse prefers the original translation but by a small difference (3.46 against 3.58), IBM direct has a strong preference for the new hypothesis. This is due to the relatively high cost for the 'longer' tuple '`reach agreement # un acuerdo`'.

It is worth mentioning that the Bilingual Model also prefers the new hypothesis. However, when decoding only with this model, the decoder chooses the first hypothesis due to a search error; in other words, the new alternative hypothesis is not explored as the sequence ending in '`to # que`'+'`reach # alcanzar un`' is pruned out.

Finally, a third improvement is to be found in the last part of the same sentence, as shown in Table 4.26, where we see that all features contradict the bilingual model and prefer the new hypothesis (on the right), except for IBM inverse, which is irrelevant.

| TUPLE | BM | IBM | IBMi | TM | WB | TUPLE | BM | IBM | IBMi | TM | WB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| occasions # ocasiones | 4.74 | 0.11 | 0.09 | 2.65 | -0.56 | occasions # ocasiones | 4.74 | 0.11 | 0.09 | 2.65 | -0.56 |
| it # NULL | 2.93 | 0.90 | 0.28 | 0 | 0 | it # NULL | 2.93 | 0.90 | 0.28 | 0 | 0 |
| seems # parece | 2.80 | 0.10 | 0.13 | 1.81 | -0.56 | seems # parece | 2.80 | 0.10 | 0.13 | 1.81 | -0.56 |
| that # NULL | 2.22 | 0.96 | 0.30 | 0 | 0 | that # que_el | 2.23 | 0.52 | 0.54 | 1.11 | -1.12 |
| | | | | | | agreement # acuerdo | 2.95 | 0.07 | 0.07 | 1.32 | -0.56 |
| agreement_is_a # es_un_acuerdo | 5.27 | 2.14 | 0.59 | 3.83 | -1.68 | is # es | 1.95 | 0.47 | 0.06 | 1.23 | -0.56 |
| | | | | | | a # NULL | 1.87 | 0.96 | 0.30 | 0 | 0 |
| very_long_way_off # larguísima | 6.83 | 6.60 | 0.48 | 4.30 | -0.56 | very # muy | 1.18 | 0.23 | 0.06 | 0.89 | -0.56 |
| | | | | | | long_way_off # lejos | 4.93 | 4.22 | 0.36 | 1.98 | -0.56 |
| . # . | 0.73 | 0.41 | 0.13 | 0.70 | -0.56 | . # . | 0.97 | 0.41 | 0.13 | 0.49 | -0.56 |
| </s> | 0.01 | | | 0.10 | | </s> | 0.01 | | | 0.01 | |
| COST/model | 25.54 | 11.21 | 1.99 | 13.39 | -3.92 | COST/model | 26.58 | 7.98 | 2.00 | 11.49 | -5.04 |
| **FINAL COST** | 48.21 | | | | | **FINAL COST** | 43.00 | | | | |

**Table 4.26:** *EuParl version 2 Eng→Spa, dev. sentence num. 18 (full system).*

Example 3

Finally, this final example shows how the contribution of feature functions leads to a worse translation. Particularly, for English sequence '`on the basis of which`', the correct translation '`sobre cuya base`' achieved by the bilingual model alone is turned into the incorrect expression '`sobre la base de que`'. See Table 4.27 for the comparison between the old and the new hypothesis, which is selected by the log-linear model combination.

| TUPLE | BM | IBM | IBMi | TM | WB | TUPLE | BM | IBM | IBMi | TM | WB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| the # las | 2.77 | 0.10 | 0.17 | 1.42 | -0.56 | the # las | 2.77 | 0.10 | 0.17 | 1.42 | -0.56 |
| political_priorities # prioridades_políticas | 3.31 | 1.32 | 0.32 | 2.09 | -1.12 | political_priorities # prioridades_políticas | 3.31 | 1.32 | 0.32 | 2.09 | -1.12 |
| on # sobre | 3.26 | 0.51 | 0.07 | 1.58 | -0.56 | on # sobre | 3.26 | 0.51 | 0.07 | 1.58 | -0.56 |
| the # NULL | 1.70 | 0.98 | 0.30 | 0 | 0 | the # la | 0.74 | 0.57 | 0.14 | 0.37 | -0.56 |
| | | | | | | basis # base | 0.99 | 0.24 | 0.04 | 0.72 | -0.56 |
| basis_of_which # cuya_base | 1.70 | 3.39 | 0.46 | 2.18 | -1.12 | of # de | 0.29 | 0.35 | 0.17 | 0.13 | -0.56 |
| | | | | | | which # que | 3.65 | 0.44 | 0.30 | 0.83 | -0.56 |
| the # los | 2.90 | 0.89 | 0.17 | 1.18 | -0.56 | the # los | 2.39 | 0.89 | 0.17 | 0.74 | -0.56 |
| parliamentary_groups # Grupos_parlamentarios | 3.94 | 1.96 | 0.35 | 2.78 | -1.12 | parliamentary_groups # grupos_parlamentarios | 4.72 | 1.60 | 0.35 | 2.51 | -1.12 |
| and # y | 0.72 | 0.26 | 0.07 | 0.63 | -0.56 | and # y_los | 2.06 | 0.45 | 0.44 | 1.14 | -1.12 |
| Members # diputados | 4.33 | 0.17 | 0.04 | 2.50 | -0.56 | Members # diputados | 2.62 | 0.17 | 0.04 | 1.38 | -0.56 |
| stood # defendido | 6.48 | 0.82 | 0.34 | 3.31 | -0.56 | stood # defendido | 6.48 | 0.82 | 0.34 | 3.62 | -0.56 |
| for # NULL | 1.80 | 0.99 | 0.31 | 0 | 0 | for # NULL | 1.80 | 0.99 | 0.31 | 0 | 0 |
| the # de_las | 2.01 | 0.73 | 0.43 | 2.28 | -1.12 | the # las | 2.08 | 1.00 | 0.17 | 1.28 | -0.56 |
| European_elections # elecciones_europeas | 2.11 | 1.66 | 0.27 | 1.76 | -1.12 | European_elections # elecciones_europeas | 3.37 | 1.66 | 0.27 | 2.03 | -1.12 |
| on # del | 2.57 | 1.14 | 0.35 | 0.84 | -0.56 | on # del | 2.57 | 1.14 | 0.35 | 0.84 | -0.56 |
| COST/model | 39.60 | 15.82 | 3.63 | 22.55 | -9.52 | COST/model | 43.10 | 13.15 | 3.65 | 20.70 | -10.64 |
| **FINAL COST** | 72.09 | | | | | **FINAL COST** | 69.95 | | | | |

**Table 4.27:** *EuParl version 2 Eng→Spa, dev. sentence num. 12 (full system).*

Clearly, the Bilingual Model prefers (by far) the first hypothesis, but the IBM, TM and WB models prefer the second (IBMi being once again irrelevant). Regarding IBM model, the preference is mainly due to the high cost assigned to the long tuple '`basis of which # cuya base`' which cannot compete against the low costs of short tuples like '`basis # base`', '`of # de`' and '`which # que`'.

As for the target language model, the situation is much more complex, as many changes are found all through this sentence segment. However, the points of stronger preference for the new hypothesis are in the regions of '`parliamentary groups`' (2.5 versus 1.38) and '`European elections`' (2.28 versus 1.28). Finally, the word bonus clearly prefers the second hypothesis, as it contains two additional words.

### 4.3.6 Error analysis

Even though automatic evaluation measures provide quality scores to assess performance improvements, there is an evident need for understanding the nature of errors that a certain system does. With this objective, in this last section we present a brief manual error analysis performed to some of the outputs provided by the full system configuration. More specifically, a detailed review of 200 translated sentences and their corresponding source sentences, in each direction, was conducted.

This analysis should be helpful to spot which translation phenomena are difficult for the *N*gram-based SMT system, diagnose possible causes and trace a pragmatic research line towards improvement. For this study, we classify most common translation errors into verb forms, omissions, word order, disagreement and bad lexical choice. These error types are explained next:

- **Bad lexical choice**. This occurs whenever the translation model selected a wrong tuple, and the resultant translation is unacceptable, possibly due to a lack of context. To illustrate this, see the following two examples, where '`either`' should be translated into '`tampoco`' (instead of '`ni`') and the correct translation of '`What`' would be '`Lo que`' instead of '`> Qué`':

| 2 | This<br>Esto | 3 | is not<br>no es | 3 | acceptable<br>aceptable | 3 | either<br>ni | 1 | .<br>. |
|---|---|---|---|---|---|---|---|---|---|

| 2 | What<br>> Qué | 3 | is<br>es | 2 | basically<br>básicamente | 1 | happening<br>sucediendo |
|---|---|---|---|---|---|---|---|

**Table 4.28:** *EuParl version 2 Eng→Spa, development sentences num. 83 and 81.*

- **Omissions**. These errors occur when a relevant part of the source sentence is not present in the translation. Two distinct cases are included here, namely whenever the translation model wrongly uses a tuple with NULL target side *and* whenever it chooses a tuple which translates only part of the source in this context (which could also be counted as bad lexical choice, but is included here). One example follows:

| 3 | Cuba<br>Cuba | 2 | no<br>NULL | 2 | debe<br>must | 2 | cambiar<br>change | 3 | .<br>. |
|---|---|---|---|---|---|---|---|---|---|

**Table 4.29:** *EuParl version 2 Spa→Eng, test sentence num. 35.*

- **Word order**. This type refers to translations which require a change in word order in order to be correct. In Spanish↔English translation, most of these cases refer to Adjective–Noun, or Subject–Verb swaps (especially for passive voice), but not exclusively. Two examples follow:

| 3 | toda<br>whole | 3 | la<br>of | 3 | sociedad<br>society | 1 | neerlandesa<br>Dutch | 2 | ,<br>, |
|---|---|---|---|---|---|---|---|---|---|

| 2 | that<br>que los | 1 | financial surpluses<br>excedentes financieros | 1 | are<br>NULL | 1 | prevented<br>evitarse |
|---|---|---|---|---|---|---|---|

**Table 4.30:** *EuParl version 2 Spa→Eng test sentence num. 14 and Eng→Spa development sentence num. 108.*

- **Verb forms**. In terms of meaning, verbs are very important in a sentence, as a mistaken verb form tends to strongly affect the message and may easily lead to confusion. For this reason, we consider any translation of a verb group with erroneous tense, person, number or lemma as a verb form error. Note that this includes verb forms being translated to NULL (ie. omitted), which are *not* counted as omissions. Examples of this type of errors follow:

| 2 | when<br>cuando | 2 | I<br>NULL | 2 | do not<br>no | 1 | rejoice<br>alegrarnos | 2 | that<br>de que |
|---|---|---|---|---|---|---|---|---|---|

| 2 | > Cree<br>Does | 1 | posible<br>possible | 2 | también<br>to | 2 | llamar<br>draw |
|---|---|---|---|---|---|---|---|

**Table 4.31:** *EuParl version 2 Eng→Spa development sentence num. 365 and Spa→Eng test sentence num. 320.*

- **Disagreement**. Also called concordance, it refers to any morphology-related (gender and number) inconsistencies between elements of the sentence. This is especially relevant when translating to morphologically richer languages, as in the Eng→Spa direction. One example can be found in Table 4.32.

Table 4.33 presents the relative number of occurrences for each of the four types or errors identified in both translation directions.

Notice from Table 4.33 that the most common errors in both translation directions are those related to verb forms. However, it is important to mention that 29.5% of verb-form errors in the

| 2 | It was<br>Es | 1 | another<br>otro | 1 | feat of strength<br>proeza | 1 | on<br>por | 2 | the<br>NULL | 3 | part<br>parte |
|---|---|---|---|---|---|---|---|---|---|---|---|

**Table 4.32:** *EuParl version 2 Eng→Spa, development sentence num. 373.*

| Type of error | English→Spanish | Spanish→English |
|---|---|---|
| Bad lexical choice | 24.0% | 21.7% |
| Omissions | 23.0% | 26.3% |
| Word order | 16.3% | 19.7% |
| Verb forms | 27.7% | 28.2% |
| Disagreement | 8.0 % | 4.1% |

**Table 4.33:** Percentage of occurrence for each type of error in the 200 sentences studied.

English→Spanish direction actually correspond to verb omissions. Similarly, 12.8% of verb-form errors in the Spanish→English direction are verb-omissions.

According to this, if errors due to omitted translations and to omitted verb forms are considered together, it is evident that errors involving omissions constitute the most important group, specially in the case of English→Spanish translations. Undoubtedly, the problem of omission needs a very thorough study, as the questions it raises are still to be answered. In fact, our system always produces a shorter output than the input (in number of length), which is clearly inconvenient when translating from English into Spanish. The use of tuples translating to NULL proves beneficial in many occasions (as shown in §3.4.2), yet harmful when certain bilingual context is missing.

On the other hand, even though Spanish and English are claimed to be a rather monotonic language pair, word order problems also prove relevant to translation quality in both directions, and tend to appear in conjunction with omissions or a bad lexical choice.

In addition to all this, Table 4.33 also shows that disagreement errors affect more than twice English→Spanish translations than Spanish→English ones. This result can be explained by the more inflected nature of Spanish.

This shallow error analysis exercise is related to the work of [Vil06], where the most common errors of a standard phrase-based SMT system are manually classified. In general terms, their findings are similar to those presented here for basically the same task.

## 4.4 Results in Evaluation Campaigns

In this section the results of the presented *N*gram-based SMT system in various international evaluation campaigns are summarised. In addition to that, a reference to reordering strategies for *N*gram-based SMT is made.

### 4.4.1 Monotone tasks

#### TC-STAR 1st evaluation

In fall 2004 the TC-Star EU-funded project organised its first evaluation campaign. Language pairs included English↔Spanish, in which UPC took part, and Chinese↔English. Roughly speaking, parallel training data consisted of the European Parliament corpus version 1 (see Table 3.14).

To study the effect of both recognition errors and spontaneous speech phenomena, particularly for the EuParl task, three types of input to the translation system were studied and compared:

- **ASR**: the output of automatic speech recognisers, without using punctuation marks
- **verbatim**: the verbatim (i.e. correct) transcription of the spoken sentences including the phenomena of spoken language like false starts, ungrammatical sentences etc. (again without punctuation marks)
- **text**: the so-called final text editions, which are the official transcriptions of the European Parliament and which do not include the effects of spoken language any more (here, punctuation marks were included)

Results are summarised in Table 4.34, where ASR Word Error Rate is 10.1% for Spanish input and 9.9% for English. RWTH stands for Rheinisch-Westfälische Technische Hochschule in Aachen (Germany), IBM for IBM Research in Yorktown Heights (NY), ITC-irst for Centro per la Ricerca Scientifica e Tecnologica in Trento (Italy), UKA for Universität Karlsruhe (Germany) and UPV for Universitat Politècnica de València (Spain).

As it can be seen, results obtained by UPC *N*gram-based SMT system (marked in bold face) are very competitive with results from other sites[6]. As these results basically correspond to those discussed in §4.3, no further discussion is made here. Further details on this evaluation campaign specification and results can be found in [Ney05].

[6]Note that a bug for RWTH Text results was reported after the evaluation

| | Spanish→English | | | English→Spanish | | |
|---|---|---|---|---|---|---|
| | Site | BLEU | NIST | Site | BLEU | NIST |
| ASR | RWTH | 41.5 | 9.12 | RWTH | 38.7 | 8.73 |
| | IBM | 39.7 | 8.81 | IBM | 34.3 | 8.13 |
| | **UPC** | 37.7 | 8.56 | **UPC** | 33.8 | 8.00 |
| | ITC-irst | 34.7 | 7.97 | UKA | 33.0 | 7.94 |
| | UKA | 32.3 | 7.85 | UPV | 19.1 | 5.46 |
| | UPV | 16.0 | 4.35 | | | |
| Verbatim | RWTH | 45.9 | 9.75 | RWTH | 42.5 | 9.32 |
| | IBM | 44.1 | 9.47 | **UPC** | 38.1 | 8.72 |
| | **UPC** | 42.1 | 9.26 | IBM | 36.8 | 8.55 |
| | ITC-irst | 38.1 | 8.46 | UKA | 33.4 | 8.29 |
| | UKA | 33.4 | 7.96 | | | |
| Text | **UPC** | 53.3 | 10.55 | **UPC** | 46.2 | 9.65 |
| | IBM | 53.1 | 10.38 | IBM | 45.2 | 9.44 |
| | ITC-irst | 47.5 | 9.60 | RWTH' | 38.9 | 8.72 |
| | RWTH' | 46.1 | 9.68 | UKA | 37.6 | 8.46 |
| | UKA | 40.5 | 8.96 | UPV | 34.1 | 7.51 |
| | UPV | 32.7 | 6.80 | | | |

**Table 4.34:** *Results of the 1st TC-STAR evaluation campaign. BLEU is presented as a percentage.*

### ACL 2005 Workshop

In June 2005, a Workshop on Building and Using Parallel Texts in the Annual Meeting of the Association for Computational Linguistics also organised a shared task in machine translation. This time translation was always directed towards English, source languages being Spanish, French, German and Finnish.

Concerning the training material, again the European Parliament Proceedings Corpus was used. However, training corpus size was smaller, ranging from 688k sentences in French–English to 751k sentences in German–English, while test set contained 2k sentences and only one single reference to perform automatic evaluation.

A summary of the results is presented in Table 4.35. UW stands for University of Washington (USA), CMU for Carnegie Mellon University (USA), GLGW for University of Glasgow (Scotland), RALI for University of Montreal (Canada), NRC for National Research Council (Canada), SAAR for Saarland University (Germany) and UJI for Universitat Jaume I (Spain). A lowercased letter indicates a different system submitted by the same site[7].

Again, results obtained by UPC's *N*gram-based SMT system (marked in bold face) rate among the best-ranked systems. Only one system outperforms UPC in all four tasks. Note the differences in evaluation scores across each language pair. Whereas taking Spanish and French as

[7]Regarding UPC, 'UPCm' and 'UPCj' stand for two phrase-based SMT systems.

| Spanish→English | | French→English | | Finnish→English | | German→English | |
|---|---|---|---|---|---|---|---|
| Site | BLEU | Site | BLEU | Site | BLEU | Site | BLEU |
| UW | 30.95 | UW | 30.27 | UW | 22.01 | UW | 24.77 |
| **UPC** | 30.07 | **UPC** | 30.20 | NRC | 20.95 | **UPC** | 24.26 |
| UPCm | 29.84 | NRC | 29.53 | **UPC** | 20.31 | NRC | 23.21 |
| NRC | 29.08 | RALI | 28.89 | RALI | 18.87 | RALI | 22.91 |
| RALI | 28.49 | CMUb | 27.65 | SAAR | 16.76 | SAAR | 20.48 |
| UPCj | 28.13 | CMUj | 26.71 | UJI | 13.79 | CMUj | 18.93 |
| SAAR | 26.69 | SAAR | 26.29 | CMUj | 12.66 | UJI | 18.89 |
| CMUj | 26.14 | GLGW | 23.01 | | | | |
| UJI | 21.65 | UJI | 21.25 | | | | |

**Table 4.35:** *Results of the ACL 2005 Workshop shared task. BLEU is presented as a percentage.*

source language renders a pretty similar translation quality, this falls off when source is German and specially Finnish.

The difficulty of these tasks lies principally in their requirement for word ordering strategies, which was ignored here with a monotone decoding for all tasks. In addition to that, Finnish is an agglutinative language, therefore having a very rich morphology which needs to be processed for translation to have success. On the other hand, German build up long compound word by catenating words, therefore increasing vocabulary size a lot. Unfortunately, no linguistic tools were available for any of these languages.

All further details on this evaluation campaign specification and results can be found in [Koe05b]. UPC *N*gram-based SMT system participation is reported in the following paper:

- [Ban05a] R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert and J.B. Mariño, "Statistical Machine Translation of Euparl Data by using Bilingual N-grams," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pps. 67–72, June 2005.

### 4.4.2 Non-monotone tasks. Reordering strategies

#### IWSLT 2005

In October 2005, the C-STAR[8] consortium organised the 2nd International Workshop on Spoken Language Translation (IWSLT). It included an evaluation campaign whose details can be found in [Eck05].

This time the task was to translate from Chinese, Arabic, Japanese and Korean into English and from English into Chinese, and 17 organisations participated. As in the previous year

[8] Consortium for Speech Translation Advanced Research, http://www.c-star.org

(see §3.2.5.2), training corpus size was about 20k parallel sentences per language pair. UPC participated in Chinese→English and Arabic→English.

Given the obtained results from 2004 with monotone decoding, a word ordering strategy was developed for the *N*gram-based SMT system [Cre05c]. Basically, this consists of relaxing the tuple extraction constraint forcing tuples to monotonically generate the source and target sentences. Unfolding the tuples is now permitted, meaning that whereas the target-sentence order is still preserved, this restriction does not apply to the source sentence anymore. Graphically, this is illustrated in Figure 4.3.

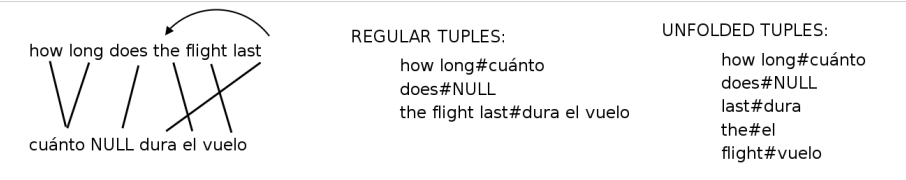



**Figure 4.3:** Differences between regular and unfolded tuple extraction.

Once this unfolded tuples are extracted, the same translation model is estimated, as well as additional features are. However, during decoding time, the decoder must allow for a reordered search. A distortion model, as described in equation 4.9, is included in the log-linear combination of features.

$$h_{DIST} = \sum_{k=1}^{K} d_k \tag{4.9}$$

where $d_k$ is the distance between the first source word of the $K^{th}$ tuple, and the last source word of the $K-1^{th}$ tuple plus 1.

In order not to suffer from a computational complexity explosion, two parameters constrain this reordered search space:

- A distortion limit ($m$): Any source word (phrase or tuple) is only allowed to be reordered if it does not exceed a distortion limit, measured in words.
- A reordering limit ($j$): Any translation path is only allowed to perform $j$ reordering jumps.

These parameters were fixed to $m = 5$ and $j = 3$ for the Chinese→English task, and to $m = 3$, $j = 3$ for the Arabic→English task. These settings suppose a necessary trade-off between quality and efficiency. As reordering is not so critical in the Arabic task and does not produce any big improvement in quality, a smaller distortion distance limit was used.

Comparative results are summarised in Tables 4.36 for Arabic and 4.36 for Chinese, which includes manual evaluation scores.

| Arabic→English | | | |
|---|---|---|---|
| BLEU | | NIST | |
| UPCph | 57.3 | RWTH | 9.78 |
| ITC-irst | 56.2 | ITC-irst | 9.66 |
| RWTH | 54.7 | UPCph | 9.33 |
| IBM | 53.8 | NTT | 9.27 |
| **UPC** | 53.3 | CMU | 8.74 |
| EDINBG | 51.1 | IBM | 8.62 |
| NTT | 44.6 | EDINBG | 7.64 |
| CMU | 40.9 | **UPC** | 6.54 |
| USC-ISI | 37.4 | USC-ISI | 2.85 |
| ITC-irst | 52.8 | RWTH | 9.57 |

**Table 4.36:** *Results of IWSLT 2005 shared task. BLEU is presented as a percentage.*

Among new acronyms, ATR-C3 stands for ATR Spoken Language Communication Research in Kyoto (Japan), EDINBG for University of Edinburgh (Scotland), MIT/AF for MIT Air Force Research Lab (USA), USC-ISI for University of Southern California (USA) and NTT for Nippon Telegraph and Telephone Cyber Space Labs (Japan). Again, UPC results are printed in bold face.

| Chinese→English | | | | | | | |
|---|---|---|---|---|---|---|---|
| Fluency | | Adequacy | | BLEU | | NIST | |
| ITC-irst | 3.15 | MIT/AF | 2.71 | ITC-irst | 52.8 | RWTH | 9.57 |
| RWTH | 3.04 | ITC-irst | 2.65 | RWTH | 51.1 | MIT/AF | 9.31 |
| CMU | 2.88 | RWTH | 2.63 | EDINBG | 46.5 | ITC-irst | 9.06 |
| ATR-C3 | 2.86 | UPCph | 2.52 | UPCph | 45.2 | IBM | 8.44 |
| **UPC** | 2.82 | IBM | 2.51 | MIT/AF | 45.0 | **UPC** | 8.40 |
| EDINBG | 2.81 | **UPC** | 2.44 | **UPC** | 44.4 | ATR-C3 | 8.00 |
| MIT/AF | 2.79 | EDINBG | 2.33 | CMU | 44.4 | UPCph | 7.97 |
| UPCph | 2.78 | ATR-C3 | 2.31 | IBM | 44.0 | NTT | 7.52 |
| IBM | 2.77 | NTT | 2.09 | ATR-C3 | 39.4 | EDINBG | 6.49 |
| USC-ISI | 2.32 | CMU | 1.95 | USC-ISI | 33.2 | CMU | 6.19 |
| NTT | 1.97 | USC-ISI | 1.90 | NTT | 27.8 | USC-ISI | 5.57 |

**Table 4.37:** *Results of IWSLT 2005 shared task. BLEU is presented as a percentage.*

As observed from the tables, results still need an improvement for tasks requiring word order, even though improvement with respect to the previous year is noticeable (in Chinese task). In the case of Arabic, bad correlation between development and test sets led to a wrong decision regarding word alignment, which ended up producing very short output translation (as reflected by the very low NIST score). Details are reported in the following publication:

- [Cre05a] J.M. Crego, A. de Gispert, and J.B. Mariño, "TALP: The UPC Tuple-based SMT System," in *Proceedings of the 2nd International Workshop on Spoken Language Translation, IWSLT'05*, pps. 191–198, October 2005.

### TC-STAR 2nd evaluation

In February 2005 the TC-Star EU-funded project organised its second evaluation campaign. As in the previous year, language pairs included English↔Spanish, in which UPC took part, and Chinese↔English. Parallel training data consisted of the European Parliament corpus version 3 (see Table 4.3), while test data was newly collected. Again three types of input to the translation system were studied and compared, namely ASR, verbatim and text (see §4.4.1).

| | Spanish→English | | | |
|---|---|---|---|---|
| | BLEU | | NIST | |
| ASR | IBM | 42.8 | IBM | 9.65 |
| | RWTH | 39.4 | RWTH | 9.38 |
| | **UPC** | 38.3 | ITC-irst' | 9.21 |
| | ITC-irst' | 37.9 | **UPC** | 9.15 |
| | LIMSI | 36.6 | LIMSI | 8.71 |
| | SystrP | 33.8 | SystrP | 8.58 |
| | UKA | 33.0 | UKA | 8.53 |
| Verbatim | IBM | 55.2 | RWTH | 10.94 |
| | RWTH | 55.1 | IBM | 10.91 |
| | ITC-irst' | 52.1 | ITC-irst' | 10.55 |
| | **UPC** | 52.0 | **UPC** | 10.45 |
| | UW | 48.0 | UW | 9.85 |
| | UKA | 46.0 | UKA | 9.85 |
| | LIMSI | 46.0 | LIMSI | 9.76 |
| | SystrP | 45.3 | SystrP | 9.68 |
| | DFKI | 42.2 | DFKI | 9.33 |
| Text | IBM | 54.1 | IBM | 10.77 |
| | RWTH | 53.1 | RWTH | 10.65 |
| | UW | 52.8 | **UPC** | 10.60 |
| | ITC-irst | 52.4 | ITC-irst | 10.56 |
| | **UPC** | 52.3 | UW | 10.55 |
| | EDINBG | 51.9 | EDINBG | 10.48 |
| | UKA | 47.0 | UKA | 9.98 |
| | SystrP | 45.7 | SystrP | 9.72 |
| | DFKI | 43.0 | DFKI | 9.47 |

**Table 4.38:** *Results of the 2nd TC-STAR evaluation campaign. BLEU is presented as a percentage.*

UPC *N*gram-based SMT system included two novel features. On the one hand, a preprocessing block reordering strategy capable of swapping the order of two input word according to

the alignment of their respective occurrences in training. This strategy, which is introduced in [Cj06], was applied to Spanish→English translation only.

On the other hand, the system included an additional feature consisting of a language model of the Part-Of-Speech sequence of the produced target sentence. This feature will be discussed in detail in §6.3.

| | English→Spanish | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Fluency | | Adequacy | | BLEU | | NIST | |
| ASR | RWTH | 3.06 | RWTH | 3.13 | ITC-irst' | 36.0 | ITC-irst' | 8.75 |
| | **UPC** | 3.04 | **UPC** | 3.09 | RWTH | 35.9 | RWTH | 8.72 |
| | IBM | 3.04 | ITC-irst' | 3.09 | IBM | 35.8 | IBM | 8.62 |
| | ITC-irst' | 2.99 | IBM | 3.05 | **UPC** | 34.8 | **UPC** | 8.56 |
| | UKA | 2.84 | UKA | 2.97 | UKA | 31.3 | UKA | 8.10 |
| | SystrP | 2.09 | SystrP | 2.33 | SystrP | 23.9 | SystrP | 7.03 |
| Verbatim | RWTH | 3.38 | RWTH | 3.55 | ITC-irst' | 46.6 | ITC-irst' | 9.91 |
| | **UPC** | 3.38 | **UPC** | 3.54 | RWTH | 45.4 | RWTH | 9.71 |
| | ITC-irst' | 3.35 | ITC-irst' | 3.54 | IBM | 45.4 | IBM | 9.66 |
| | IBM | 3.35 | IBM | 3.51 | **UPC** | 44.1 | **UPC** | 9.50 |
| | UW | 3.13 | UW | 3.43 | UW | 43.6 | UW | 9.36 |
| | UKA | 3.07 | UKA | 3.37 | UKA | 40.1 | UKA | 9.08 |
| | SystrP | 2.34 | SystrP | 2.77 | SystrP | 33.0 | SystrP | 8.10 |
| Text | EDINBG | 3.62 | EDINBG | 3.79 | ITC-irst | 49.8 | ITC-irst | 10.23 |
| | RWTH | 3.58 | RWTH | 3.74 | EDINBG | 49.5 | RWTH | 10.16 |
| | IBM | 3.50 | **UPC** | 3.69 | RWTH | 49.4 | EDINBG | 10.11 |
| | **UPC** | 3.48 | ITC-irst | 3.67 | UW | 48.8 | UW | 10.03 |
| | ITC-irst | 3.46 | UW | 3.62 | **UPC** | 48.2 | **UPC** | 10.00 |
| | UW | 3.40 | IBM | 3.60 | IBM | 47.7 | IBM | 9.93 |
| | DFKI | 3.31 | DFKI | 3.53 | UKA | 44.0 | UKA | 9.56 |
| | UKA | 3.17 | UKA | 3.49 | DFKI | 36.3 | DFKI | 8.70 |
| | SystrP | 2.46 | SystrP | 2.93 | SystrP | 36.3 | SystrP | 8.57 |

**Table 4.39:** *Results of the 2nd TC-STAR evaluation campaign. BLEU is presented as a percentage.*

Results are summarised in Tables 4.38 and 4.39, where ASR Word Error Rate is 6.2% for Spanish input and 6.9% for English. LIMSI stands for LIMSI-CNRS Paris (France), SystrP for Systran Product (not Systran Research) and DFKI for German Centre for Artificial Intelligence (Germany).

Results obtained by UPC *N*gram-based SMT system are competitive with results from other sites. In fact, nearly 6 systems perform very similarly, with slight differences depending on translation direction and input data. The novel reordering strategy applied to Spanish→English does not seem to yield a remarkable boost in terms of system ranking. On the other hand, human evaluation conducted for English→Spanish seems to give UPC system slightly better comparative results than automatic evaluation scores.

In addition to these tasks, a complementary Spanish→English task was included in this evaluation for portability assessment. This data consisted of transcriptions from Spanish Parliament, for which no parallel training was provided. As results and system ranking correlate strongly with EuParl data[9], they are not reported here.

Further details on this evaluation campaign specification and results can be found in [Ney06], whereas more details on UPC experiments are reported in the following publication:

- [Mn06b] J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, M.R. Costa-jussà and M. Khalilov, "UPC's Bilingual N-gram Translation System," in *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, pps. 43–48, June 2006.

### HLT/NAACL 2006 Workshop

In June 2006, a Workshop on Statistical Machine Translation of the HLT/NAACL conference (Human Language Technologies / North American Chapter of the Association for Computational Linguistics) organised a shared task following the guidelines of ACL 2005 workshop discussed above. Training material consisted of the same subset of the European Parliament Proceedings Corpus used in 2005. This time language pairs were French↔English, German↔English and Spanish↔English, making a total of 6 translation directions. Apart from an in-domain test set, an out-of-domain test set was also used.

In this evaluation, UPC *N*gram-based system introduced a novel reordering strategy described in [Cre06b]. This strategy extends the monotone search graph with a limited number of promising source-reordered hypotheses. The decision as to produce a new reordered path is taken according to reordering patterns extracted from training alignments and source Part-Of-Speech sequences. [Cre06b] reports a significant performance increase at nearly no additional computational cost.

Some results are summarised in Table 4.40. Among new acronyms, LCC stands for Language Computer Corporation (USA), UTX for University of Texas (USA) and MS for Microsoft Research (USA). Again, a lowercased letter indicates a different system submitted by the same site[10]. The table shows that UPC results belong to the group of 5 leading sites in terms of BLEU score for this task, which perform very similarly. This is confirmed by the varying ranking order depending on translation direction and input test.

Further details of this evaluation campaign, including automatic and human evaluation scores for all translation directions and test sets, can be found in [Koe06]. UPC *N*gram-based SMT

[9] Generally, lower performance is observed for all systems, reflecting the test-train data mismatch, as well as increased ASR WER rates, from 6.2% to 9.8%.

[10] Regarding UPC, 'UPCm' and 'UPCj' stand for two phrase-based SMT systems.

| in-domain test | | | | out-domain test | | | |
|---|---|---|---|---|---|---|---|
| Spanish→English | | English→Spanish | | Spanish→English | | English→Spanish | |
| Site | BLEU | Site | BLEU | Site | BLEU | Site | BLEU |
| LCC | 31.46 | UPCm | 31.06 | **UPC** | 27.92 | UPCm | 26.62 |
| NTT | 31.29 | NTT | 30.93 | UTX | 27.41 | NTT | 26.52 |
| UTX | 31.10 | UTX | 30.73 | LCC | 27.18 | MS | 26.15 |
| **UPC** | 31.01 | **UPC** | 30.44 | NTT | 26.85 | **UPC** | 25.59 |
| RALI | 30.80 | NRC | 29.97 | UPCm | 25.62 | NRC | 25.58 |
| NRC | 30.04 | MS | 29.76 | NRC | 25.40 | UTX | 25.26 |
| UPCm | 29.43 | RALI | 29.38 | EDINBG | 25.20 | RALI | 24.03 |
| EDINBG | 29.01 | EDINBG | 28.49 | RALI | 25.03 | EDINBG | 23.18 |
| UPCj | 28.03 | UPCj | 27.46 | UPCj | 23.42 | UPCj | 22.04 |
| UPV | 23.91 | UPV | 23.17 | UPV | 19.17 | UPV | 16.83 |

**Table 4.40:** *Some results of NAACL 2006 Workshop. BLEU is presented as a percentage.*

system participation is reported in the following publication:

- [Cre06a] J.M. Crego, A. de Gispert, P. Lambert, M.R. Costa-jussà, M. Khalilov, R. Banchs, J.B. Mariño and J.A.R. Fonollosa, "N-gram-based SMT System Enhanced with Reordering Patterns," in *Proceedings of the HLT/NAACL Workshop on Statistical Machine Translation*, pps. 162–165, June 2006.

Most recently, UPC *N*gram-based system participated in NIST 2006 MT evaluation, both for Chinese→English and Arabic→English large-data tasks, and IWSLT 2006 evaluation campaign, in all translation directions. As official results are being or have not been released yet, these are not reported here.

Further details on the description of these tasks, as well as UPC participation will be easily accessible in the corresponding publications.

## 4.5 Chapter Summary and Conclusions

This chapter introduced in detail *N*gram-based SMT, a maximum-entropy approach to provide the bilingual *N*-gram translation model from chapter 3 with additional feature functions in order to obtain state-of-the-art MT performance. All the training stages are thoroughly explained, as well as the optimisation procedure.

Following the tendency of the field towards a log-linear combination of feature functions whose respective weights need to be optimised according to automatic evaluation measures on a development data set (minimum-error training), this chapter showed that additional features do certainly improve the bilingual *N*-gram model performance.

In particular, the contribution of lexical models based on IBM model 1 probabilities achieve the biggest quality improvement. Their impact is double: on the one hand, they represent a constant tuple ranking which is complementary to the bilingual *N*-gram model; and on the other, they tend to favour shorter tuples, thus producing less number of words. These effects seem to boost automatic evaluation scores significantly.

Additionally, the target language model (which is combined with a word bonus model) enlarges the translation context on a monolingual scale. In other words, whenever the bilingual context is useless, the target language model can help produce a grammatically correct target sentence, since the target language model is less sparse.

Finally, the combination of the translation model plus all these four models yields the best performance. In order to illustrate each model contribution, a study of examples is presented, in which some bad examples (whose translation is worse when feature combination is at play) are purposely introduced.

When it comes to tuple segmentation, the contrastive impact of the different strategies from §3.4.2 appears reduced under the optimised combination framework, hinting that features may help compensate for bad segmentation decisions. Apart from that, experiments on tuple pruning show that it is more adequate to prune given the log-linear combination than only taking the bilingual *N*-gram model into account, as done in §3.3.1.2.

When manually studying the system outputs (for the English↔Spanish pair) and when comparing the system to other MT systems (in many evaluation campaigns), the following conclusions can be drawn:

- Despite the positive impact of features in translation quality, many errors still exist
- The system falls short of words, systematically outputting less words than the number of input words. Omission errors therefore demand for future studies on modelling tuples

translating to NULL

- Verb forms are sparse and hard to translate correctly, especially when target is Spanish (or any rich language in terms of morphology)
- The system obtains a very good performance in monotone tasks; however, word order errors persist and are relevant even in these cases. For language pairs in need of severe word order modification, the system looses competitiveness

As noted in the publications already mentioned, a big amount of work from this chapter has been co-authored with the members of UPC SMT group. Apart from the papers being referred to throughout the chapter, the *N*gram-based SMT system is further reported in the following journal publication:

- [Mn06a] J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa and M.R. Costa-jussà, "N-gram Based Machine Translation," Accepted for publication in *Computational Linguistics*, December 2006.

# Chapter 5

# Linguistic Knowledge into Word Alignment

## 5.1 Introduction

This chapter presents a set of experiments conducted with the aim of improving Alignment Error Rate by introducing a certain degree of linguistic knowledge into the statistical training of SMT systems.

Mainly two approaches were followed; on the one hand, a one-to-one cooccurrence-based word alignment algorithm was implemented and extended with linguistic phrases; and on the other, word alignment based on IBM models (ie. using GIZA++) was extended by several linguistically-guided classification schemes affecting input words.

All experiments were carried out in Spanish↔English tasks, with large variations in corpus size and domain. In all tests, word alignment quality is assessed by measuring Alignment Error Rate (see §2.5.1) against a human gold standard reference set. For each corpus, a reference on the development of the gold standard set will be made.

The chapter is organised as follows:

- §5.2 reports the development of a word alignment tool based on word cooccurrences and link probabilities, and its extension to many-to-many links by including certain linguistic phrases, namely verb forms, regular time and data expressions and idiomatic expressions.
- Given the positive results provided by this many-to-many extension, IBM-based alignment can also be constrained by previously classifying verb forms in an analogous way. This is studied in §5.3.
- Following this classification principle, §5.4 thoroughly evaluates the impact of various word classes (including base forms, stems and morphological word derivations) on alignment

error rate for both large- and small-data tasks. Furthermore, a discussion on correlation with translation evaluation scores is also included.

Lastly, §5.5 closes this chapter with a summary of discussed topics and main conclusions.

## 5.2 Cooccurrence-based word alignment extended with linguistic phrases

### 5.2.1 Related work

Word alignment systems based on IBM models suffer from two structural flaws that pose a severe limitation to their performance. Due to the model definition of alignment as a function from positions in the target sentence to positions in the source sentence, the result is strictly asymmetric, generating one-to-many word alignments that do not account for many translation phenomena. Several kinds of symmetrisation heuristics (all of them linguistically blind) have been proposed to deal with this effect, seeking the most accurate result to pass onto posterior phrase-based translation systems. Furthermore, the mathematical complexity of IBM models and their overload of parameters to estimate make it very hard to introduce linguistic information into this setting in a comfortable way, although some efforts have been done in [Tou02].

From a considerably different standpoint, a word alignment can be produced using information on word cooccurrences and link probabilities, as was introduced in [Che03a]. The relative simplicity of this approach, its flexibility to introduce more knowledge sources, its structural symmetry and its promising results shown in [Mih03] make it attractive despite its dependence on empirical data and tuning strategies. However, the most important disadvantage of the approach is the one-to-one constraint, producing high precision alignments with low recall, limiting thus its use in practical translation systems.

In the face of this, a novel alignment strategy was proposed during the course of this Ph.D. research work. The strategy is also based on bilingual cooccurrences, but aims at finding phrase-to-phrase alignments directly from the corpus cooccurrence counts by using linguistic knowledge, thereby overcoming the one-to-one limitation. This knowledge is introduced by means of very simple rules made by non-linguists.

A similar approach is followed in the framework of example-based machine translation in [Bro99], in order to improve the coverage of the examples during translation. As it will be detailed in the following sections, in our case the rules provide a classification that leads to improved statistical word alignment performance.

Another approach to directly generate phrase alignments from the corpus without symmetrising IBM-based alignments was presented in [Mar02]. In contrast to their *open* (and computationally costly) approach, here very high-confidence links between phrases are performed before proceeding onto word alignment, based on linguistic knowledge. This search for phrase links is limited to a small adequate set of possible phrases, as discussed below.

### 5.2.2 Word and phrase association measures

Association or cooccurrence measures extracted from parallel corpora give strong evidence of so-called translation equivalence [Mel01], or simply alignment adequacy, between a pair of phrases or words. Among these measures we find Dice-score, $\phi^2$ score and some others, offering a similar performance. In this work, we used $\phi^2$ as presented in [Gal91], which is defined by the following equation:

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{5.1}$$

where $a$ is the number of times that two words (or phrases) cooccur in a document (the addition of the cooccurrence count for each sentence), $b$ and $c$ are the number of sentences where one word occurs and the other does not, and finally $d$ accounts for the number of sentences neither one nor the other occur in the data set. All these counts are defined at document level, meaning that they are the addition of the count for each sentence.

In our implementation, we defined the cooccurrence of a word (or phrase) appearing $x$ times in a sentence and a word (or phrase) occurring $y$ times in its translation as $min(x, y)$, for two reasons: on the one hand, the alternative option given by the product $xy$ leads to confusing results when computing $b$ and $c$, as these can be negative because the times a word cooccurs with another can overweight the total occurrences of the word. On the other hand, the word alignment algorithm used estimates link probabilities from existing one-to-one links (see §5.2.3.3), being $min(x, y)$ the maximum number of links that can be established between the two words (in which case their probability is the highest). This way stochastic consistency is preserved.

Even though this score can be easily computed for each possible pair of words from both languages, computational problems arise when dealing with every bigram, trigram or, in general, phrase for each language. However, these scores can convey a useful complementary information in many cases.

To illustrate this idea, consider the examples of Table 5.1, where the phrase-to-word score -in bold face- is comparatively much better than all word-word scores[1] for all words involved in Spanish idioms '`por favor`' and '`a lo mejor`'. Furthermore, it is reasonable to expect that the longer the phrases considered, the stronger the evidence of a correct alignment, so long as we have a reasonable number of occurrences of the phrase.

The main problem is then the practical impossibility to compute all combinations for even relatively small corpora. To tackle this, one can try to extract as much useful information from these phrase cooccurrence measures by performing a selection of only a subset of all possible

[1] Note that $-10log$ is assumed when referring to $\phi^2$

| | please | |
|---|---|---|
| por | 22.4 | **0.9** |
| favor | 1.2 | |

| | maybe | |
|---|---|---|
| a | 23.1 | |
| lo | 18.2 | **8.0** |
| mejor | 12.2 | |

**Table 5.1:** *Examples of $\phi^2$ scores between words and phrases.*

phrases. This selection, which is made using linguistic criteria, was used in the phrase alignment strategy presented next.

### 5.2.3 A phrase alignment strategy

We propose a phrase alignment strategy in four stages, as shown in Figure 5.1.

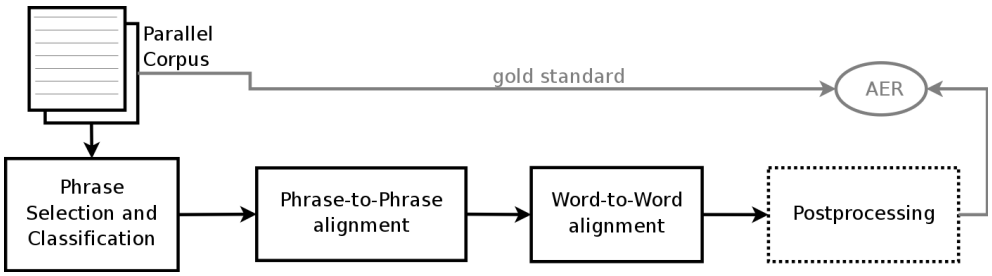



**Figure 5.1:** Flow diagram of the proposed phrase-based word alignment strategy.

Firstly, from all possible sets of words we select a small set of 'interesting' phrases for each language. This selection is linguistically guided and should produce a set of phrases containing words that play a unique semantic role.

Secondly, a high-precision phrase alignment algorithm links these phrases together (with phrases or single words) using cooccurrence measures, and discarding uncertain links.

After these phrase alignments have been produced, we run a word-to-word alignment algorithm based on link probabilities and shallow syntactic information in the fashion of [Che03a]. This stage takes advantage of the complexity reduction derived from the previous linking.

Finally, in the fourth stage a postprocessing takes final decisions on unaligned words with certain linguistic criteria. As illustrated in Figure 5.1, Alignment Error Rate is then computed against the gold standard reference set.

Details on each of these four blocks are given next.

#### 5.2.3.1 Phrase selection and classification

The objective of this stage is double. Given the exponential nature of the amount of different phrase cooccurrences shown above, which makes it impossible to work with cooccurrences between all combinations for each language, a first objective is the reduction of this huge space to those being 'interesting' from an alignment/translation point of view. Our criterion is one of so-called translational equivalence, so we define as interesting those phrases expressing a same concept or being semantically linked in one language, as it is reasonable to expect that these might be aligned to (or might translate into) a single word (or phrase) in another language.

On the other hand, a semantic classification of these phrases should improve cooccurrence measures by adding different instantiations of the same concept to a same measure. As for the selection of phrases, we have followed a linguistically-guided strategy. Specifically, we have implemented three selection criteria using complementary knowledge.

Firstly, we detect **verb phrases** using deterministic automata that implement a few simple rules using word forms, POS-tags and word lemmas as input, and mapping the resulting phrase to the lemma of the head verb (see Figure 5.2 for some rules and examples of detected verb phrases). This way, the classification improves cooccurrence counts for verb phrases no matter how their full form is expressed, as long as they share the same base form of the head verb. Therefore, forms like

```
we have brought
will we bring
I would not bring
she probably brings
```

are considered equivalent and add a cooccurrence count for the base form '`bring`', increasing its evidence and reducing evidence for function words like '`have`' and '`will`' that act as modifiers and may therefore be expressed in many ways in the other language, as they do not convey a stand-alone meaning in the sentence.

One can expect this to produce a special gain in languages using heavily inflected verb forms, like languages belonging to the Romance family (as Spanish).

The automata allow detection of non-sequential verb phrases, that is, phrases containing words that will not be part of the selected phrase, and therefore will not be linked together (like the underlined words in the last two examples above). At the moment, we restrict this case to adverbs modifying English verb phrases (or negative forms), but other linguistic phenomena could be tackled similarly, such as separable phrasal verbs in English, provided a list of these verbs is available.

Since the rules detect only phrases with at least two tokens, all remaining words with

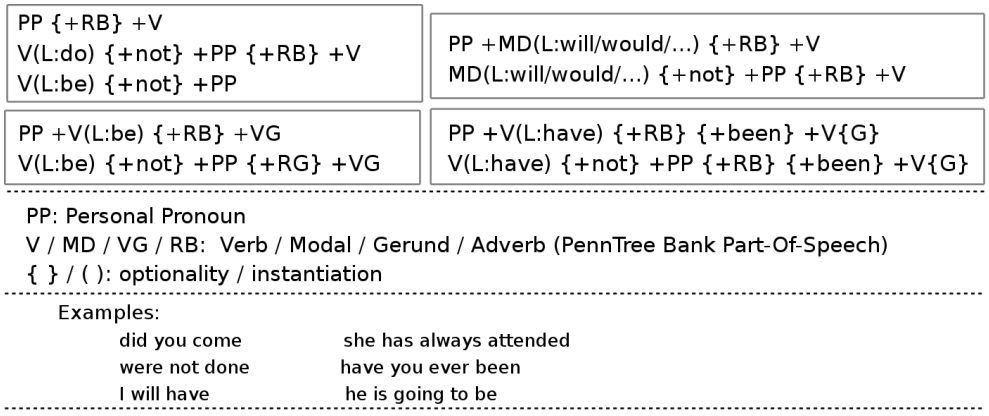



**Figure 5.2:** Some verb phrase detection rules and detected forms in English.

POS-tag of a Verb are substituted by their base form to enforce the verb's cooccurrence evidence, after detection rules have been applied. This way, in the following example,

`Stuart` **will come** `if Henry` *comes* `too`
`Stuart` *vendrá* `si Henry` *viene* `también`

only the phrase '`will come`' -in bold face- is detected (and classified to lemma '`come`'), but words '`comes`', '`vendrá`' and '`viene`' -in italics- are also substituted by their base forms '`come`', '`venir`' and '`venir`', respectively, improving the cooccurrence measure between the pair '`come,venir`' during the word-alignment stage.

Of course, this is only possible if tagging gives a Verb tag for each of these tokens. Lemmatisation ambiguity did not affect the corpus used in our experiments.

Secondly, we also implemented a selection based on **idiomatic expressions**. Specifically, the corpus was matched against a list of 1497 and 50 usual idiomatic expressions, available from the *FreeLing* language analysis tool [Car04] for Spanish and English, respectively.

These expressions (containing examples such as '`on the other hand`' for English, and '`sin embargo`' or '`a lo mejor`' for Spanish) tend to convey a single meaning and we can expect them to be aligned *together* to one or more words in the other language. No additional dictionary was used, so these expressions were not classified according to their meaning.

Finally, **date expressions** are detected as well, by implementing a basic automaton detecting expressions like '`on Monday the tenth`', '`Sunday the fifth`' or '`on the second`' for English, and mapping the resulting phrase to a unified token, so that all date expressions contribute to the same cooccurrence measure with all the dates in the other language.

Furthermore, we keep the *value* of the expression (meaning the day of the week and day of the month) for further use in the alignment phase discussed in the next section. Similarly for Spanish, expressions like '`el lunes día diez`' or '`el domingo cinco`' are detected.

Other possible linguistically-guided selection rules could include more regular expressions such as numbers or times of the day (that could also be classified) or even collocations, phrasal verbs or more complex structures. As this selection is language-dependent, every language will define its own adequate rules.

If no linguistic knowledge is available, statistical procedures could also be used to obtain a set of possible phrases. For example, selecting the N most frequent bigrams, trigrams and Ngrams in general, or the ones having a very high bigram, trigram or Ngram probability (defining phrases of words that consistently appear together in the text).

It is important to note that we do not expect this selection to be exhaustive, nor does it imply that the selected phrase will necessarily be linked *together* at the next stage (it is not a hard decision in terms of alignment). It is the phrase alignment stage that decides whether a phrase should be linked together, or whether the words should be left free to be linked word-to-word.

#### 5.2.3.2 Phrase alignment

In this stage cooccurrence measures are computed for each selected phrase in one language against all selected phrases and single words in the other language. Then, a competitive linking strategy is used, but not until all words or phrases are linked, but until a certain threshold is surpassed.

Basically, this greedy strategy produces an alignment solution by iteratively choosing the link with best phrase-phrase or phrase-word cooccurrence measure as long as this is better than the threshold [Mel01]. Links are only selected if they link positions which have not been linked before.

This strategy relies on the fact that phrase cooccurrence measures are a stronger evidence of translational equivalence than word, and the threshold (which has to be empirically tuned) ensures that we generate only the highest-confidence links. This way, not all selected phrases will be linked, but only those having a high cooccurrence evidence in the data.

Once the linking of two phrases is decided, one can use several strategies to determine the internal links between words inside the phrases, if that is desired. For example, internal links can be solved using the general word alignment algorithm, but restricting the search inside the phrase positions.

However, in our case we have introduced all internal links between linked phrases, as this is a consistent alignment given the manual reference. In the case of 'date' expressions, we constrain their alignment to those having equal values (same day of the week and same day of the month).

After the initial phrase alignment, an one-to-one word alignment algorithm tries to link all unaligned tokens (including tokens inside unaligned detected phrases, which will be now considered separately). The details are presented in the following section.

#### 5.2.3.3 Word alignment

As for the word alignment algorithm, we implemented an iterative algorithm similar to the one presented in [Che03a]. Basically, an initial alignment is generated using word cooccurrence measures, from which link probabilities are estimated. Then, a best first search is performed, following an heuristic function based on the global aligned sentence link probabilities.

The search is further improved with a syntactic constraint (also called cohesion constraint [Che03b]) and can introduce features on the links, such as a dependence on adjacent links. Our implementation allows certain positions to be prohibited, so that previous phrase alignment is fixed, although its links also contribute to the link probability estimation at each iteration.

Given the enormous space of possible word alignments to choose from, the heuristic function becomes the key to efficiency, so long as it is correctly defined. Basic parameters are:

- the initial **null probability**, or the probability that a word links to null (no word), which is necessary to make fully- and partially-aligned solutions comparable
- and the **minimum score** to accept a link between two words (hereafter referred to as *mscore*)

These parameters must be set empirically for the optimal performance of the algorithm. We also found that restricting the search of possible links to a window in the other language not only made the algorithm much more efficient (turning it from exponential time to linear time with input sentence lengths), but also improved results by discarding the ambiguities generated by the repetition of frequent words (mostly function words).

We define this window in the neighbourhood of the diagonal defined by the division between both sentence lengths. Of course, this window is completely dependent on the pair of languages considered (might even be eliminated for certain pairs), but in our case (English–Spanish) and with the corpus used turned out to be optimal considering 8 words.

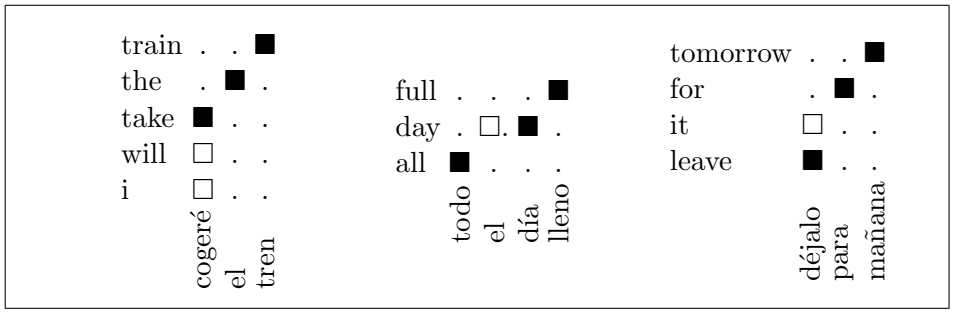



**Figure 5.3:** *Three linguistically-guided postprocessing strategies: completing selected phrases with only one token aligned (left), alignment of Spanish unaligned articles together with the noun they precede (centre) and alignment of English personal pronouns following verbs linked to Spanish verbs with pronoun (right)*

#### 5.2.3.4 Postprocessing

As postprocessing, three basic strategies based on linguistic criteria were implemented. First of all, we automatically align together those phrases selected during the first stage that have been left unaligned except for one of their tokens. This aligned token defines the alignment that is automatically generated for the other tokens of the phrase. The rationale behind this is to recover those detected phrases that did not have a good cooccurrence measure during the phrase alignment stage and were not linked, but got one token linked during posterior word alignment.

On the other hand, we also look for those Spanish articles preceding a noun which is linked to an English word, according to POS tags. For all these tokens, we automatically generate a link from the Spanish article to that English word, as these common Spanish articles are sometimes omitted in English. Finally, a similar approach is followed with English personal pronouns following a verb. If this verb is aligned to an Spanish verb form containing a pronoun, we conclude that the English pronoun must be aligned to the Spanish verb form as well. Figure 5.3 shows graphically one example of each of these postprocessing strategies (where black and white boxes represent links produced before and during postprocessing, respectively), whereas their impact is evaluated empirically in §5.2.4.6.

The postprocessing stage could take other final alignment decisions using sentence-level information (ie. deciding whether unlinked words should be linked, looking for long-distance links, reconsidering the links for a word/phrase given all its links in all sentence pairs, etc.). Ideally, it could also feedback into the phrase selection/alignment blocks to reconsider previous decisions using global information of all sentences. Undoubtedly, this stage is strongly connected to the posterior translation model. Although alignment can be and must be evaluated separately, we are of the opinion that it is not completely independent from the translation model.

### 5.2.4 Experimental work

#### 5.2.4.1 Experiment setup

Experiments with the alignment strategy described above were conducted on English–Spanish VerbMobil parallel corpus, whose main statistics were already shown in Table 3.8. Data preprocessing included:

- Normalisation of contracted forms for English (ie. wouldn't = would not, we've = we have) and Spanish (del = de el)
- English data tagging using freely-available *TnT* tagger [Bra00], and base forms were obtained using *wnmorph*, included in the WordNet package [Mil91].
- Spanish data tagging using *FreeLing* package already mentioned. This software also generates a lemma or base form for each input word.
- Regular date and time expressions (numerous in this domain) were substituted by a unified tag using a semi-automatic technique [Gis03]. Specifically, only dates containing a month of the year in both languages were substituted, in contrast to date expressions dealt with in section 5.2.3.1.
- Finally, punctuation marks were left out.

In order to assess the quality of the algorithm using Recall, Precision and AER measures, we randomly selected from the corpus *two sets*: a validation set of 100 sentences (for tuning of parameters) and a test set of 400 sentences.

These sets were manually aligned, following the criterion of producing Possible links *only* when they allow combinations which are considered equally correct, as a reference with too many Possible links suffers from a resolution loss, causing several different alignments to be equally rated. The result was that 80% of the links were Sure and 20% were Possible.

Figure 5.4 shows an example of a manual reference alignment used in evaluation, where two alignments between verb phrases can be found ('`I said - he dicho`', '`you would send - tú enviarías`').

As the interest was in aligning full verb phrases (with their pronoun, if any) as single meaning-bearing units during phrase alignment, the criterion used in reference was to separate verb phrases into two parts, that is, personal pronoun (if any) and remaining tokens. These parts were aligned with Sure links to their counterparts in the other language, whereas Possible links were introduced between personal pronouns in one language and remaining tokens in the other (and vice versa). Introducing these latter links as Sure would favour our alignment result, since

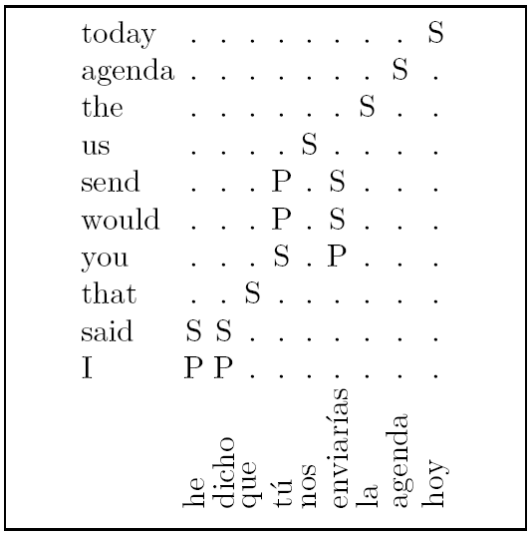



**Figure 5.4:** Example of manual reference alignment (S for Sure links and P for Possible links).

we produce all links between all tokens inside verb phrases. Therefore, the Possible links in the reference allow the generation of all internal links without falsely improving AER results with a misleading boost in Recall.

#### 5.2.4.2 Phrase alignment results

In this section, we evaluate separately the phrase selection, classification and alignment blocks described in sections 5.2.3.1 and 5.2.3.2. First we present results of the phrase selection and alignment stage previous to word alignment. Specifically, statistics on phrase selection and alignment are provided when dealing with verb phrases, idiomatic expressions and date expressions separately. Finally, complete alignment results are reported (phrase alignment + word alignment + postprocessing), comparing performance against state-of-the-art word alignment models.

#### 5.2.4.3 Verb phrases

Verb phrase detection rules include 14 basic rules for English language and just 6 for Spanish, which usually employs inflected verb forms omitting thus personal pronouns and using thus a single token. Verb phrase rules have detected a total of:

- English: 20,556 verb phrases (1,907 different), classified into 276 different verb lemmas
- Spanish: 2,007 verb phrases (674 different), classified into 190 different verb lemmas

In the case of English, 5% of detected phrases are 'gapped' in that they include adverbs modifying the verb, which will not be aligned together with the verb phrase (see section 5.2.3.1).

Note that these figures include only phrases with more than one token (which are linked in the phrase alignment stage). For this reason, we get such a big difference between languages, given the usual omission of personal pronoun in Spanish by using inflected single-word verb forms. However, if we include also words with POS-tag of Verb that are substituted by their base form before word alignment (see section 5.2.3.1), we detect:

- English: 36,556 verb forms (2,584 different), classified into 482 different verb lemmas, and representing ∼31% of all English tokens
- Spanish: 37,605 verb forms (2,763 different), classified into 539 different verb lemmas, and representing ∼20% of all Spanish tokens

Results of the phrase alignment with only verb phrases for the development set are shown in the upper rows of Table 5.2, changing the value of the threshold to accept phrase links from more restrictive to less restrictive. A restriction that the linked pair cooccurs at least twice has also been used. The last two columns correspond to the percentage of verb phrases linked over the total number of phrases detected in the train set for English and Spanish, respectively. The margin of confidence at the 95% for Recall and Precision measure is 1.0%.

| | Recall | Precision | LkEng | LkSpa |
|---|---|---|---|---|
| Verb phrases $\phi_v^2 < 5$ | 6.00 | 98.94 | 51% | 40% |
| Verb phrases $\phi_v^2 < 10$ | 11.06 | 98.97 | 81% | 77% |
| Verb phrases $\phi_v^2 < 15$ | 12.14 | 98.92 | 87% | 86% |
| Verb phrases $\phi_v^2 < 20$ | 12.64 | 96.93 | 93% | 91% |
| Idioms $\phi_v^2 < 5$ | 2.19 | 98.61 | 19% | 47% |
| Idioms $\phi_i^2 < 10$ | 3.24 | 99.06 | 32% | 77% |
| Idioms $\phi_i^2 < 15$ | 3.68 | 97.54 | 63% | 84% |
| Idioms $\phi_v^2 < 20$ | 3.92 | 95.59 | 89% | 90% |
| Date exprs. $\phi_i^2 < 5*$ | 4.11 | 98.82 | 88% | 86% |

**Table 5.2:** *Separate phrase alignment results for development data set.*

The proposed selection strategy provides very positive results, as Precision is consistently nearly 99 % for the three most-restrictive thresholds used, whereas Recall is just over 12% and nearly 87% of all selected blocks are being linked with a high precision.

We have to keep in mind that these phrase links will necessarily boost Recall with respect to the isolated word aligner, as it is a one-to-one algorithm, unable to produce these links. As about Precision, the high figures are due to the greater statistical evidence of phrase cooccurrence measures with respect to single word cooccurrences.

Given the significant descent in Precision without significant improvement in Recall setting the threshold to 20, a threshold of 15 has been selected as optimal for further experiments.

#### 5.2.4.4 Idiomatic expressions

Regarding idiomatic expressions, the phrase selection stage detects the following cases by matching with the available lists presented in §5.2.3.1:

- English: 525 idiomatic expressions (21 different), representing ∼1% of all English tokens
- Spanish: 2,760 idiomatic expressions (98 different), representing ∼3% of all Spanish tokens

The results when aligning only idiomatic expressions phrase-phrase and phrase-words links are shown in the middle rows of Table 5.2, again for different thresholds.

In this case, although the impact in Recall is much smaller than when considering verb phrases, two points are worth raising. First, we have again a nearly error-free alignment using a relatively small set of phrases (with thresholds up to 15). And second, but not less important, that we expect this Recall to complement the previous experiment and further boost the global alignment Recall, as we find no verb phrases among the idiomatic expressions considered. Again, a threshold of 15 has been selected as optimal for further experiments.

#### 5.2.4.5 Date expressions

Finally, regarding date expressions, which are all classified into one single class, the selection stage detects:

- English: 4,221 date expressions (336 different), representing ∼5% of all English tokens
- Spanish: 4,303 date expressions (410 different), representing ∼5% of all Spanish tokens

The result when aligning only date expressions is shown in the bottom-most row of Table 5.2. This result does not change when increasing the threshold. As we only allow links between date expressions having the same 'value' (same day of the week and same day of the month), this result is constant once surpassed the cooccurrence measure between the tags 'date' of each language. Note that the 12% of expressions not linked correspond to not respecting the aforementioned constraint.

#### 5.2.4.6 Complete alignment results and discussion

All in all, when aligning verb phrases, idiomatic expressions and date expressions before word alignment (using each optimal threshold), we achieve a **19.93%** of Recall with an outstanding

**99.03%** Precision. This means that nearly 20% of the links are already done before proceeding to the statistical one-to-one word aligner. Now the complete alignment results are presented.

For comparison purposes, we aligned our data using GIZA++ from English to Spanish and vice versa (performing 5 iterations of model IBM1 and HMM, and 3 iterations of models IBM3 and IBM4), and evaluated two symmetrisation strategies, namely the union and the intersection. Another symmetrisation technique presented in [Och00b] was also tested, without improvement over the union given this data and reference.

The obtained alignment results are shown in the first four rows of Table 5.3. Note that the giza++ union rates the best among them, an expected result given the proportion of Sure and Possible links in the manual alignment, which favours high-recall alignments. The margin of confidence at the 95% for Recall and Precision measures is 1.0%, and 0.8% for AER.

| | Recall | Precision | AER |
|---|---|---|---|
| giza++ eng2spa | 76.70 | 92.90 | 15.78 |
| giza++ spa2eng | 78.52 | 94.02 | 14.14 |
| giza++ union | **84.24** | 90.64 | 12.52 |
| giza++ intersection | 70.98 | **97.37** | 17.78 |
| one-to-one word align | 72.78 | 95.85 | 17.15 |
| full phrase align | 80.75 | 96.37 | **11.42** |

**Table 5.3:** *Comparison of final alignment results for test set.*

Complementarily, we also used the word alignment algorithm presented in section 5.2.3.3 to align the data without any kind of previous phrase selection and alignment, thus producing the one-to-one alignment shown in the fifth row. For this result, we ran three iterations with a $mscore = 30$, and three iterations further restricting it to 8 to achieve high precision, always using cohesion constrain (initial NULL cost being set to 15).

In contrast to giza++ intersection (the only baseline alignment that is also one-to-one and thus subject to a fair comparison), we observe a reduction in Precision and an increase in Recall, leading to a non-significant AER reduction, that is, comparable performance. However, both alignments skip around 30 % of good links (far below the other alignments recall), which make them unpractical for posterior statistical translation modelling.

The results of our phrase-based word aligner are shown in the 'full phrase align' row (including phrase selection with each of the phrase optimal thresholds and postprocessing). As it can be seen, it rates the best AER with 1.1 absolute points lower than the giza++ union (a relative 8.8% reduction).

Our alignment achieves a very high Precision level (just one point lower than giza++ intersection), while providing a Recall increase of over 10 absolute points with respect to intersection,

and of about 8 points with respect to the one-to-one aligner. The achieved Recall figure is still 3.5 points lower than that of the giza++ union (that nevertheless offers 6.3 points worse Precision results).

It is interesting to note the Precision improvement when comparing the word-to-word aligner with the phrase aligner proposed, which is due to two factors. On the one hand, the previous phrase alignment introduces links with a higher precision than that of the one-to-one aligner, and on the other hand, this previous linking results in a complexity reduction (less ambiguity) that simplifies the task of the one-to-one aligner, improving its performance.

| | Recall | Precision | AER |
|---|---|---|---|
| phrase align, no postrocess. | 78.71 | 97.20 | 12.35 |
| postprocess. (3 strategies) | 2.04 | 85.20 | |
| postprocess. strategy 1 | 0.80 | 83.33 | |
| postprocess. strategy 2 | 0.90 | 86.14 | |
| postprocess. strategy 3 | 0.40 | 77.78 | |

**Table 5.4:** *Impact of postprocessing strategies in AER results.*

To analyse the impact of the postprocessing in this result, Table 5.4 shows the AER without any postprocessing, which is not significantly lower than that of the giza++ union (with as high a Precision as the intersection). It is the addition of the three linguistically-guided postprocessing strategies presented in section 5.2.3.4 that provides a further boost to Recall for achieving the best AER.

To sum up, the presented alignment strategy improved state-of-the-art results while still making a relatively small use of linguistic knowledge. As its architecture is open to an easy introduction of more information, many other knowledge sources could be used. Furthermore, we performed the union between our results and the highest precision alignment, namely the giza++ intersection, obtaining a significant improvement in terms of Recall (over 3.5% absolute) and AER (1.3% absolute), as shown in Table 5.5.

| | Recall | Precision | AER |
|---|---|---|---|
| phr. align U giza++ intsct. | **84.36** | 95.11 | **10.10** |

**Table 5.5:** *Results when combining two highest-precision alignments.*

This result leads to two conclusions. On the one hand, the links produced by the proposed algorithm are complementary to those provided by the state-of-the-art statistical approaches, as linguistic knowledge plays a generalisation role where mere statistics are limited. And on the other hand, this means that there is still room for improvement, which could be achieved by introducing more linguistic knowledge, as mentioned in the next section.

Finally, an additional conclusion which can be drawn from this result points towards the

inclusion of verb detection rules into alignments based on IBM models. Given the existence of complementary information between alignments based on IBM models and our cooccurrence-based aligner (including full verb detection rules forcing their treatment as single alignment tokens), it is reasonable to expect that the former will possibly benefit from verb detection rules as well. This research line is documented in the following two sections.

## 5.3 Verb form classification for constraining IBM-based alignment

In this section an experiment of verb form classification for constraining IBM-based word alignment is reported. Basically the approach aims at leveraging the positive effect of full verb form classification from the previous section, where it produced a big boost in recall at nearly no precision cost.

For this, the same verb detection rules (implemented with a deterministic automaton) from §5.2.3.1 were used to unambiguously classify the English–Spanish parallel text before word alignment.

Again, it is worth mentioning that with these rules one can detect verbs containing adverbs and negations (underlined in Figure 5.2), which are ordered before the verb to improve word alignment with Spanish, but once aligned they are reordered back to their original position *inside* the detected verb, representing the real instance of this verb.

Experiments were carried out using the LC-Star Spanish–English parallel corpus (see detailed statistics in Table 3.4). Preprocessing included:

- Normalisation of contracted forms for English (ie. wouldn't = would not, we've = we have)
- English POS-tagging using freely-available *TnT* tagger [Bra00], and lemmatisation using *wnmorph*, included in the WordNet package [Mil91].
- Spanish POS-tagging using *FreeLing* analysis tool [Car04], which also generates a lemma or base form for each input word.

### 5.3.1 Verb Phrase Detection/Classification

Table 5.6 shows the number of detected verbs using the detection rules presented in §5.2.3.1, and the number of different lemmas they map to.

| | verbs | lemmas |
|---|---|---|
| Train set | | |
| English | 56419 | 768 |
| Spanish | 54460 | 911 |

**Table 5.6:** *Detected verb forms in LC-Star parallel corpus.*

In average, detected English verbs contain 1.81 words, whereas Spanish verbs contain 1.08 words. This is explained by the fact that we are including the personal pronouns in English and modals for future, conditionals and other verb tenses.

### 5.3.2 Word alignment results

In order to assess the quality of the word alignment, we randomly selected from the training corpus *350 sentences*, and a manual gold standard alignment was created with the criterion of Sure and Possible links, in order to compute Alignment Error Rate (AER) as described in §2.5.1, together with appropriately redefined Recall and Precision measures.

Word alignment was performed using GIZA++ [Och03a] from English to Spanish and vice versa (performing 5 iterations of model IBM1 and HMM, and 3 iterations of models IBM3 and IBM4), and evaluated two symmetrisation strategies, namely the union and the intersection, the union always rating the best. Once again, the refined symmetrisation technique presented in [Och00b] was also tested, without improvement over the union given this data and reference.

Table 5.7 compares the result when aligning words (current baseline), and when aligning classified verb phrases. In this latter case, after word alignment we substitute the class for the original verb form and each new word gets the same links the class had. Of course, adverbs and negations are kept apart from the verb and have separate links.

| | Recall | Precision | AER |
|---|---|---|---|
| baseline | 74.14 | 86.31 | 20.07 |
| with class. verbs | 76.45 | 89.06 | **17.37** |

**Table 5.7:** *Results in statistical alignment for the LC-Star corpus.*

Results show a significant improvement in AER, which proves that verbal inflected forms and auxiliaries do harm alignment performance in absence of the proposed classification.

The following section deepens into this approach by evaluating this classification strategy in a large-vocabulary task and comparing it to other classification approaches.

## 5.4 Linguistic classifications for IBM-based alignment

Extending the approach following in the previous section, here a thorough study of the contribution of linguistic information for classifying words before word alignment is reported. Experiments were carried out on the EuParl version 3 task, including an additional small-data corpus containing 1% of the whole Spanish–English parallel corpus.

This study was conducted in cooperation with Deepa Gupta (from ITC-irst in Trento), Maja Popović (from RWTH in Aachen) and Patrik Lambert (from UPC in Barcelona).

With the goal of finding out which linguistic features are relevant for improving statistical word alignment, we followed a corpus transformation approach, ie. data was modified using morphosyntactic information before word alignment, as shown in the flow diagram in Figure 5.5.

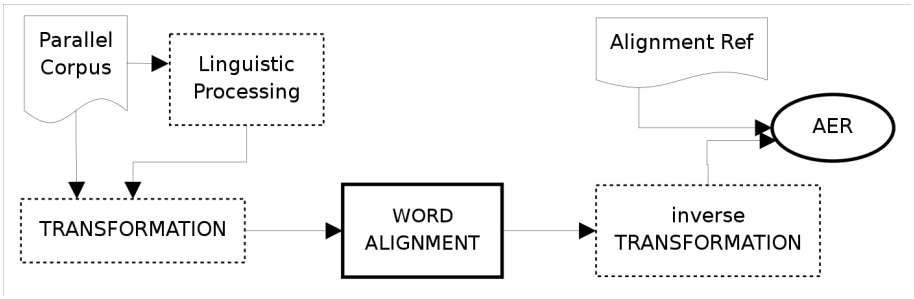



**Figure 5.5:** Experimental configuration to evaluate impact of using morphological information on word alignment.

Then, the obtained alignment of the transformed parallel corpus is mapped to the original sentence pairs in order to evaluate Alignment Error Rate against a manual reference. The same word alignment algorithm and configuration is used in all cases, therefore acting as a black-box.

In many cases, the corpus transformation can be seen as a classification from words to linguistically-enriched tokens, be it of all words or just some groups of words. However, we have also considered linguistically-motivated word order modifications, as well as combinations of both. This is equivalent to the approach followed in §5.3.

Two basic types of transformations have been considered, namely word classifications and word order modifications. Now each of these transformations is motivated and fully described.

### 5.4.1 Word Classifications

In general, word classifications aim at reducing data sparseness, by mapping some words to a unique token according to a certain criterion. In our case, criteria are based on the linguistic

information provided by state-of-the-art language tools, in the particular case of processing the Spanish and English languages.

#### Base forms

Also known as lemmas, base forms lack details on morphological derivation of the word (gender, number, tense, and so on) and only provide information on the head of the word. Therefore, they represent a meaning-bearing reduced version of each word, especially in the case of high morphological derivation, such as verbs, nouns or adjectives in Spanish. In English, verbs and nouns are also reduced by taking the base form, even though in lesser degree.

#### Stems

Same as lemmatisation, stemming is another method of word transformation which truncates inflected word forms by a single stem without morphological suffixes or derivations. However, a stemmer may not necessarily produce any meaning-bearing word form, whereas a lemmatiser returns the base form, usually associated with a dictionary citation of the given word form.

Table 5.10 gives examples of stemming and lemmatisation results illustrating the differences between the two processes.

#### Spanish Adjective Base Forms

Spanish adjectives, in contrast to English, have gender and number inflections so that one base form can have four different full forms. For instance, the adjective '*bonito*' (beautiful/pretty) has four inflected forms ('*bonita*', '*bonitas*', '*bonito*', '*bonitos*').

Therefore, reducing the inflection from the Spanish adjectives might simplify the process of word alignment between two languages. All Spanish adjectives are replaced with its base forms whereas the English corpus remains the same.

#### Reduced Spanish Verbs

Spanish language has an especially rich inflectional morphology for verbs. Person and tense are expressed via suffix so that many different full forms of one verb exist, many of them without the corresponding equivalent in English. Therefore, reducing the POS information of Spanish verbs could be helpful for improving word alignments.

Each verb has been reduced into its base form and reduced POS tag: parts of POS tag describing tense and/or mode which does not exist in English are removed. For example, the

tag for the subjunctive mode has been removed, and the two tags representing two types of the past tense are replaced with the unique past tense tag.

#### Lemma plus reduced Spanish POS Morpho-attributes

As already mentioned, Spanish is a morphologically richer language than English. However, all inflected forms of Spanish are not relevant for translation into English. For instance, whereas Spanish adjectives may have four inflected forms, English adjectives have only one form. Therefore, it might be possible that all inflected forms of Spanish adjectives are not required for translation. Similar cases are possible to a limited extent with other words also, such as nouns, verbs, etc.

To handle this morphology-related problem of Spanish with respect to English, we can count for each Spanish part of speech (POS) tag which additional morphological attributes (morpho-attributes) do not affect the translation from Spanish to English. For this purpose, we extract bilingual lexicons from original word-based statistical word alignment for large training data from both directions (Spanish to English and English to Spanish), where each Spanish original word is replaced with its lemma plus morpho-syntactic tag. On this bilingual lexicons, entropy was calculated with respect to each morpho-attribute corresponding to each Spanish POS tag. As a result, Table 5.8 reports that irrelevant and relevant morpho-attributes corresponding to some Spanish POSs. Other Spanish POS (adverbs, conjunctions and interjections) have not been reported in the table as they do not convey enough morphological information. In case of some morpho-attributes for Spanish POS, the value of the entropy was not significantly reduced with respect to the value of the entropy considering only with lemma form. In this situation, we tried different combination of morpho-attributes for that POS. For instance, Table 5.8 reports relevant morpho-attributes for determiner are gender and number. We observed that for small data track, these morpho-attributes do not make significant effect on the translation. Therefore, in case of small data track, we have not provided this information with lemma form.

In general, Spanish words are replaced with lemma and its relevant POS tag information. The remaining ones are transformed into lemma forms in small as well as in large data (see Table 5.10 for example).

#### Full Verb Forms

Undoubtedly, given a verb meaning, tense and person, each language *implements* each verbal form independently from the other language. For example, whereas the personal pronoun is compulsory in English unless the subject is present, this does not occur in Spanish, where the morphology of the verb expresses the same aspect.

| POS-tag | Irrelevant POS morpho-attributes | Relevant POS morpho-attributes |
|---|---|---|
| Verb | type (principal, auxiliary) | mode, time, person, number, gender |
| Noun | type (common, proper, etc.), gender, grade | number (singular, plural, invariable) |
| Adjective | type, grade, gender, number, function | – |
| Pronoun | person, possessor, politeness | type, gender, number, case |
| Determiner | type (demonstrative, possessive, etc.) person, possessor | gender, number |
| Preposition | type, form, gender, number | – |

**Table 5.8:** *Irrelevant and Relevant Part-Of-Speech Morphological Attributes for Spanish.*

Therefore, aiming at simplifying the work for the word alignment, another word classification strategy can be devised to address the rich variety of verbal forms. For this, we group all words that build up a whole verbal form (including pronouns, auxiliary verbs and head verb) into the lemma of the head verb. This is a knowledge-based detection taken using deterministic automata implementing a few simple rules. These rules require information on word forms, POS-tags and lemmas in order to map the resulting expression to the lemma of the head verb, as done in [Gis05a]. Examples of such mappings can be found in Table 5.9.

| English | | Spanish | |
|---|---|---|---|
| full form → lemma | | full form → lemma | |
| has been found | find | introdujeran | introducir |
| we will find | find | han cometido | cometer |
| do you think | think | dijo | decir |
| offered | offer | está haciendo | hacer |
| I am doing | do | haremos | hacer |

**Table 5.9:** *Full verb forms are mapped to the lemma of the head.*

### 5.4.2 Word Order Modification

It is commonly known that non-monotonicity poses difficulties for word alignment, not to mention for statistical machine translation. The more differences in word order between two languages, the more difficult to extract a good alignment and the more challenging the translation task is. Although English and Spanish exhibit a quite remarkable monotonicity (compared to other pairs such as English and Chinese), here we study two techniques, exploring the possible gain in alignment quality of reordering one language to make word alignment more monotone.

#### POS-based Reordering of Spanish Nouns and Adjectives

Adjectives in Spanish are usually placed after the corresponding noun, whereas in English it is the other way round. Therefore local reordering of nouns and adjective groups might be helpful

for monotonising word alignments between two languages. POS-based local reordering has been used: each Spanish noun has been moved behind the correspondent adjective group. If there are two adjectives connected with a coordinate conjunction "and" or "or", the noun is moved behind the whole group of words.

#### Noun–Adjective swapped realignment

An alternative strategy consists of deciding which Spanish 'Noun + Adjective' structures need to be swapped according to classes extracted from an initial statistical word alignment in the original order, as introduced in [Cj06].

Given this baseline alignment, we build up classes of nouns preceding the same adjectives and having crossed links[2]. The same classes can be extracted for the adjectives following the same nouns. From these classes, we filter out those pairs occurring less than 6 times or having a low crossed-link probability, ie. being more often monotonically linked.

Finally, we swap all remaining 'Noun + Adjective' belonging to seen pairs of classes, and realign, as we expect the increase in monotonicity to reduce the word alignment complexity and improve quality.

| English | Asian countries have followed our example too . |
|---|---|
| base forms | Asian country have follow our example too . |
| stems | asian countri have follow our exampl too . |
| Spa Adj base | Asian countries have followed our example too . |
| Spa V reduced | Asian countries have followed our example too . |
| Spa lem+redPOS | Asian countries have followed our example too . |
| full verbs | Asian countries V[follow] our example too . |
| word order | Asian countries have followed our example too . |
| **Spanish** | **Los países asiáticos han seguido también nuestro ejemplo .** |
| base forms | El país asiático haber seguir también nuestro ejemplo . |
| stems | los país asiátic han segu también nuestr ejempl . |
| Spa Adj base | Los países asiático han seguido también nuestro ejemplo . |
| Spa V reduced | Los países asiáticos haber#P seguido también nuestro ejemplo . |
| Spa lem+redPOS | el país_NP asiático haber_VIP3P0 seguir_VP00SM (...) |
| full verbs | Los países asiáticos V[seguir] también nuestro ejemplo . |
| word order | Los asiáticos países han seguido también nuestro ejemplo . |

**Table 5.10:** *Some English and Spanish corpus transformations as described in corresponding sections.*

Obviously, one can combine two (or more) presented approaches to produce a new transformation. For example, any word order modification can be done together with stemming, base

[2]By crossed links, we mean that Spanish word in position $n$ is linked to English word in position $m+1$, and Spanish word in $n+1$ is linked to English word in $m$.

form substitution or full verb classification. Verb classification can also be combined with other transformation for all words outside the verb groups.

### 5.4.3 Experimental work

#### 5.4.3.1 Experiment setup

As already mentioned, experiments were carried out using the Spanish–English European Parliament parallel corpus version 3 (more details on it in Table 4.3). In order to extract the linguistic information needed to perform the presented corpus modifications, data was preprocessed as follows:

- English POS-tagging using freely-available *TnT* tagger [Bra00].
- English lemmatisation using *wnmorph*, included in the WordNet package [Mil91].
- Spanish POS-tagging and lemmatisation using *FreeLing* analysis tool [Car04].
- English and Spanish stemming using the Snowball stemmer[3], which is based on Porter's algorithm.

Table 5.11 shows the main statistics of this parallel corpus, including number of sentences, number of words, vocabulary and average sentence length for each language. The lower part of the table shows the statistics for the 1% division used in the small data track.

| | sent | words | vocab. | avg len |
|---|---|---|---|---|
| English | 1.28 M | 34.9 M | 106 k | 27.2 |
| Spanish | | 36.6 M | 153 k | 28.5 |
| English 1% | 13.4 k | 366 k | 16.3 k | 27.4 |
| Spanish 1% | | 385 k | 22.4 k | 28.8 |

**Table 5.11:** *Parallel corpus statistics for large and small data tracks.*

For evaluation, an ample set of bilingual sentences was aligned manually (see Table 5.12), by computing a consensus gold standard between three human alignments, as described in [Lam05]. Out of the set of gold standard links, 67% are Sure and 33% are Possible. This alignment test set is a subset of the training data, both in the large and the small data tracks.

#### 5.4.3.2 Alignment results

As word alignment core algorithm (baseline), GIZA++ [Och03a] was used. Two baseline configurations are compared.

[3] http://www.snowball.tartarus.org/

| | sent | words | vocab. | avg len |
|---|---|---|---|---|
| English | 400 | 11.7 k | 2.7 k | 29.1 |
| Spanish | | 12.3 k | 3.1 k | 30.4 |

**Table 5.12:** *Alignment test data statistics.*

On the one hand, model iterations were set to $1^5H^54^34^3$ (meaning 5 iterations of IBM model 1, 5 iterations of HMM model and 3 iterations of IBM models 3 and 4) without using word classes and respecting original case. On the other hand, we used the $1^4H^54^4$ configuration (meaning 4 iterations of IBM model 1, 5 iterations of HMM model and 4 iterations of IBM model 4), included 50 word classes per language as estimated by 'mkcls', a freely-available tool along with GIZA++[4], and worked with lowercase text before aligning.

As it will be seen in alignment results, the latter strategy (denoted simply as 'baseline') always produced significantly lower AER results that its true-case no-class counterpart (denoted as 'baseline*'), which is shown as a means of comparison. For this reason, this better configuration applies for all experiments that have been done, except the one noted as baseline*.

| | Eng→Spa | | | Spa→Eng | | | Union | | |
|---|---|---|---|---|---|---|---|---|---|
| | $R_S$ | $P_P$ | AER | $R_S$ | $P_P$ | AER | $R_S$ | $P_P$ | AER |
| baseline* | 59.97 | 75.05 | 33.09 | 59.11 | 78.16 | 32.31 | 69.33 | 67.65 | 31.56 |
| baseline | 63.10 | 77.11 | 30.34 | 64.12 | 80.21 | 28.38 | 73.37 | 69.43 | 28.77 |
| base forms | 66.37 | 83.50 | 25.75 | 68.06 | 83.72 | 24.69 | 73.93 | 75.01 | 25.51 |
| stems | 67.02 | 84.30 | 25.01 | 68.61 | 83.80 | 24.32 | 74.66 | 75.65 | 24.82 |
| Spa Adj base | 63.96 | 78.29 | 29.33 | 64.17 | 80.31 | 28.31 | 73.59 | 70.19 | 28.25 |
| Spa V reduced | 64.25 | 78.39 | 29.13 | 64.09 | 80.16 | 28.44 | 73.17 | 70.05 | 28.51 |
| Spa lem+redPOS | 64.36 | 80.63 | 28.06 | 64.51 | 79.08 | 28.70 | 73.71 | 70.76 | 27.87 |
| full verbs | 66.50 | 79.72 | 27.13 | 65.44 | 81.30 | 27.10 | 73.96 | 71.36 | 27.45 |
| Spa N-A reord | 63.44 | 77.27 | 30.08 | 64.57 | 80.39 | 28.04 | 73.40 | 69.68 | 28.61 |
| N-A swap realign | 63.63 | 77.41 | 29.91 | 64.27 | 80.00 | 28.38 | 73.43 | 69.59 | 28.65 |
| verbs + stems | 69.58 | 83.17 | 23.89 | 67.33 | 83.96 | 24.85 | 75.47 | 75.17 | 24.69 |

**Table 5.13:** *Word Alignment results for small-data task.*

Results with the 1% data set are shown in Table 5.13, where both directions and the symmetrisation through union are evaluated. Each row refers to each of the corpus transformations presented.

As it can be seen, both **base forms** and **stems** produce a very significant quality improvement, especially reflected in a more than 5 point absolute precision improvement in union alignment, whereas recall is also very high in these two cases for all alignment directions. It looks like their classifications reduce sparseness and help the word alignment algorithm perform better. This improvement is best in the case of stems.

[4]See http://www.fjoch.com for details on both tools.

Whereas '**Spa lem+redPOS**' transformation also achieves significant improvements in recall and precision for all directions, leading to an approximate 1 point AER reduction, improvements due to '**Spa Adj base**' and '**Spa V reduced**' transformations are very slight. Yet all three cases fall short compared to stemming or lemmatising, indicating that for data-sparse situations, classifying all words regardless of their class is a more effective strategy.

'**Full verb**' classification achieves a 1.5 AER reduction, basically thanks to an important recall increase in all alignment directions, due to the grouping effect of this classification, so that all words belonging to a verb form become linked to the same tokens. Finally, **reordering** experiments produce very slight improvements, and apparently the result is equal no matter if the reordering is *a priori* forced as in '**Spa N-A reord**' or learnt from data as in '**N-A swap realign**'.

Combining full verb classification and stemming (of the words outside verb forms) we obtain the best AER results.

| | Eng→Spa | | | Spa→Eng | | | Union | | |
|---|---|---|---|---|---|---|---|---|---|
| | $R_S$ | $P_P$ | AER | $R_S$ | $P_P$ | AER | $R_S$ | $P_P$ | AER |
| baseline* | 69.13 | 88.81 | 21.94 | 67.25 | 90.04 | 22.60 | 73.98 | 84.41 | 20.92 |
| baseline | 73.20 | 90.78 | 18.65 | 72.18 | 92.17 | 18.64 | 78.42 | 86.43 | 17.56 |
| base forms | 72.80 | 91.70 | 18.54 | 71.84 | 93.17 | 18.50 | 76.73 | 87.90 | 17.82 |
| stems | 73.56 | 92.40 | 17.79 | 72.72 | 93.78 | 17.68 | 77.81 | 88.94 | 16.74 |
| Spa Adj base | 73.01 | 90.78 | 18.77 | 72.40 | 92.47 | 18.39 | 78.30 | 86.70 | 17.50 |
| Spa V reduced | 73.07 | 90.69 | 18.77 | 72.07 | 92.22 | 18.70 | 77.97 | 86.43 | 17.80 |
| Spa lem+redPOS | 72.72 | 90.46 | 19.06 | 71.94 | 92.06 | 18.82 | 77.87 | 86.16 | 17.97 |
| full verbs | 74.27 | 90.77 | 17.85 | 73.03 | 93.31 | 17.56 | 78.60 | 87.37 | 16.97 |
| Spa N-A reord | 72.69 | 90.06 | 19.25 | 72.23 | 91.85 | 18.73 | 78.10 | 85.93 | 17.97 |
| N-A swap realign | 72.52 | 90.41 | 19.22 | 72.13 | 91.80 | 18.81 | 77.91 | 86.10 | 17.99 |
| verbs + stems | 74.74 | 91.83 | 17.14 | 73.23 | 93.84 | 17.23 | 78.36 | 88.82 | 16.42 |

**Table 5.14:** *Word Alignment results for large-data task.*

Results with the full parallel corpus are shown in Table 5.14. Interestingly, conclusions regarding base forms and stems do not hold in this case. Whereas base forms are not useful anymore and even degrade alignment quality, stems still provide significant improvement in AER. This is expressed in a 2.5 point absolute precision increase at a cost of 0.6 recall decrease. One possible reason for this is the harder classification of stems, especially for English, where initial vocabulary of 95K words is reduced to 81K with base forms and only 69K for stems (in Spanish, from baseline 138K vocabulary we end up with 78K base forms and 79K stems). Apparently, this involves a sparseness reduction, which makes word alignment more robust to non-literal translations. On the other hand, frequent words such as auxiliary verbs are not mapped to the same stem, thus possibly helping the aligner to discriminate compared to the case with base

forms.

Partial transformations such as '**Spa lem+redPOS**', '**Spa Adj base**' and '**Spa V reduced**' do not help improve alignment quality anymore. On the other hand, '**full verb**' classification is still producing significant improvements, again reflected in the best recall figures for all alignment directions. This recall can countermeasure the recall loss when stemming and achieves the best AER (16.42) when combining these two approaches.

As about word order modification experiments, again results are not encouraging, and in this case they are even harmful for alignment quality. This holds both for deterministic Noun–Adjective reordering ('**Spa N-A reord**') and for reordering according to an initial word alignment. All combinations of order modification and stemming, base form or verb forms classification that have been tested did not yield improvements and are not reported.

#### 5.4.3.3 Discussion

Remarkably, and even though quality improvements due to morphological information are bigger in case of data scarceness, alignment error rate can be reduced by using these informations even in case large amounts of data are available. Specifically, stemming and verb forms classification achieve significantly better recall and precision figures in all situations.

These experiments provide different alignment sets which can contain complementary information, so alignment quality can be further improved if they are combined. For the large data task, the best 3, 4 and 5 best union sets were combined with a consensus criterion. For each link present in at least one of the sets, if this link is present in a majority of sets, then it is selected for the combined set. Otherwise it is absent from the combined set. For the combination of an even number of sets, the criterion can be strict (more than half of the sets must agree) or weak (a half is enough).

| | $R_S$ | $P_P$ | AER |
|---|---|---|---|
| 3 best | 78.50 | 90.04 | 15.79 |
| 4 best (weak) | 80.29 | 87.35 | 16.10 |
| 4 best (strict) | 76.51 | 92.59 | 15.87 |
| 5 best | 78.37 | 89.70 | 16.07 |

**Table 5.15:** *Combination, with a consensus criterion, of the best union alignment sets obtained in the large data task (in order: the verbs+stems, stems, full verbs, spa adj base and baseline sets).*

Results are shown in Table 5.15. While all combinations improve the best AER presented in Table 5.14 (that of the verbs+stems experiment), the combination of best 3 sets is particularly interesting since both recall and precision are also improved. In the 4 sets combinations, the weak criterion gives a high recall and lower precision combination, whereas the strict criterion

gives a high precision but lower recall combination.

### 5.4.4 Correlation with SMT quality

Since word alignment represents the first step in the training of any SMT system, it is reasonable to expect that a better word alignment (as expressed by a lower AER figure) should generate more accurate translation units (tuples), which would in turn make for a better estimated $n$-gram translation model.

However, given the wide range of additional aspects affecting final translation performance, such as segmentation decisions, additional feature models, optimisation runs, etc. it is *a priori* unclear whether a certain alignment with lower AER will end up boosting BLEU or other automatic translation scores.

In addition to that, we would like to know how much gain in AER is needed to achieve significant gains in translation scores. Aiming at this correlation study, translation experiments were carried out by comparing 5 selected alignment configurations, ranging from worst to best AER for both large and small data tracks.

#### Small data track

Results for both translation directions in the small data track are shown in Table 5.16, where the result when translating only with the Bilingual Model (onlyBM) and the full log-linear combination (full) are shown.

| | AER | Eng→Spa | | | | Spa→Eng | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | onlyBM | | full | | onlyBM | | full | |
| | | BLEU | NIST | BLEU | NIST | BLEU | NIST | BLEU | NIST |
| baseline* | 31.56 | 0.2815 | 7.450 | 0.3198 | 7.922 | 0.3685 | 8.670 | 0.4159 | 9.163 |
| baseline | 28.77 | 0.2824 | 7.453 | 0.3215 | 7.943 | 0.3733 | 8.734 | 0.4209 | 9.204 |
| full verbs | 27.45 | 0.2884 | 7.507 | 0.3251 | 8.015 | 0.3651 | 8.603 | 0.4218 | 9.249 |
| stems | 24.82 | 0.2859 | 7.483 | 0.3254 | 8.031 | 0.3719 | 8.717 | 0.4283 | 9.319 |
| verbs + stems | 24.69 | 0.2897 | 7.491 | 0.3290 | 8.048 | 0.3597 | 8.567 | 0.4190 | 9.229 |

**Table 5.16:** *Translation scores for small-data task.*

At first glance, we can already conclude that strong variations in AER do not end up producing a strong variation in translation quality. While AER shows a nearly 22% relative decrease from worst to best alignment (about 6.87 points absolute), BLEU experiences an increase of at most 3% relative (Eng→Spa) and 4% relative (Spa→Eng). In Eng→Spa, biggest BLEU difference is 0.009 absolute (full system) whereas in Spa→Eng, it is about 0.014 absolute.

According to a 95% confidence level for this task, BLEU measures may have a variation of

around ±0.015 in Eng→Spa and around ±0.012 in Spa→Eng. Therefore, the impact of improved AER figures in the Eng→Spa translation direction is below the 95% BLEU confidence level, even though there seems to be a tendency to positively correlate with AER in this task.

Opposite to that, in the Spa→Eng direction, BLEU variations do achieve the confidence level, though minimally. Unfortunately, in this case correlation with AER is unclear. Particularly, stems and baseline achieve pretty similar results in the onlyBM configuration even though they present a nearly 4-point absolute AER difference. In addition to that, the use of full verb forms for alignment, which always generates a lower-AER alignment solution, does not help and even harm translation performance when comparing 'stems' versus 'verbs+stems'.

### Large data track

Results for the large data track are shown in Table 5.17. Again, the relatively big AER difference between both baselines (from 20.92 to 17.56, a 16% relative decrease) does not yield any significant change in translation performance in any translation direction, as shown in the first two rows.

| | AER | Eng→Spa | | | | Spa→Eng | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | onlyBM | | full | | onlyBM | | full | |
| | | BLEU | NIST | BLEU | NIST | BLEU | NIST | BLEU | NIST |
| baseline* | 20.92 | 0.4368 | 9.345 | 0.4794 | 9.887 | 0.4782 | 9.963 | 0.5519 | 10.797 |
| baseline | 17.56 | 0.4366 | 9.331 | 0.4802 | 9.909 | 0.4769 | 9.919 | 0.5526 | 10.763 |
| full verbs | 16.97 | 0.4293 | 9.219 | 0.4790 | 9.922 | 0.4728 | 9.871 | 0.5514 | 10.779 |
| stems | 16.74 | 0.4284 | 9.223 | 0.4787 | 9.883 | 0.4760 | 9.902 | 0.5553 | 10.788 |
| verbs + stems | 16.42 | 0.4264 | 9.187 | 0.4785 | 9.889 | 0.4748 | 9.882 | 0.5525 | 10.765 |

**Table 5.17:** *Translation scores for large-data task.*

Surprisingly, further improvements in AER achieved by 'full verbs', 'stems' and 'verbs+stems' configurations do not improve performance, but even produce worse translation results, as can be especially observed when only the bilingual translation model is used. In the case of full log-linear combination, BLEU differences are reduced to 0.4% relative (Eng→Spa) and 0.7% relative (Spa→Eng), or 0.0017 and 0.0039 absolute, far away from significance thresholds.

This unexpected behaviour of the bilingual model demands careful study. Clearly the translation $N$-gram model is not being able to leverage the improvements in alignment quality to build up a more robust system. Alternatively, one could also hypothesise that AER may be failing to measure improvements in word alignment quality (implying that differences in AER are not relevant in terms of alignment), though given the amount of previous work on the area we are inclined not to believe so[5].

[5]Note that recent works claiming low correlation between AER and SMT performance, as mentioned in §2.5.1,

In order to investigate this situation, a comparison of number of extracted tuples, tuple vocabulary and percentage of tuples translating to NULL (in the training parallel corpus) was carried out. Statistics are shown in Table 5.18, where we observe a clear tendency to extract more tuples as AER goes down (see columns named '# tups' for number of running tuples).

| | | Eng→Spa | | | Spa→Eng | | |
|---|---|---|---|---|---|---|---|
| | AER | # tups | toNULL | vocab | # tups | toNULL | vocab |
| baseline* | 20.92 | 19.91 M | 9.2% | 2.18 M | 20.55 M | 12.0% | 2.20 M |
| baseline | 17.56 | 19.89 M | 8.1% | 2.24 M | 20.66 M | 11.4% | 2.26 M |
| full verbs | 16.97 | 20.90 M | 9.5% | 2.32 M | 21.63 M | 12.5% | 2.34 M |
| stems | 16.74 | 21.89 M | 9.9% | 2.18 M | 22.62 M | 12.8% | 2.23 M |
| verbs + stems | 16.42 | 21.60 M | 9.9% | 2.32 M | 22.46 M | 13.3% | 2.34 M |

**Table 5.18:** *Training tuple statistics for large-data task.*

In many occasions, previously extracted long tuples are now broken into a sequence of tuples, as illustrated in the example from Tables 5.19 and 5.20 (comparing 'baseline*' and 'stems'). The model probabilities are therefore modified by these new shorter units (which include an important percentage of tuples to NULL).

| | a | building | contractor | in | another | Member | State | who | were | drawing | benefit |
|---|---|---|---|---|---|---|---|---|---|---|---|
| subsidio | . | . | . | . | . | . | . | . | . | . | + |
| un | . | . | . | . | . | . | . | . | . | . | . |
| percibían | . | . | + | . | . | . | . | . | . | + | . |
| que | . | . | . | . | . | . | . | + | . | . | . |
| , | . | . | . | . | . | . | . | . | . | . | . |
| miembro | . | . | . | . | . | + | + | . | . | . | . |
| Estado | . | . | . | . | . | + | + | . | . | . | . |
| otro | . | . | . | . | + | . | . | . | . | . | . |
| en | . | . | . | + | . | . | . | . | . | . | . |
| obras | . | + | + | . | . | . | . | . | . | . | . |
| de | . | . | . | . | . | . | . | . | . | . | . |
| aparejador | . | . | + | . | . | . | . | . | . | . | . |
| un | + | . | . | . | . | . | . | . | . | . | . |

**Extracted tuples:**

a # un

bldg. contr. in another Mem. State who were drawing # un apjdr. de obras en otro Est. miem. , q. percibían

benefit # un subsidio

**Table 5.19:** *Eng→Spa example of a long tuple extracted with 'baseline*' configuration.*

As a result, tuple vocabulary augments as well, though not in the 'stems' configuration, and markedly in the case of using verb classification. This tuple vocabulary increase is derived from

do not find a negative correlation as in this experiment. Besides, they are presented for phrase-based SMT systems, an approach which depends differently from word alignment than *N*gram-based SMT. As an example, consider two different alignments with a common long-distance link from the first source word to the last target word. In this case, extracted tuples will be the same, whereas phrases will depend on the quality of short-distance links.

| | a | building | contractor | in | another | Member | State | who | were | drawing | benefit |
|---|---|---|---|---|---|---|---|---|---|---|---|
| subsidio | . | . | . | . | . | . | . | . | . | . | + |
| un | . | . | . | . | . | . | . | . | . | . | . |
| percibían | . | . | . | . | . | . | . | . | . | + | . |
| que | . | . | . | . | . | . | . | . | + | . | . |
| , | . | . | . | . | . | . | . | . | . | . | . |
| miembro | . | . | . | . | . | + | . | . | . | . | . |
| Estado | . | . | . | . | . | . | + | . | . | . | . |
| otro | . | . | . | . | + | . | . | . | . | . | . |
| en | . | . | . | + | . | . | . | . | . | . | . |
| obras | . | + | + | . | . | . | . | . | . | . | . |
| de | . | . | . | . | . | . | . | . | . | . | . |
| aparejador | . | . | + | . | . | . | . | . | . | . | . |
| un | + | . | . | . | . | . | . | . | . | . | . |

**Extracted tuples:**

a # un
building contractor # aparejador de obras
in # en
another # otro
Member State # Estado miembro ,
who # que
were # NULL
drawing # percibían
benefit # un subsidio

**Table 5.20:** *Eng→Spa example of a long tuple extracted with 'stems' configuration.*

imposing the full verb form segmentation, forcing the whole verb form to belong to a single tuple. Apparently, this situation may create sparseness and explain the slight performance decrease.

On the other hand, the percentage of tuples translating to NULL (shown in 'vocab' columns) does also increase as AER goes down, with the clear exception of the 'baseline' configuration. The impact of these novel NULL-tuples appearance can be measured by studying the outputs produced by the system when only the bilingual model is used.

Table 5.21 shows statistics from this translation, including (apart from AER and BLEU score) number of output words (trgwrds), number of tuples used (tups), percentage of these translating to NULL (%toNULL) and percentage of tuples used as 1gram, 2gram and 3gram in Eng→Spa translation. Note that English input number of words is 26,917, and both Spanish references contain 27,4k and 27,7k words respectively.

| | | | Eng→Spa onlyBM | | | |
|---|---|---|---|---|---|---|
| | AER | BLEU | trgwrds | tups | %toNULL | %1–2–3grams |
| baseline* | 20.92 | 0.4368 | 25,853 | 23,214 | 9.6 | 18.1–43.4–38.5 |
| baseline | 17.56 | 0.4366 | 25,899 | 22,793 | 8.3 | 18.6–44.2–37.2 |
| full verbs | 16.97 | 0.4293 | 25,440 | 22,771 | 11.4 | 18.7–43.7–37.6 |
| stems | 16.74 | 0.4284 | 25,519 | 23,205 | 11.0 | 17.0–44.2–38.8 |
| verbs + stems | 16.42 | 0.4264 | 25,313 | 22,881 | 11.9 | 18.2–43.8–38.0 |

**Table 5.21:** *Translation output study (onlyBM, Eng→Spa).*

As it can be seen, the percentage of used tuples translating to NULL grows remarkably in the three bottom configurations, which implies a negative reduction of output word production.

Apart from that, in those cases using verb forms ('full verbs' and 'verbs + stems'), the translation model tends to use a shorter memory length when compared to 'baseline*' (higher percentage of 1grams, lower percentage of 3grams), signalling a sparseness increase. On the contrary, although the 'stems' configuration achieves an important context usage (lowest percentage of 1grams, highest of 3grams), the excessive number of tuples to NULL seems to be limiting its translation performance.

When comparing both baselines we observe that 'baseline' produces a slightly longer output (around 40 extra words), even though it uses less tuples in translation. This is due to the use of longer tuples and a much lower percentage of tuples to NULL. However, this has the cost of losing context during translation (much lower percentage of 3grams usage), which ends up obtaining same BLEU scores.

Finally, Table 5.22 shows the same statistics in the opposite translation direction. In this case, Spanish input number of words is 22,774, and both English references contain 22,8k and 23,0k words respectively.

| | | | Spa→Eng onlyBM | | | |
|---|---|---|---|---|---|---|
| | AER | BLEU | trgwrds | tups | %toNULL | %1–2–3grams |
| baseline* | 20.92 | 0.4782 | 21,253 | 20,001 | 11.3 | 14.3–40.6–45.1 |
| baseline | 17.56 | 0.4769 | 21,272 | 19,671 | 10.7 | 14.7–41.3–44.0 |
| full verbs | 16.97 | 0.4728 | 20,998 | 19,566 | 12.5 | 14.9–41.2–43.9 |
| stems | 16.74 | 0.4760 | 21,000 | 19,994 | 12.5 | 13.5–41.1–45.5 |
| verbs + stems | 16.42 | 0.4748 | 20,963 | 19,657 | 13.0 | 14.5–40.9–44.6 |

**Table 5.22:** *Translation output study (onlyBM, Spa→Eng).*

As it can be seen, tendencies regarding an increasing percentage of tuples to NULL (leading to shorter translation output) and context usage correlate with the opposite direction.

To sum up, we can conclude that current *N*gram-based SMT system proves incapable of taking full advantage of improved alignments. The trade-off between model sparseness in training (equivalent to context usage during translation) and usage of tuples to NULL seems to keep stable, especially when log-linear model combination is conducted.

Further study to understand the implications of these tuples translating to NULL in translation modelling and generation need to be conducted in the future.

## 5.5 Chapter Summary and Conclusions

This chapter was devoted to a study of how linguistic information can help improve statistical word alignment quality. For evaluation, Alignment Error Rate was used, and details on this measure as well as on development of manual references were discussed.

Three alternative approaches were followed; firstly, extending a one-to-one coocurrence-based statistical word alignment model with many-to-many linguistic groups; secondly, using these many-to-many groups as a segmentation preprocessing for constraining GIZA-based word alignment; and thirdly, extending this approach with several linguistically-driven data classification aiming at data sparseness reduction for word alignment IBM models.

Overall results prove that the use of linguistic information does indeed reduce alignment error rates in all tasks. As it is reasonable to expect, the effect of linguistic classification techniques gets reduced as data availability grows.

Additionally, a study of the impact of AER reduction on final translation performance was included in §5.4.4. Unfortunately, the main conclusion is that our current $N$gram-based SMT system does not leverage these alignment changes to construct a significantly better translation model.

Some research work reported in this chapter was published in the following contributions:

- [Gis04b] A. de Gispert, J.B. Mariño and J.M. Crego, "Phrase-based alignment combining corpus cooccurrences and linguistic knowledge," in *Proceedings of the 1st International Workshop on Spoken Language Translation, IWSLT'04*, pps. 107–114, October 2004.
- [Gis05a] A. de Gispert, "Phrase Linguistic Classification and Generalisation for Improving Statistical Machine Translation," in *Proceedings of the ACL Student Research Workshop 2005*, pps.67–72, June 2005.
- [Gis06c] A. de Gispert and J.B. Mariño, "Linguistic knowledge in statistical phrase-based word alignment," in *Natural Language Engineering*, Vol. 12, num. 1, pps.91–108, March 2006.
- [Gis06a] A. de Gispert, D. Gupta, M. Popovic, P. Lambert, J. B. Mariño, M. Federico, H. Ney and R. Banchs, "Improving Statistical Word Alignments with Morpho-syntactic Transformations," in *Proceedings of 5th International Conference on Natural Language Processing, FinTAL'06*, Springer Verlag, LNCS, pps. 368–379, August 2006.

# Chapter 6

# Linguistic Knowledge into Translation Modelling

## 6.1 Introduction

This chapter is devoted to techniques incorporating linguistic information directly into the translation model. In particular, and in the face of the error analysis study conducted in Chapter 4, we mainly address verb form translation, disagreement improvement and conduct a thorough analysis of the effects of morphology in the English→Spanish task. The chapter is organised as follows:

- Given that an important number of translation errors are related to verb forms and taking into account that they are very important to translate the meaning of a sentence, we followed a classification strategy aiming at reducing sparsity and introduce generalisation capabilities. This approach is reported in §6.2.
- In addition to that, §6.3 briefly addresses the problem of morphology disagreement by introducing a complementary feature function into the log-linear combination.
- Finally, §6.4 concludes this chapter with a study of the possible translation gain produced by modelling morphology-reduced translation models. For each morphology word category, Spanish morphology derivation is reduced before SMT training and oracle results are compared to post-processing oracles.

## 6.2 Verb classification for SMT

### 6.2.1 Introduction

As observed from SMT translation outputs, there is big difficulty in translating verb forms correctly. Their mode, tense, person and number morphological derivations imply a large vocabulary size, thus generating sparseness when training translation models. Unless a sufficiently large data representing all forms is available (which rarely happens), all verb forms will not be represented, not to mention their usual contexts.

This is very relevant for Romance languages, as Spanish or Catalan. For English, some of these morphology variations are expressed via the appearance of new words (obligatory personal pronouns, modals such as 'will' or 'would'), whereas others are not even explicitly found in English words (such as Spanish subjunctive present tense).

In order to tackle this, we devised a classification strategy which is described next. The objective is to group all verb forms from a same verb to a single token for bilingual $N$-gram modelling, independently of the actual form representation (with or without auxiliaries, with or without pronouns, etc.).

This way, the bilingual context of a certain verb will be much more represented in the data, and hopefully better estimated. Once a verb form is detected in the input text, this will be classified and the bilingual model will work on the verb-classified language. Additionally, if the target text contains classified verbs, an additional model (instance model) will have to determine which target instance is to be selected given the source instance.

Details and experiments are reported next.

### 6.2.2 Verb classification model

Suppose we want to translate a source sentence $f$ to a target sentence $e$. By defining $\tilde{e}_i$ as a certain source phrase and $\tilde{f}_j$ as a target phrase (where phrases are just sequences of contiguous words), the phrase translation model $Pr(\tilde{e}_i|\tilde{f}_j)$ can be decomposed as:

$$\sum_T Pr(\tilde{e}_i, T|\tilde{f}_j) =$$

$$= \sum_T Pr(\tilde{e}_i|T, \tilde{f}_j) Pr(\tilde{E}_i, \tilde{F}_j|\tilde{f}_j) =$$

$$= \sum_T Pr(\tilde{e}_i|T, \tilde{f}_j) Pr(\tilde{E}_i|\tilde{F}_j, \tilde{f}_j) Pr(\tilde{F}_j|\tilde{f}_j) \quad (6.1)$$

where $T = (\tilde{E}_i, \tilde{F}_j)$ is the pair of source and target classes used (called Tuple), and $\tilde{E}_i$, $\tilde{F}_j$ are the generalised classes of the source and target phrases, respectively. In our current implementation, we consider a classification of phrases that is:

- *Linguistic*, ie. based on linguistic knowledge
- *Unambiguous*, ie. given a source phrase there is only one class (if any)
- *Incomplete*, ie. not all phrases are classified, but only the ones we are interested in
- *Monolingual*, ie. it runs for every language independently

The second condition implies $Pr(\tilde{F}_j|\tilde{f}_j) = 1$, leading to the following approximation:

$$Pr(\tilde{e}_i|\tilde{f}_j) \approx \max_T Pr(\tilde{E}_i|\tilde{F}_j) Pr(\tilde{e}_i|T, \tilde{f}_j) \tag{6.2}$$

where we have just two terms, namely a standard phrase translation model based on the classified parallel data, and an instance model assigning a probability to each target instance given the source class and the source instance. The latter helps us choose among target words in combination with the language model.

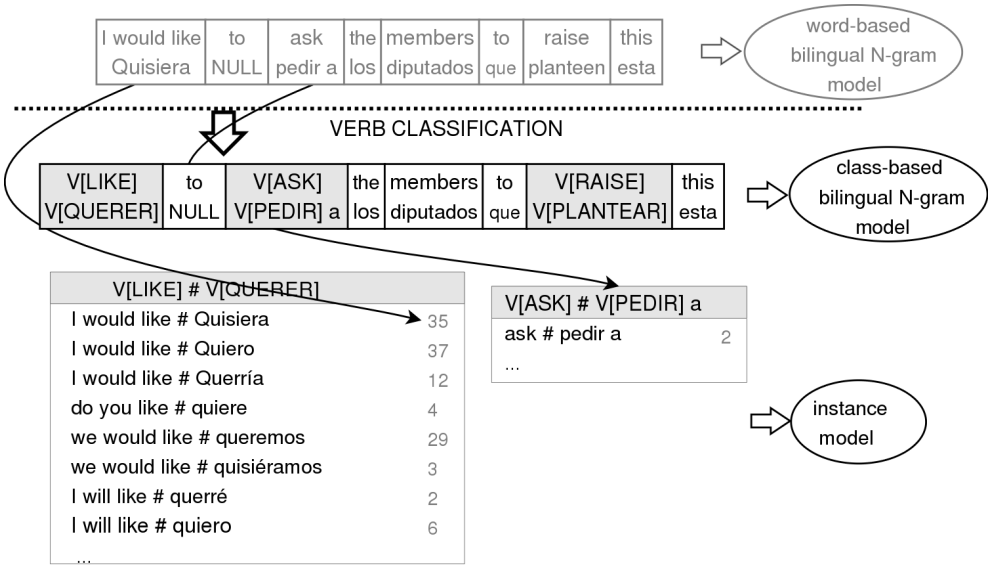



**Figure 6.1:** Example of verb classification for bilingual translation model and instance model estimation via study of observed instances for each tuple containing a verb class.

Therefore, once a standard translation model over the classified text (without verb forms, but only verb heads) is estimated, counts of observed instances in each classified tuple serve as data for the instance model. Figure 6.1 exemplifies this classification approach.

#### 6.2.2.1 Instance model

In order to estimate this instance model $Pr(\tilde{e}_i|T, \tilde{f}_j)$, we propose a simple approach based on the relative frequency of each instance across all tuples that share the same source phrase, as expressed by equation 6.3.

$$Pr(\tilde{e}_i|T, \tilde{f}_j) = \frac{N(T, \tilde{e}_i, \tilde{f}_j)}{N(T, \tilde{f}_j)} \tag{6.3}$$

thus weighing each target verb form given the source form, and the translation tuple or phrase containing the source and target classes.

#### 6.2.2.2 Generalisation of unseen verb forms

Usually, a number of verb forms appearing in the test set will be unseen in the training data. In these cases, they will be classified to the lemma of their head verb and, if this has been seen, will be translated into a target phrase. However, the instance model probability given this source verb form is not defined, and a generalisation strategy must be followed.

To produce a target instance $\tilde{e}_i$ given the tuple $T$ and an unseen source instance $\tilde{f}_j$, the approach followed has been to make use of the information of verb forms that are seen in the training, seeking among seen instances those that are identical except on the personal pronoun (or verb suffix).

| | | |
|---|---|---|
| $T_1$ = (V[pay] , V[pagar]) | | |
| I would have payed | habría pagado | 3 |
| you would have payed | habrías pagado | 1 |
| you would have payed | pagarías | 1 |
| $T_2$ = (V[pay] , V[hacer] el pago) | | |
| * would have payed | — | 0 |
| $T_3$ = (V[pay] it , lo V[pagar]) | | |
| I would have payed it | lo habría pagado | 1 |

**Table 6.1:** *Seen instances in the tuples translating V[pay] that are useful to generalise 'we would have payed'.*

For example, suppose we want to translate the sentence '*we would have payed* it' from English to Spanish and we see tuples $T_1$=(V[pay],V[pagar]), $T_2$=T(V[pay],V[hacer] el pago) and $T_3$=T(V[pay] it, lo V[pagar]) translating the class V[pay] in the training data. However, among all seen instances of these three tuples, the verb form 'we would have payed' is not to be found. In this case, for each tuple we look among its seen instances for identical instances (in words,

POS-tags and lemmas) except for the information regarding the person, as shown in Table 6.1, where no useful instance has been found for $T_2$.

For each of these instances, we generate a new Spanish verb form, by changing all the information on the person in the seen form (*habría pagado*, 1stSingular) for the detected person of the expression to translate (*we*, 1stPlural). Furthermore, each new translation alternative is weighed according to the number of times the seen instance has appeared in training, shown in the last column of Table 6.1. This weight acts as the instance probability for these new forms. In the example, the following new forms would be generated, with probability:

| | | | |
|---|---|---|---|
| $T_1$ | we would have payed | habríamos pagado | 4/6 |
| $T_1$ | we would have payed | pagaríamos | 1/6 |
| $T_3$ | we would have payed it | lo habríamos pagado | 1/6 |

Note that in the case of ambiguity (for example when generalising a form with 'you', it can be translated into 2nd person singular or plural in Spanish), our approach is to over-generate all possible forms and let the SMT combination of models choose the most convenient one. Actually, we expect the target Language Model to help decide the best translation alternative.

#### 6.2.2.3 Extended generalisation

In many cases, we observe only one exact realisation of the test verb form in the training set. If this instance is found in a highly-improbable tuple $T_i$, the translation system will be forced to produce this translation, ignoring the fact that there may be several other tuples $T_k$ translating the class with much higher probability.

Then, another approach to generalisation is to look for generalisation instances in all tuples, no matter whether there already is one exact seen instance of the test verb form in one tuple $T_i$. We will call this approach Extended Generalisation. A comparison of translation results for these alternative approaches is performed in the next section.

### 6.2.3 Advantages and difficulties

This classification strategy has three advantages:

- **Better alignment**. By reducing the number of words to be considered during first word alignment (auxiliary words in the classes disappear and no inflected forms used), we lessen the data sparseness problem and can obtain a better word alignment (as was done in §5.3). In a secondary step, one can learn word alignment relationships inside aligned classes by realigning them as a separate corpus, if that is desired.

- **Improvement of translation probabilities**. By considering many different phrases as different instances of a single phrase class, we reduce the size of our phrase-based (now class-based) translation model and increase the number of occurrences of each unit, producing a model $Pr(\tilde{E}|\tilde{F})$ with less perplexity.

- **Generalising power**. Phrases not occurring in the training data can still be classified into a class, and therefore be assigned a probability in the translation model. The new difficulty that rises is how to produce the target phrase from the target class and the source phrase, if this was not seen in training.

On the other hand, these are the main difficulties that need to be faced which will hopefully lead to improved translation performance if tackled conveniently.:

- **Verb detection and classification**. To solve this, we use an in-house rule-based deterministic classification automaton (already introduced in §5.2.3.1). This is done both in the English and the Spanish side, and before word alignment. Note that we detect verbs containing adverbs and negations, which are ordered before the verb to improve word alignment with Spanish, but once aligned they are reordered back to their original position *inside* the detected verb, representing the real instance of this verb.

  Even though no formal quantitative evaluation of its accuracy was conducted, in all informal subjective examinations, no errors were detected. This task is assumed to be successfully completed.

- **Instance probability**. On the one hand, when a phrase of the test sentence is classified to a class, and then translated, how do we produce the instance of the target class given the tuple T and the source instance? This problem is mathematically expressed by the need to model the term of the $Pr(\tilde{e}_i|T, \tilde{f}_j)$ in Equation 6.2.

  At the moment, we learn this model from relative frequency across all tuples that share the same source phrase, dividing the times we see the pair $(\tilde{f}_j, \tilde{e}_i)$ in the training by the times we see $\tilde{f}_j$. However, this solution is not dependent on the bilingual context anymore. In other words, only the source instance serves to weight the target instance (independently of the bilingual model), so that we rely only on the target language model to cope with the contextual factor.

- **Unseen instances**. To produce a target instance $\tilde{f}$ given the tuple $T$ and an unseen $\tilde{e}$, our idea is to combine both the information of verb forms seen in training *and* off-the-shelf knowledge for generation. A translation memory can be built with all the seen pairs of instances with their inflectional affixes separated from base forms.

For example, suppose we translate from English to Spanish and see the tuple T=(V[go],V[ir]) in training, with the following instances:

| | |
|---|---|
| I will go<br>PRP(1S) will VB | iré<br>VB 1S F |
| you will go<br>PRP(2S) will VB | irás<br>VB 2S F |
| you will go<br>PRP(2S) will VB | vas<br>VB 2S P |

where the second row is the analysed form in terms of person (1S: 1st singular, 2S: 2nd singular and so on) and tense (VB: infinitive and P: present, F: future). From these we can build a generalised rule independent of the person ' PRP(X) will VB ' that would enable us to translate 'we will go' to two different alternatives (present and future form):

| | |
|---|---|
| we will go | VB 1P F |
| we will go | VB 1P P |

These alternatives can be weighted according to the times we have seen each case in training. An unambiguous form generator produces the forms 'iremos' and 'vamos' for the two Spanish translations.

### 6.2.4 LC-Star experiment

To evaluate the proposed verb classification approach, a first experiment was carried out using the parallel corpus developed in the framework of the LC-Star project (already introduced in §3.2.3). In this case, English→Spanish translation direction was studied, as it contains more verb form translation errors according to error analyses. Data preprocessing included:

- Normalisation of contracted forms for English (ie. wouldn't = would not, we've = we have)
- English POS-tagging using freely-available *TnT* tagger [Bra00], and lemmatisation using *wnmorph*, included in the WordNet package [Mil91].
- Spanish POS-tagging using *FreeLing* analysis tool [Car04]. This software also generates a lemma or base form for each input word.

Table 6.2 shows the statistics of the LC-Star corpus used, where each column shows number of sentences, number of words, vocabulary, and average sentence length, respectively.

There are 20 unseen words in the English development set (0.3% of all words), and 48 unseen words in the English test set (0.7% of all words). Three Spanish reference translations are available for both the development and the test set.

| | sent | words | vocab | avglen |
|---|---|---|---|---|
| Train set | | | | |
| English | 29998 | 419113 | 5940 | 14.0 |
| Spanish | | 388788 | 9791 | 13.0 |
| Dev set | | | | |
| English | 350 | 6645 | 841 | 19.0 |
| Test set | | | | |
| English | 500 | 7412 | 963 | 14.8 |

**Table 6.2:** *LC-Star English–Spanish Parallel corpus statistics.*

#### 6.2.4.1 Verb Phrase Detection/Classification

Table 6.3 shows the number of detected verbs using the rule-based detection automaton, and the number of different lemmas they are mapped to. For the development and test sets, the percentage of unseen verb forms and lemmas are also shown.

| | verbs | unseen | lemmas | unseen |
|---|---|---|---|---|
| Train set | | | | |
| English | 56419 | | 768 | |
| Spanish | 54460 | | 911 | |
| Dev set | | | | |
| English | 856 | 3% | 120 | 0% |
| Test set | | | | |
| English | 1076 | 5.2% | 146 | 4.7% |

**Table 6.3:** *Detected verb forms in corpus.*

In average, detected English verbs contain 1.81 words, whereas Spanish verbs contain 1.08 words. This is explained by the fact that we are including the personal pronouns in English and modals for future, conditionals and other verb tenses, whereas Spanish tends to omit personal pronouns and contract tense information in a single inflected form.

#### 6.2.4.2 Translation results

In order to evaluate the proposed classification scheme, we integrated it into an $N$gram-based SMT system implementing a log-linear combination of:

- class-based bilingual translation model
- instance model
- target language model

- word bonus

Four translation experiments were conducted, whose results are shown in Table 6.4. On the one hand, a baseline experiment without verb forms classification (**baseline**). Secondly, an experiment with the classification but without dealing with unseen verb forms, which are not translated (**verb class**).

Later on, the same experiment including the generalisation of unseen verb forms described in Section 6.2.2.2 (**verb class + gen**). Finally, a last experiment also generalising regardless of the form appearing in the training data, as discussed in Section 6.2.2.3 (**verb class + genEX**) and shown in the last row of the table. For all four experiments, the weights of each model have been optimised according to the BLEU score in the development set.

| | dev set | | test set | |
|---|---|---|---|---|
| | WER | BLEU | WER | BLEU |
| baseline | 21.32 | 0.698 | 23.16 | 0.671 |
| verb class | 19.37 | 0.728 | 22.22 | 0.686 |
| verb class + gen | 19.27 | 0.727 | 21.65 | 0.692 |
| verb class + genEX | 19.25 | 0.729 | 21.62 | 0.689 |

**Table 6.4:** *Verb classification translation results. LC-Star corpus, English→Spanish.*

#### 6.2.4.3 Discussion

As it can be seen, the classification produces a significant improvement both in WER and BLEU, even when not dealing with unseen verb forms (around 60 verb forms in the test set). When generalising unseen forms we achieve a further boost in performance. Note that this could hardly be achieved by a strictly statistical model, since the form to be translated is not present in the training data. Finally, even though the idea of generalising tuples when the verb form is seen too does not harm the performance, it does not seem to provide any significant improvement either, leading to a practically identical output.

The different behaviour between development and test sets can be explained in terms of the percentage of verb forms that are unseen (which is higher in the test, as shown in Table 6.3), leading to a bigger improvement when performing generalisation.

On the other hand, in the test set we have 4.7% of the lemmas which are unseen and therefore cannot be translated at all unless a dictionary is provided. This effect is not present in the development set, which indicates that there is room for improvement in the final results.

Examples of translated sentences can be found in Figure 6.2.

The Research work reported in this chapter has been published in the following contributions:

| | |
|---|---|
| **source:** | I WAS TOLD that the service IS very good |
| **baseline:** | yo estaba dicho que el servicio está muy bien |
| **verb class:** | me habían dicho que el servicio está muy bien |
| **source:** | In two days' time , if YOU HAVE NOT CALLED me I WILL CANCEL the reservation |
| **baseline:** | pasado mañana fuera tiempo , si no hemos llamado anular la reserva |
| **verb class:** | en dos días tiempo , si UNSEEN UNSEEN la reserva |
| **verb class+gen:** | en dos días tiempo , si no ha llamado la anularé la reserva |

**Figure 6.2:** Examples of translated sentences. English detected verb forms are shown in capital letters.


- [Gis05a] A. de Gispert, "Phrase Linguistic Classification and Generalisation for Improving Statistical Machine Translation," in *Proceedings of the ACL Student Research Workshop 2005*, pags.67–72, June 2005.
- [Gis05c] A. de Gispert, J.B. Mariño and J.M. Crego, "Improving Statistical Machine Translation by Classifying and Generalising Inflected Verb Forms," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pags.107–114, September 2005.
- [Gis05b] A. de Gispert, J.B. Mariño and J.M. Crego, "Clasificación y generalización de formas verbales en sistemas de traducción estocástica," in *XXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural, SEPLN 2005*, pags.69–76, September 2005.


### 6.2.5 European Parliament experiment

Extending this experimental work, a complementary study using the European Parliament corpus (version 3) was carried out. The objective is to evaluate this same classification technique on a much larger corpus (see statistics in Table 4.3) with in a different domain and style. Again English→Spanish direction will be studied.

Apart from that, given their importance for state-of-the-art translation performance, we want to include lexicon models (based on IBM model 1 probabilities over word classes) into the log-linear combination, as for simplification this was excluded from the previous experiment.

| | dev set | | test set | |
|---|---|---|---|---|
| | WER | BLEU | WER | BLEU |
| baseline | 39.12 | 0.5552 | 40.22 | 0.4789 |
| verb class | 39.43 | 0.5504 | 40.88 | 0.4696 |
| verb class + gen | 39.51 | 0.5505 | 40.77 | 0.4707 |
| verb class + genEX | 39.43 | 0.5509 | 41.11 | 0.4696 |

**Table 6.5:** *Verb classification translation results. EuParl v3 corpus, English→Spanish.*

Results for this task are shown in Table 6.5. Contrary to the LC-Star results from the previous section, here the verb classification strategy does not seem to help improve translation results, yielding slightly worse scores. Reasons explaining this undesired situation may include:

- Corpus size. The much larger training data set may be reducing the data sparseness problem regarding verb forms which the classification strategy attempts to address.
- IBM model 1 influence. When verb classes are used, including this model in the log-linear combination is not trivial, as we can either consider it with respect to the classes (ie. computing IBM model 1 probabilities for the classified text) or with respect to the word sequences inside each class, which has implications in the information each bilingual unit must carry. In this experiment, given the augmented complexity of the latter approach, the first solution was selected.
- Bad influence of verb segmentation on bilingual model.

In order to assess the validity of these explanations, we carried out the equivalent experiment without lexicon models (equivalently to LC-Star experiment), and results are showed in Table 6.6.

| | dev set | | test set | |
|---|---|---|---|---|
| | WER | BLEU | WER | BLEU |
| baseline noLEX | 42.06 | 0.5223 | 44.34 | 0.4422 |
| verb class noLEX | 42.14 | 0.5214 | 44.80 | 0.4376 |

**Table 6.6:** *Verb classification translation results excluding lexicon models in log-linear combination. EuParl v3 corpus, English→Spanish.*

As it can be seen, the classification approach is already obtaining worse scores without the introduction of lexicon models (ie. in the equivalent experiment to LC-Star results). This suggests that, in this case, the influence of IBM model 1 features is not explaining the bad performance. Clearly, the verb-classified bilingual model is not producing better estimations of the translation process, possibly due to the word grouping effect of the classification (generating different history length $N$ depending on word identity).

This result also suggests that classifying **all** verb forms, independently of their mode or tense, into a single token may be producing a model with far too low resolution. Indeed, assuming that infinitives, past participles (which act as adjectives in many occasions) or a future or present tense should appear in the same bilingual contexts may be too strong an assumption.

Complementarily, we performed a final evaluation in terms of classified text, in order to ignore actual verb instances. The objective is to evaluate whether the verb instances are being wrongly generated, thus signalling a verb instance model estimation problem.

| | dev set | | test set | |
|---|---|---|---|---|
| | WER | BLEU | WER | BLEU |
| baseline | 37.75 | 0.5747 | 38.42 | 0.5055 |
| verb class | 37.61 | 0.5783 | 38.49 | 0.5013 |

**Table 6.7:** *Verb classification translation results. EuParl v3 corpus, English→Spanish.*

For this, we classified verb forms in both translation outputs (for the full system including lexicon models) and in human references, obtaining the scores from Table 6.7. These figures represent the scores that would be achieved if, for each translated verb also appearing in the references, the final verb form instance matches the reference.

As results are quite similar for both approaches (better for 'verb class' in development set, and better for 'baseline' in test set), we can conclude that the instance model is solving the instance with sufficient accuracy. As similar scores would be obtained either by manually matching baseline or class-based verb forms to the references, this confirms that the main problem lays in class-based bilingual model estimation, which is not taking advantage of the proposed word grouping scheme.

## 6.3 Target Part-Of-Speech Language Model

In order to tackle disagreement errors (detected in error analysis from §4.3.6), a new feature was added to the log-linear combination, corresponding to the 5-gram language model of the target POS-tag sequence. This feature does not require POS-tagging of the output sentence, as the POS information is carried within the tuple.

Accordingly, the bilingual unit is redefined in terms of a triplet comprising the source word sequence, the target word sequence, and the Part-Of-Speech sequence representing that target word sequence. For simplicity, we only allow one single POS representation for each target word sequence (*and* given a fixed source word sequence).

Note that the POS information contained in the triplet is not actually used for computing the bilingual translation model probabilities, thus keeping the $N$-gram model unchanged. This information is only used during decoding, when the target POS language model is computed for each hypothesis and included in the log-linear combination with its own weight.

The goal for this feature is to help the global system choose among different morphological variations when both the bilingual model and the word-based target language model lack context. An example of such a situation is shown in Figure 6.3, where the POS language model would favour the masculine alternative 'establecidos' due to its being preceded by the masculine Spanish noun 'objetivos.'

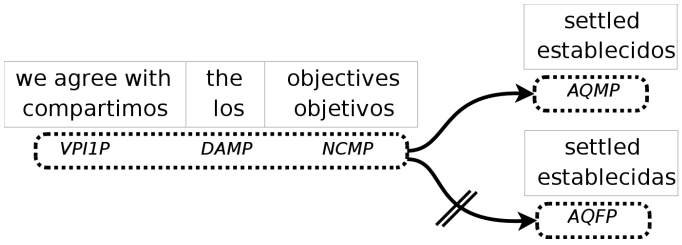



**Figure 6.3:** Augmented tuple and target POS-tag language model implementation.

Of course, this will not be useful whenever the correct alternative is missing (ie. it does not exist as a tuple), since no morphological generation is being performed.

Table 6.8 shows the incidence of the target POS-tag language model on translation quality for the experimental translation tasks, by adding to the full log-linear combination (full) the new proposed feature (+trgPOSlm) in EuParl version 2 task[1].

As we can see, a slight improvement is observed only in the translation direction into Spanish. This is a reasonable result, according to two facts; first, that English does not present such

[1]For historical reasons, IBM1weight tuple segmentation is used in this experiment. Therefore, 'full' results correspond to IBM1weight results from Table 4.6

| | | BLEU | mWER | NIST |
|---|---|---|---|---|
| Eng→Spa | full | 0.4714 | 40.22 | 9.83 |
| | +trgPOSlm | 0.4750 | 40.42 | 9.87 |
| Spa→Eng | full | 0.5470 | 34.41 | 10.74 |
| | +trgPOSlm | 0.5453 | 34.53 | 10.73 |

**Table 6.8:** *Impact of including target POS-tag language model on translation scores.*

morphology variation and is not susceptible to benefit from this technique, and second that error analysis from §4.3.6 showed disagreement errors are not among the most significant.

In addition to that, the approach is simple in that no morphology generation is included and if the correct alternative does not exist in training, it cannot be favoured by the target POS-tag model anyway.

Following a similar idea and with similar results, in [Kir05] an alternative approach is introduced, consisting of incorporating a factorised target language model. Instead of using a standard word-based target language model, the authors factorise it (in other words, interpolate it) with language models based on Part-Of-Speech and stem information.

## 6.4 Morpho-reduced Translation Models and Morphology Post-processing

This section presents a study of the impact of morphology derivation on Ngram-based SMT models. For this purpose, we define a framework under the assumption that a certain degree of morphology information can be decoupled from the standard bilingual $n$-gram model, and introduced by performing a feature-based classification task.

Experiments assessing the validity of this assumption are carried out for the English→Spanish task, showing how much benefit SMT models can obtain by reducing Spanish morphology for each category, and describing the morphology classification task to be integrated with translation models.

### 6.4.1 Introduction

As it is well known, and as it can be seen throughout all the previous experiments reported in this Ph.D. dissertation, automatic evaluation scores when translating from English to Spanish are always lower than when translating in the opposite direction. The most reasonable explanation for this is that Spanish language, having a richer morphology than English, tends to be represented by a larger vocabulary set, making decisions harder for SMT systems (ie. models have more perplexity).

Indeed, whereas from Spanish to English several input words may share the same (or very close) translation probability distributions and translate into the same target words, from English to Spanish a single input word may present a wider range of possible translations. In other words, while in the Spa→Eng direction sparsity problems may arise in the source language (higher percentage of OOVs, few translation examples for each input word, etc.), in Eng→Spa these problems arise in the target language (higher perplexity in translation and target language models, etc.). Obviously, this also holds for language pairs such as English and Arabic, Catalan, Finnish, German, Italian or French, to name a few.

In the face of these facts, the question of how much morphology derivation is weakening the translation model estimates needs to be addressed. Is all morphology relevant for the Ngram-based SMT models? How much of this information is not being captured in our bilingual model? Is it feasible to learn independent models to generate this morphological information?

To illustrate this, consider the following bilingual training example, where Part-Of-Speech (POS) information is included for the last Spanish words. We note that 'POS' refers to Part-Of-Speech category ('VM','DA' and 'NC' meaning Main Verb, Determinant Article and Common Noun), 'M' refers to verb mode ('subj' meaning subjunctive), 'T' to tense ('pres' meaning present), 'P' to person ('3rd' meaning third), 'N' to number ('sing' and 'pl' meaning singular

and plural) and 'G' to gender ('fem' meaning feminine)[2]. Additionally, the base form for these words is shown in the bottom row. Of course, this same analysis could be conducted for the initial part of this example sentence.

| I ask you and your party to give support for the release | | | |
|---|---|---|---|
| Les pido a usted y a su partido que | respalden | la | liberación |
| Part-Of-Speech information → | POS:VM<br>M:sjve<br>T:pres<br>P:3rd<br>N:pl | POS:DA<br>G:fem<br>N:sing | POS:NC<br>G:fem<br>N:sing |
| base form → | [respaldar] | [el] | [liberación] |

**Table 6.9:** *Example of Eng→Spa bilingual training sentence.*

Regarding Spanish verb form '`respalden`', certain considerations need be taken into account. On the one hand, the reason for it to be in third person plural (`P:3rd`, `N:pl`) is the necessary subject-verb agreement, where the subject of the relative clause is '`usted y su partido`'. It seems obvious that this specific dependency cannot be learnt independently from lexical instances using a bilingual $n$-gram model. In other words, unless the exact input sequence '`you and your party to give`' is found in a test set, the bilingual model will not find an adequate estimate here.

On the other hand, the reason for this verb to be in present subjunctive (`M:sjve`,`T:pres`) is the structure '`pedir a alguien que HAGA algo`', or equivalently in English, '`to ask someone TO DO something`', where the capitalised word **must be** a verb in subjunctive mode, its tense depending on the tense of the preceding verb '`pedir`'. Again, it is evident that these complex dependencies cannot be captured by a bilingual $n$-gram model. During test decoding, a change in the person *being asked to do something* will lead to a very uninformed translation solution.

In addition to that, during training these morphology variations will cause various such subjunctive examples to be different if the number and person information differs (as the Spanish verb form changes), weakening the chances of this complex structure to be correctly translated.

When it comes to Spanish noun '`liberación`', its gender information is invariant, whereas it is reasonable to expect that its number information (`N:sing` or `N:pl`) will somehow depend on the English noun '`release`'.

And finally, regarding Spanish article '`la`', the reason for it being in feminine singular is solely its being followed by a feminine singular Spanish noun ('`liberación`'). No relevant information on this gender and number decision can be extracted neither from the English sentence nor from

[2]This set of Part-Of-Speech tags including rich morphology information (gender, number, person, tense, etc.) is used by the FreeLing toolkit mentioned earlier. Full details on the tag set can be freely obtained at http://garraf.epsevg.upc.es/freeling

the preceding Spanish verb '`respalden`'. Therefore, assuming a certain tuple segmentation as in Table 6.10, it is clear that the trigram defined by tuples $(T_3, T_4, T_5)$ is not useful to generate Spanish article '`la`'. Only the bigram defined by $(T_5, T_6)$ will be useful, or an additional target language model in case this bigram is not estimated in training.

| $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ |
|---|---|---|---|---|---|
| your | party | to | give support for | the | release |
| su | partido | que | respalden | la | liberación |

**Table 6.10:** *Example of tuple-segmented Eng→Spa bilingual sentence.*

One possible way to address this problematic situation derived from big Spanish vocabulary size is to increase English vocabulary size as done in [Uef03], where personal pronouns and modals are attached to English verbs to build new full forms, as already discussed in §2.6.

However, defining which relevant information must be attached to English verbs is not trivial (consider the previous example, where English verb '`give`' should be modified with markers indicating subjunctive, present and third person plural information). Furthermore, the vocabulary increase will, in many occasions, prevent the bilingual model from generalisation.

Therefore, we propose to follow a different strategy, which consists of excluding this morphological information from bilingual $n$-gram model, and estimating it as an *independent* classification NLP task. This approach is explained next.

### 6.4.2 Morpho-reduced Translation Models

#### 6.4.2.1 Training architectures

With the aim of assessing how much the morphology-based problems stated above are affecting the translation model, we propose the framework defined by the architecture from Figure 6.4.

As it can be seen, after standard word alignment and tuple extraction, we proceed to substituting target language words (Spanish) by a morphology-reduced version of them. Then we estimate the bilingual $n$-gram translation model with these new tuples. Optionally, this morphology reduction can also be performed at the tuple source-language side (English).

Morphology reduction is independent from word alignment and tuple extraction and could indeed be carried out before word alignment (possibly affecting alignment quality). However, given the lack of alignment-translation correlation in results from §5.4.4, we obviate this time-consuming process by reducing words directly in extracted tuples.

Several types of morphology reductions can be applied, depending on which Part-Of-Speech

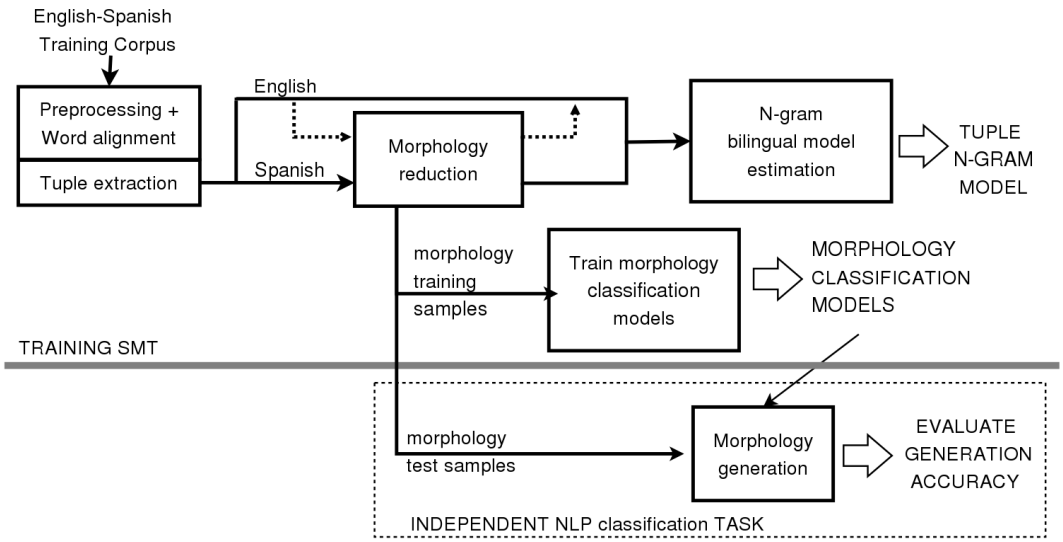



**Figure 6.4:** Training morphology-reduced translation models. Flow diagram.

category is modified (Verbs, Nouns, Adjectives, etc.) or which morphology attribute is modified. For instance, mode, tense, person or number for verbs, or person, number and gender of adjectives. Combinations of these reductions can make the set of possible morpho-reduced models grow endlessly. Section 6.4.4 presents all the reduction configurations investigated.

The result is a standard bilingual model producing morphology-reduced Spanish. This translation process may be independently evaluated if compatible morphology-reduced references are provided. We note that this result may in some cases[3] represent an oracle estimate of the score that could be achieved in case all reduced morphology was eventually instanced matching the original reference(s).

Additionally, the morphology reduction module produces a set of morphology samples including correct morphology class plus a set of features. These samples can be used to train morphology classification (or generation) models. Any strategy capable of estimating the correct morphology class given the sample and its features can be implemented here. For example, manual rules or machine learning techniques. The advantage is that this morphology generation task can be independently evaluated if a certain number of training samples is reserved as test (or development) set, as illustrated below the gray line from Figure 6.4.

To illustrate this process, let us consider again the example from Table 6.9 and assume that only Spanish verb mode and tense are reduced. In this case, the verb form '`respalden`' is transformed into '`VMmt3P[respaldar]`', indicating reduced POS and base form. Under this reduction, the POS keeps information on word category ('VM'→Main Verb), person and number ('3P'→third plural), whereas 'm' and 't' represent *any* mode and tense.

[3]Experiments in §6.4.4 will show that this is not always a *really achievable* oracle.

In addition to that, as the correct mode and tense for this verb is known beforehand, this serves as a classification training sample. Assuming that a certain number of statistical or linguistic features describing this sample can be useful to induce its mode and tense, we can train morphology classification models.

#### 6.4.2.2 Decoding architectures

During translation decoding, two possible architectures can be followed (see Figure 6.5). On the one hand, the sequential approach depicted above produces a single 1-best morphology-reduced translation output, performing a posterior final morphology generation independently.

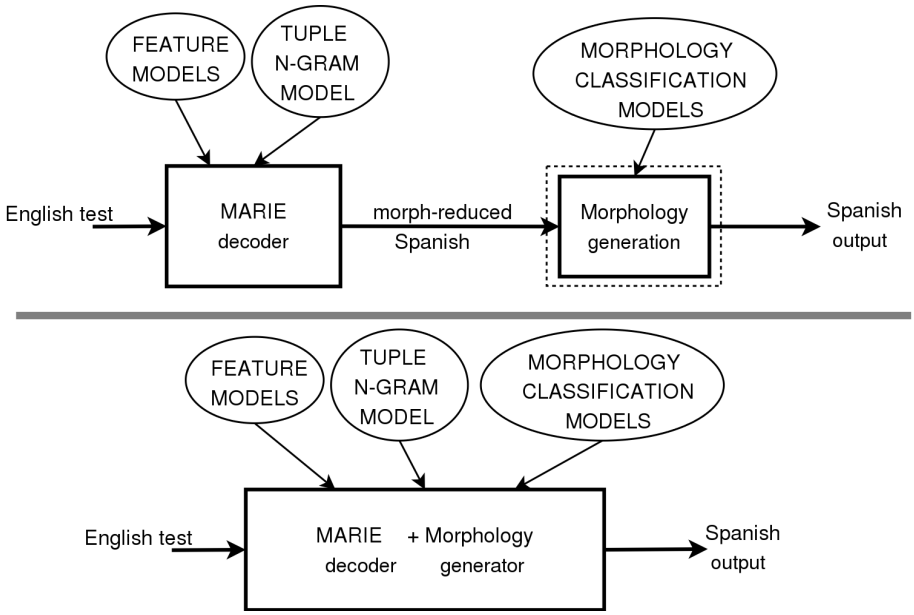



**Figure 6.5:** Sequential and integrated translation architectures for morphology-reduced translation models.

On the contrary, the integrated strategy depicted below would take morphology generation into account at decoding time. Regarding additional feature functions (lexicon models and target language model), if the sequential architecture is followed, these must be estimated using morphology-reduced parallel texts. In the integrated architecture, these can either be estimated using morphology-reduced parallel texts or using standard texts, so long as the decoding tool *is aware* of the criterion followed.

Finally, an intermediate architecture would apply morphology generation during N-best hypothesis re-ranking (see §2.4.2), but as this technique was not used in the system for this Ph.D. research work (see §4.2.2), we did not apply it for this experimental study.

In the present study the strictly sequential architecture will be followed. Apart from its simple integration with the Ngram-based SMT system, we believe that it serves as a good approximation of the possible modelling gains due to morphology reduction.

### 6.4.3 Morphology Post-processing

For the sake of comparison, the current 1-best standard translation output can also be passed through a morphology reduction module. In order to do that, Part-Of-Speech tagging and lemmatisation of the Spanish output is needed. This linguistic processing will probably contain more errors than usual due to the fact of the sentence not being a correct Spanish but a machine translation hypothesis.

By evaluating this morphology-reduced version of the current translation output against compatible morphology-reduced references, we obtain an *oracle estimate for morphology post-processing*. The difference between this oracle estimate and the morpho-reduced bilingual $n$-gram model oracle estimate will represent how much gain we can expect to obtain by estimating morpho-reduced translation models.

### 6.4.4 Experimental study

This experimental work was conducted using the European Parliament corpus version 3 (see details in Table 4.3). The following morphology reduction configurations were considered:

| | |
|---|---|
| S:$D$ | Spa determiners: gender and number |
| S:$A$ | Spa adjectives: gender and number |
| S:$N$ | Spa nouns: gender and number |
| S:$P$ | Spa pronouns: person, number and gender |
| S:$V$ | Spa verb: person and number |
| S:$V_M$ | Spa verb: person, number and mode |
| S:$V_{MT}$ | Spa verb: person, number, mode and tense |
| S:$V_{MT}$+E:$VP$ | S:$V_{MT}$ + Eng pronouns: person, number and verb: person, number |
| S:$DAV_{MT}$ | S:$V_{MT}$ + S:$D$ + S:$A$ |
| S:$full$ | S:$V_{MT}$ + S:$D$ + S:$A$ + S:$N$ + S:$P$ |
| S:$full$+E:$full$ | S:$full$ +E:$VP$ + Eng nouns: number |

**Table 6.11:** *Morphology reduction configurations considered in Eng→Spa translation.*

We note that in all cases the original form is transformed into its POS reduced tag plus its base form. For each base form, only those attributes exhibiting different values in Spanish are allowed. For this, we use human-written lists of possible tagging and lemmatising solutions from the FreeLing package.

For example, most Spanish nouns do not vary in gender (eg. '`liberación`' is always feminine), so that only those nouns varying will reduce their gender information. Table 6.12 presents an

example to illustrate these configurations. Note that all personal pronouns are assumed to have a unique base form (indicated by '[]') and that English verb reduction only applies to the distinction between 3rd person singular and others in present tense.

| English | she | has | a | strong | interest | in |
|---|---|---|---|---|---|---|
| Eng POS | PRP | VBZ | DT | JJ | NN | IN |
| Spanish | ella | tiene | el | máximo | interés | en |
| Spa POS | PP3SF | VMIP3S | DAMS | AQMS | NCMS | SPS |
| S:$D$ | ella tiene DAgn[el] máximo interés en | | | | | |
| S:$A$ | ella tiene el AQgn[máximo] interés en | | | | | |
| S:$N$ | ella tiene el máximo NCMn[interés] en | | | | | |
| S:$P$ | PPpng[] tiene el máximo interés en | | | | | |
| S:$V$ | ella VMIPpn[tener] el máximo interés en | | | | | |
| S:$V_M$ | ella VMmPpn[tener] el máximo interés en | | | | | |
| S:$V_{MT}$ | ella VMmtpn[tener] el máximo interés en | | | | | |
| S:$DAV_{MT}$ | ella VMmtpn[tener] DAgn[el] AQgn[máximo] interés en | | | | | |
| S:$full$ | PPpng[] VMmtpn[tener] DAgn[el] AQgn[máximo] NCMn[interés] en | | | | | |
| E:$VP$ | PRPpng[] VBPp[have] a very strong interest in | | | | | |
| E:$full$ | PRPpng[] VBPp[have] a very strong Nn[interest] in | | | | | |

**Table 6.12:** *Examples for each morphology reduction configuration.*

#### 6.4.4.1 Post-processing oracles

The post-processing oracle results for each morphology configuration are shown in Table 6.13, contrasting BLEU and NIST scores when applying only the bilingual model (onlyBM) and the full log-linear model combination (full). For each case, corresponding morphology-reduced references are used.

Furthermore, we show for each case number of output generated words (trgwrds) and percentage of these being modified due to morphology reduction. For this post-processing case, the number of output words is obviously fixed. The amount of words reduced represents the amount of generation decisions that need be taken in order to obtain the final Spanish text.

As it can be seen, the most promising oracles when reducing one single class are determiners and verbs (when reducing mode and tense as well). The bilingual model oracle for determiners ('S:$D$') improves BLEU in 1 point absolute[4] over baseline, and only around 0.8 points with the full configuration. Verbs when reducing person, number, mode and tense ('S:$V_{MT}$') present the same oracle in bilingual modelling, whereas oracle grows up to 1.3 points for full configuration[5]. Determiners morphology reduction involves ∼14% of generated words, whereas in the case of verbs this amount is only ∼9%.

[4] We denote one point as 0.01 BLEU.

[5] It is important to bear in mind that whereas results for the onlyBM configuration assess the quality of the translation model estimation, the full result is affected by the use of an imperfect model weight optimisation stage.

| | onlyBM | | | | full | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU | NIST | trgwrds | (%red) | BLEU | NIST | trgwrds | (%red) |
| baseline | 0.4264 | 9.187 | 25,313 | (0%) | 0.4785 | 9.889 | 25,733 | (0%) |
| S:$D$ | **0.4354** | 9.252 | 25,313 | (13.3%) | **0.4853** | 9.925 | 25,733 | (13.7%) |
| S:$A$ | 0.4306 | 9.253 | 25,313 | (8.1%) | 0.4823 | 9.947 | 25,733 | (7.9%) |
| S:$N$ | 0.4269 | 9.171 | 25,313 | (21.7%) | 0.4789 | 9.880 | 25,733 | (21.8%) |
| S:$P$ | 0.4273 | 9.194 | 25,313 | (0.8%) | 0.4793 | 9.893 | 25,733 | (0.8%) |
| S:$V$ | 0.4312 | 9.242 | 25,313 | (8.7%) | 0.4833 | 9.941 | 25,733 | (8.3%) |
| S:$V_M$ | 0.4337 | 9.277 | 25,313 | (8.7%) | 0.4863 | 9.977 | 25,733 | (8.3%) |
| S:$V_{MT}$ | **0.4369** | **9.315** | 25,313 | (8.7%) | **0.4916** | **10.036** | 25,733 | (8.3%) |
| S:$DAV_{MT}$ | 0.4529 | 9.478 | 25,313 | (31.6%) | 0.5042 | 10.148 | 25,733 | (31.4%) |
| S:$full$ | **0.4576** | **9.525** | 25,313 | (54.1%) | **0.5093** | **10.203** | 25,733 | (54.0%) |

**Table 6.13:** *Morphology post-processing oracles for each reduction configuration (Eng→Spa).*

Adjectives ('S:$A$') present a more reduced oracle gain of only 0.5 absolute BLEU (for both onlyBM and full), requiring to take morphology decisions on up to ∼8% of the target words. In contrast to that, reduction of nouns involves modifying up to ∼22% target words without a promising oracle result. Finally, pronouns involve very few words, with a very reducing impact on translation oracle scores.

By reducing all Spanish categories ('S:$full$') BLEU oracle reaches an approximate 3 point increase by modifying more than half the target words (54.0%). However, reducing determiners, adjectives and verbs oracle results are only 0.5 point lower, but reducing only 31% of the generated words (see 'S:$DAV_{MT}$').

#### 6.4.4.2 Study of post-processing oracles

Post-processing oracles should be correlated with manual error analysis conclusions as drawn in §4.3.6, even though only morphological differences between translation and references are considered here. Certainly the error analysis reflected an important amount of verb errors, without distinguishing between morphology or lexical errors.

Regarding determiners, error analysis did not reflect such a strong importance, except a few disagreement errors (see Table 4.33). For this reason, we now analyse the obtained oracles in detail. In order to do that, we study 100 random sentences whose oracle Word Error Rate is better than baseline.

The result is shown in Table 6.14, where four basic cases are distinguished. Firstly, an adjacent or far disagreement error is denoted whenever the adequate Spanish noun is placed and its adjacent or nearby determiner does not present agreement. In most of the cases, the use of monotone search limits the capacity of the bilingual model.

| Case | Percentage | Example |
|---|---|---|
| Adjacent disagreement error | 9% | `nombrar a un Comisión (REF:una Comisión)` |
| Far disagreement error | 39% | `las atroces situación (REF:la atroz situación)` |
| Wrong Translated noun | 29% | `las cautiverio (REF:los cautivos ← the captives)` |
| Reference noun mismatch | 23% | `las Naciones Unidas (REF:la ONU)` |

**Table 6.14:** Morphology reduction for determiners. Post-processing oracle analysis.

On the other hand, whenever the adequate Spanish noun is wrongly produced or omitted (29% of the cases), the determiner contributes to a false oracle, as there is no information to instantiate it correctly. Finally, whenever a different correct noun is produced (22.6% of the cases), the morphology reduction leads again to a false oracle increase.

In contrast with these findings, a similar overview study of 100 sentences reveals that oracle for verbs represents in 70% of the cases a true verb form morphology error (see Table 6.15). The remaining cases include a third person confusion and differences between reference and correct translation.

| Case | Percentage | Example |
|---|---|---|
| Verb error | 69% | `la Unión Europea , que legalizaron (REF:legalizó)` |
| 3rd person confusion | 14% | `como sabe (REF:como saben ← as you know)` |
| Reference mismatch | 17% | `el pueblo prefiere (REF:los ciudadanos prefieren)` |

**Table 6.15:** Morphology reduction for verbs. Post-processing oracle analysis.

We note that English word 'you' is either translated as third person singular or plural in Spanish, depending on the context. Unless the English sentence introduces the subject (as in '`mister President , as you know`'), then both singular and plural translations are valid, which causes mismatch with references in 14% of the cases. Together with the remaining mismatch cases, these situations do not represent a true morphology oracle.

#### 6.4.4.3 Morpho-reduced model oracles

In general terms, oracles from Table 6.13 are reduced, especially taking into account the analyses from the previous section. It seems clear that morphology post-processing will not produce a very strong impact on translation quality, especially regarding determiners, nouns and adjectives.

To evaluate the possible gain of estimating morphology-reduced models, as introduced in §6.4.2, the same translation oracles for each reduction configuration are presented in Table 6.16.

Interestingly, translation oracles for determiners ('S:*D*') do not improve with respect to the post-processing case. In fact, whereas BLEU score is slightly lower, NIST is slightly higher. This fact indicates that morphology reduction for determiners does not contribute to estimate

| | onlyBM | | | | full | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU | NIST | trgwrds | (%red) | BLEU | NIST | trgwrds | (%red) |
| baseline | 0.4264 | 9.187 | 25,313 | (0%) | 0.4785 | 9.889 | 25,733 | (0%) |
| S:$D$ | 0.4349 | 9.263 | 25,369 | (13.5%) | 0.4840 | 9.936 | 25,697 | (14.3%) |
| S:$A$ | 0.4306 | 9.267 | 25,310 | (8.1%) | 0.4819 | 9.984 | 25,833 | (8.2%) |
| S:$N$ | 0.4270 | 9.184 | 25,309 | (21.9%) | – | – | – | |
| S:$P$ | 0.4279 | 9.198 | 25,357 | (0.9%) | – | – | – | |
| S:$V$ | 0.4322 | 9.258 | 25,418 | (9.1%) | 0.4888 | 9.916 | 26,733 | (9.0%) |
| S:$V_M$ | 0.4366 | 9.318 | 25,388 | (9.1%) | 0.4882 | 10.025 | 25,628 | (8.9%) |
| S:$V_{MT}$ | **0.4394** | **9.353** | 25,446 | (9.4%) | **0.4972** | **10.085** | 26,189 | (9.5%) |
| S:$DAV_{MT}$ | 0.4576 | 9.575 | 25,470 | (32.3%) | 0.5060 | 10.227 | 25,433 | (32.8%) |
| S:$full$ | **0.4632** | **9.637** | 25,491 | (54.8%) | **0.5168** | **10.325** | 26,028 | (55.3%) |
| S:$full$+E:$full$ | 0.4635 | 9.618 | 25,645 | (54.7%) | 0.5143 | 10.250 | 26,209 | (55.1%) |

**Table 6.16:** *Morphology-reduced model oracles for each reduction configuration (Eng→Spa).*

a better translation model. The same conclusions can be drawn regarding adjectives ('S:$A$').

On the other hand, verb morphology reduction ('S:$V_{MT}$') does improve translation oracles, though not significantly (0.3 point difference with only bilingual model, 0.7 in full system). Verb reduction involves having to take morphology generation decisions for ∼9% of the produced words.

Additionally, combined reductions ('S:$DAV_{MT}$' or 'S:$full$') yield similar difference margins against post-processing (∼0.5 point difference with only bilingual model, 0.7 in full system), by reducing around 32% and 55% of Spanish words, respectively. Therefore, we can conclude that oracle improvement is basically due to verb morphology reduction. Note that English morphology reduction ('+E:$full$') does not produce any relevant oracle improvement.

Oracle difference is in many cases higher in the full case than in the onlyBM case. This result indicates that the sequential approach from §6.4.2.2 would not harm the chances of a posterior morphology generation module.

Finally, we observe that morphology reduction also has an effect on target sentence length. In other words, in all experiments morphology-reduced models tend to produce longer output sentences than the post-processing case, or equivalently, to use less tuples translating to NULL.

This indicates that even though it is positive to map Spanish words to their morphology reduced version, morphology is not the main reason explaining translation errors. Syntax-aware models are required, dealing not only with word order issues but also with long-range lexical and morphology dependencies.

### 6.4.5 Conclusions

This section presented a thorough study of the incidence of morphology in English→Spanish Ngram-based SMT models. We defined a framework to evaluate, for each morphology word category, the difference between estimating morphology-reduced models against post-processing the current baseline morphology.

Results reveal that Ngram-based SMT models estimated when reducing Spanish verb morphology produce longer target sentences and increase translation oracles. For the remaining word categories (Determiners, Adjectives, Nouns or Pronouns) no relevant positive effect of reduction is produced.

On the other hand, oracle difference against the post-processing approach is not significant. This indicates that although verb morphology contributes to more perplex models, this is probably not the main source of errors. Long-distance lexical and morphology dependencies (such as those highlighted in example from Table 6.10) demand for a structurally different translation model.

Therefore, in combination with morphology reduction, syntax-aware features need to be incorporated into the *N*gram-based SMT system. Future research work should be directed towards this goal, as morphology treatment only accounts for reduced error margins.

## 6.5 Chapter Summary and Conclusions

This chapter considered the application of linguistic-aware techniques in the framework of statistical machine translation modelling.

Section §6.2 was devoted to a translation model containing verb form classification, realising actual verb forms via the inclusion of a relative-frequency instance model in the log-linear model combined search. Experiments for small- and large-data English→Spanish situations were reported.

While results were very promising for a small-data task showing a significant performance boost, improvement is unfortunately cancelled when scaling to a large-data training situation. Although the simple instance model is behaving sufficiently well, the verb classification approach (which includes word grouping) is harming the bilingual model in the European Parliament task.

Possibly, the classification of all verb forms into a single token is excessive, and should better be converted into a set of tokens including certain types of verb forms. Alternative ways to handle morphology variation are introduced in the following section.

In Section 6.3, a simple target Part-Of-Speech language model to address disagreement problems is introduced, with moderate improvement, especially for English→Spanish translation.

To conclude, Section 6.4 investigated the effects of Spanish morphology in English→Spanish Ngram-based translation modelling. In conclusion, it is positive to reduce Spanish verb morphological information in order to estimate a bilingual $n$-gram model. The resultant model is capable of producing more target words and with a slightly better lexical accuracy. However, impact is reduced and syntax-aware techniques must be incorporated for the Ngram-based SMT approach to capture the complex lexical and morphology dependencies of the English–Spanish language pair.

# Chapter 7

# Conclusions and Future Work

This Ph.D. dissertation has considered the use of linguistic knowledge into statistical machine translation. Alternatively to most of current SMT systems based on phrase-based translation models, a joint-probability approach estimating an $N$-gram of bilingual tuples has been used, with competitive results.

Regarding techniques for taking advantage of linguistic knowledge, these have been applied to tuple segmentation, word alignment, verb form classification and translation model estimation. The main conclusions from this work are summarised next.

## 7.1 Conclusions

**Chapters 3 and 4** are devoted to explain in full detail the statistical machine translation model based on tuple $N$-grams, or $N$gram-based SMT. This approach is an evolution of a previous Finite-State Transducer implementation of $X$-grams, which adapted speech recognition tools for speech-oriented MT.

Despite its bilingual nature, the tuple translation model behaves similarly to monolingual language models when it comes to pruning and smoothing options, so that standard techniques can be efficiently applied. As widely done in literature, additional feature functions can be combined under a maximum entropy approach to the translation model in order to achieve significant quality gain.

The result is a competitive statistical machine translation model whose basic unit is the tuple. We studied the impact of alternative tuple definitions on the translation model, and proposed a segmentation criterion based on Part-Of-Speech entropy information for solving the problem of source NULLs.

The system achieves high-quality translation for relatively monotone language pairs

(Catalan↔Spanish, English↔Spanish), whereas more elaborate word order models need to developed for more complex language pairs (Chinese↔English).

Still, a human error analysis proves that, even in a situation of large-data availability (∼35M words), severe morphology and syntax errors persist, which derive from the model inability to handle word derivation, multi-word expressions, long-dependency syntax relationships, or semantic disambiguation, among other linguistic phenomena. Without providing somehow this information to our models, we can only rely on increasing training data indefinitely for improving current translation quality.

In this direction, **Chapter 5** is devoted to introducing alternative techniques to include information on morphology derivation and verb group information into word alignment algorithms. In terms of Alignment Error Rate, stemming and verb form classification achieve the most remarkable error reduction for both small and large-data tasks.

Regarding incidence on translation, unfortunately the bilingual $N$-gram model shows insensitivity to these improvements when large amounts of data are available, while showing a negative tendency towards reducing translation output as more training tuples are produced.

**Chapter 6** develops a verb classification strategy directly into the statistical translation model by decoupling the instanced verb form from the bilingual context, and incorporating a novel instance model into the log-linear combination. This approach is specially devised for situations of data sparseness, as it even allows for generalisation, thus producing verb forms which were unseen in training material.

English→Spanish experiments show significant improvements in a situation of small-data availability. However, as more data is available, the approach fails to increase performance over the best log-linear feature combination result. By mapping structurally different verb forms into a single token, the proposed classification seems to agglutinate too many bilingual contexts into a unique case, obtaining a worse translation model performance.

The Chapter also introduces a simple target Part-Of-Speech language model to address disagreement problems. Impact is moderate yet positive for both translation directions, indicating the positive smoothing effect brought to actual words by their POS tags.

Finally, the Chapter concludes with a detailed analysis of the impact of morphology variation in bilingual translation modelling. We introduce a framework to evaluate the impact of morphology derivates for each Spanish category onto the Ngram-based SMT system, based on morphology reduction. Results indicate that there is a slight gain in reducing Spanish verbs in terms of person, number, mode and tense information, obtaining longer translation outputs with better scores. However, more significant improvement will only be achieves if syntax-aware

features are incorporated into the system.

## 7.2 Future Work

There exist several lines for future research that can be taken as an extension of the work carried out in this dissertation. Without aiming at completeness, some of them are mentioned here.

- Concerning the *N*gram-based approach to statistical machine translation, some modelling issues are still in need of clarification, such as tuples translating to NULL. Further research is needed in order to optimally define these tuples and fully understand their contribution to the bilingual model.

  This research can be addressed via *a priori* decisions when segmenting parallel sentences (as preliminary done in §3.4.2.3), to differently weighing these examples in language modelling (and/or smoothing), or even introducing target NULL penalties in the log-linear model combination.

  The drawback of always producing less output words than the number of input words can thus be tackled. In combination with this, the preference of lexicon feature models and certain optimisation scores for short sentences should be compensated.

- Given that correct word order proves to be a big problematic issue for Ngram-based SMT (as shown not only in Chinese→English experiments, but also in Spanish↔English), syntax-aware word-order models need to be defined, as long as parsing tools are being made available.

  In this sense, the works in [Cre06b, Cre06c] show significant improvement by including reordering rules defined over POS-tag sequences. Extending this approach, rules based on syntactic information could be defined to capture long-dependency relationships.

- Concerning the use of linguistic information (and always taking into account that any technique is subject to be language-dependent), the proposed verb form classification can be improved for large-data tasks by adding some granularity to the classes. In other words, classifying all verb forms from a certain verb into more than just one class, which proves to be too gross for estimating the bilingual model. Alternative classification schemes can be applied.

  In addition to that, the proposed instance model based on relative frequency in §6.2.2.1 could be estimated via machine learning techniques, such as boosting [Sch99] or Support Vector Machines. The underlying assumption is that certain morphology information related to the verb form can be transferred by alternative models which do not necessarily need an Ngram-based SMT formulation, and in fact, may limit this *N*-gram capabilities.

- Including deeper linguistic knowledge into SMT, class-based bilingual $N$gram-models could be devised, in which classes would contain not only words (or morphology-reduced words) but also word groups (such as chunks or *phrases*, in a linguistic sense).

  As long as a shallow parser or a chunker is available, the bilingual model could benefit from estimating longer dependencies, for instance, by noun phrase classification. This information could be forced directly into the model by classifying input data, or via defining factorised bilingual translation models.

  In combination with this, the proposed framework of morphology reduction should be extended to deal with syntactic dependencies. By reducing not only morphological information but also lexical information, syntactic features could be defined to better capture long-distance word relationships that determine their base form or morphology.

- Finally, from a more statistical point of view but aiming at the same objective, tuple definition could also be revised and extended by allowing hierarchical classes, in the fashion of [Chi05] for phrase-based approaches.

Several other strategies could be followed. All in all, given current state-of-the-art performance (which still includes relevant mistakes) and given the increasing amount of research effort focused on SMT, it is reasonable to expect significant improvements in the field in the years to come.

# Bibliography


[Abe06] Antonio J. Abellán, *Traducción automática estadística castellano–catalán de gran vocabulario usando N-gramas*, Masters Thesis in Telecommunication Engineering, Dep. de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya, 2006.

[Aki04] Y Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii, "Overview of the iwslt04 evaluation campaign", *Proc. of the 1st Int. Workshop on Spoken Language Translation, IWSLT'04*, pags. 1–12, October 2004.

[Als00] H. Alshawi, Sh. Douglas, and S. Bangalore, "Learning dependency translation models as collections of finite-state head transducers", *Computational Linguistics*, Vol. 26, nº 1, pags. 45–60, March 2000.

[AO99] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F.J. Och, D. Purdy, N.A. Smith, and D. Yarowsky, "Statistical machine translation: Final report", Tech. rep., Johns Hopkins University Summer Workshop, Baltimore, MD, USA, 1999.

[Arn95] D. Arnold, and L. Balkan, "Machine translation: an introductory guide", *Comput. Linguist.*, Vol. 21, nº 4, pags. 577–578, 1995.

[Arr04] V. Arranz, E. Comelles, D. Farwell, C. Nadeu, J. Padrell, A. Febrer, D. Alexander, and K. Peterson, "A speech-to-speech translation system for catalan, spanish and english", *Proc. of the 6th Conf. of the Association for Machine Translation in the Americas*, October 2004.

[Arr05] V. Arranz, E. Comelles, and D. Farwell, "The fame speech-to-speech translation system for catalan, english and spanish", *Proc. of the MT Summit X*, pags. 195–202, September 2005.

[Aru06] A. Arun, A. Axelrod, A. Birch Mayne, Ch. Callison-Burch, H. Hoang, P. Koehn, M. Osborne, and D. Talbot, "Edinburgh system description for the 2006 tc-star spoken language translation evaluation", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 37–41, Barcelona, Spain, June 2006.

[Ats98] J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo, "Morphosyntactic analysis and parsing of unrestricted spanish text", *1st Int. Conf. on Language Resources and Evaluation, LREC'98*, 1998.

[Aya06] N.F. Ayan, and B.J. Dorr, "Going beyond aer: An extensive analysis of word alignments and their impact on mt", *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 9–16, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Bab04] B. Babych, and T. Hartley, "Extending the bleu mt evaluation method with frequency weightings", *42nd Annual Meeting of the Association for Computational Linguistics*, pags. 621–628, July 2004.

[Ban01] S. Bangalore, and G. Riccardi, "Finite-state models for lexical reordering in spoken language translation", *2nd Meeting of the North American Chapter of the Assoc. for Comp. Linguistics., NAACL*, June 2001.

[Ban05a] R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, and J.B. Mariño, "Statistical machine translation of euparl data by using bilingual n-grams", *Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pags. 67–72, June 2005.

[Ban05b] S. Banerjee, and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments", *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pags. 65–72, June 2005.

[Bat04] Montserrat Batlle, *Generació d'unitats bilingües per a la traducció estadística*, Masters Thesis in Informatics, Dep. de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, 2004.

[Ber96] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, Vol. 22, nº 1, pags. 39–72, March 1996.

[Ber05] N. Bertoldi, and M. Federico, "A new decoder for spoken language translation based on confusion networks", *IEEE Automatic Speech Recognition and Understanding Workhsop, ASRU'05*, December 2005.

[Ber06] N. Bertoldi, R. Cattoni, M. Cettolo, B. Chen, and M. Federico, "Itc-irst at the 2006 tc-star slt evaluation campaign", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 19–24, Barcelona, Spain, June 2006.

[BH60] Y. Bar-Hillel, "The present state of automatic translation of languages", *Advances in Computers*, Vol. 1, pags. 91–163, 1960.

[Bla04] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing, "Confidence estimation for machine translation", *Proc. of the 20th Int. Conf. on Computational Linguistics, COLING'04*, pags. 315–321, August 2004.

[Blu06] Ph. Blunsom, and T. Cohn, "Discriminative word alignment with conditional random fields", *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 65–72, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Bon96] A. Bonafonte, and J.B. Mariño, "Language modeling using X-grams", *Proc. of the 4th Int. Conf. on Spoken Language Processing, ICSLP'96*, pags. 394–397, October 1996.

[Bon98] A. Bonafonte, and J.B. Mariño, "Using X-gram for efficient speech recognition", *Proc. of the 5th Int. Conf. on Spoken Language Processing, ICSLP'98*, pags. 2559–2562, December 1998.

[Bra00] T. Brants, "TnT – a statistical part-of-speech tagger", *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000.

[Bro90] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, and P.S. Roossin, "A statistical approach to machine translation", *Computational Linguistics*, Vol. 16, nº 2, pags. 79–85, 1990.

[Bro93] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics*, Vol. 19, nº 2, pags. 263–311, 1993.

[Bro99] R. D. Brown, "Adding linguistic knowledge to a lexical example-based translation system", *Proc. of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pags. 22–32, August 1999.

[Car04] X. Carreras, I. Chao, L. Padró, and M. Padró, "Freeling: An open-source suite of language analyzers", *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, May 2004.

[Cas00] F. Casacuberta, "Inference of finite-state transducers by using regular grammars and morphisms", A.L. Oliveira (ed.), *Grammatical Inference: Algorithms and Applications*, Vol. 1891 of *Lecture Notes in Computer Science*, pags. 1–14, Springer-Verlag, 2000, 5th International Colloquium Grammatical Inference -ICGI2000-. Lisboa. Portugal. Septiembre.

[Cas01] F. Casacuberta, "Finite-state transducers for speech-input translation", *IEEE Automatic Speech Recognition and Understanding Workhsop, ASRU'01*, December 2001.

[CB04] Ch. Callison-Burch, D. Talbot, and M. Osborne, "Statistical machine translation with word- and sentence-aligned parallel corpora", *42nd Annual Meeting of the Association for Computational Linguistics*, pags. 176–183, July 2004.

[CB06] Ch. Callison-Burch, M. Osborne, and Ph. Koehn, "Re-evaluating the role of bleu in machine translation research", *13th Conf. of the European Chapter of the Association for Computational Linguistics*, pags. 249–246, April 2006.

[Che96] S. Chen, and J. Goodman, "An empirical study of smoothing techniques for language modeling", *34th Annual Meeting of the Association for Computational Linguistics*, pags. 310–318, San Francisco, July 1996.

[Che98] S. Chen, and J. Goodman, "An empirical study of smoothing techniques for language modeling", Tech. Rep. TR-10-98, Computer Science Group, Harvard University, Harvard, USA, 1998.

[Che03a] C. Cherry, and D. Lin, "A probability model to improve word alignment", *41st Annual Meeting of the Association for Computational Linguistics*, pags. 88–95, July 2003.

[Che03b] C. Cherry, and D. Lin, "Word alignment with cohesion constraint", *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, pags. 49–51, May 2003.

[Che05] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico, "The ITC-irst SMT system for IWSLT-2005", *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pags. 98–104, October 2005.

[Che06] C. Cherry, and D. Lin, "Soft syntactic constraints for word alignment through discriminative training", *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 105–112, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Chi05] D. Chiang, "A hierarchical phrase-based model for statistical machine translation", *43rd Annual Meeting of the Association for Computational Linguistics*, pags. 263–270, June 2005.

[Cj06] M.R. Costa-jussà, J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J.B. Mariño, J.A.R. Fonollosa, and R. Banchs, "Talp phrase-based statistical translation system for european language pairs", *Proceedings of the Workshop on Statistical Machine Translation*, pags. 142–145, Association for Computational Linguistics, New York City, June 2006.

[CO04] S. Corston-Oliver, and M. Gamon, "Normalizing german and english inflectional morphology to improve statistical word alignment", *Proc. of the 6th Conf. of the Association for Machine Translation in the Americas*, pags. 48–57, October 2004.

[Col05] M. Collins, Ph. Koehn, and I. Kucerova, "Clause restructuring for statistical machine translation", *43rd Annual Meeting of the Association for Computational Linguistics*, pags. 531–540, June 2005.

[Cre05a] J.M. Crego, A. de Gispert, and J.B. Mariño, "TALP: The UPC tuple-based SMT system", *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pags. 191–198, October 2005.

[Cre05b] J.M. Crego, J.B. Mariño, and A. de Gispert, "An ngram-based statistical machine translation decoder", *Proc. of the 9th European Conference on Speech Communication and Technology, Interspeech'05*, pags. 3193–3196, September 2005.

[Cre05c] J.M. Crego, J.B. Mariño, and A. de Gispert, "Reordered search and tuple unfolding for ngram-based smt", *Proc. of the MT Summit X*, pags. 283–89, September 2005.

[Cre06a] J.M. Crego, A. de Gispert, P. Lambert, M.R. Costa-jussà, M. Khalilov, R. Banchs, J.B. Mariño, and J.A.R. Fonollosa, "N-gram-based SMT system enhanced with reordering patterns", *Proceedings of the Workshop on Statistical Machine Translation*, pags. 162–165, Association for Computational Linguistics, New York City, June 2006.

[Cre06b] J.M. Crego, and J.B. Mariño, "Integration of postag-based source reordering into smt decoding by an extended search graph", *Proc. of the 7th Conf. of the Association for Machine Translation in the Americas*, pags. 29–36, August 2006.

[Cre06c] J.M. Crego, and J.B. Mariño, "Reordering experiments for n-gram-based smt", *1st IEEE/ACL Workshop on Spoken Language Technology*, December 2006.

[Dod02] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics", *Proc. ARPA Workshop on Human Language Technology*, 2002.

[Eck05] M. Eck, and Ch. Hori, "Overview of the IWSLT 2005 Evaluation Campaign", *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pags. 11–32, October 2005.

[EI06] A. El Isbihani, Sh. Khadivi, O. Bender, and H. Ney, "Morpho-syntactic arabic preprocessing for arabic to english statistical machine translation", *Proceedings of the Workshop on Statistical Machine Translation*, pags. 15–22, Association for Computational Linguistics, New York City, June 2006.

[FM05] José A. Fernández Molina, *Traductor estadístico castellano–catalán basado en corpus*, Masters Thesis in Telecommunication Engineering, Dep. de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya, 2005.

[Fra05] A. Fraser, and D. Marcu, "Isi's participation in the romanian-english alignment task", *Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pags. 91–94, June 2005.

[Fra06a] A. Fraser, and D. Marcu, "Measuring word alignment quality for statistical machine translation", Tech. Rep. ISI-TR-616, ISI/University of Southern California, May 2006.

[Fra06b] A. Fraser, and D. Marcu, "Semi-supervised training for statistical word alignment", *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 769–776, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Fri03] Pedro Frieros, *Traducción automática de texto mediante técnicas estadísticas*, Masters Thesis in Telecommunication Engineering, Dep. de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya, 2003.

[Gal91] W.A. Gale, and K.W. Church, "Identifying word correspondences in parallel texts", *4th Speech and Natural Language Workshop*, pags. 152–157, 1991.

[Ger01] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada, "Fast decoding and optimal decoding for machine translation", *39th Annual Meeting of the Association for Computational Linguistics*, pags. 228–235, July 2001.

[Gim06] J. Giménez, and E. Amigó, "Iqmt: A framework for automatic machine translation evaluation", *5th Int. Conf. on Language Resources and Evaluation, LREC'06*, pags. 22–28, May 2006.

[Gis02a] A. de Gispert, and J.B. Mariño, "Análisis de las relaciones cruzadas en el alineado estadístico para la traducción automática", *II Jornadas en Tecnología del Habla*, December 2002.

[Gis02b] A. de Gispert, and J.B. Mariño, "Using X-grams for speech-to-speech translation", *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, pags. 1885–1888, September 2002.

[Gis03] A. de Gispert, and J.B. Mariño, "Experiments in word-ordering and morphological preprocessing for transducer-based statistical machine translation", *IEEE Automatic Speech Recognition and Understanding Workhsop, ASRU'03*, pags. 634–639, November 2003.

[Gis04a] A. de Gispert, and J.B. Mariño, "TALP: Xgram-based Spoken Language Translation System", *Proc. of the 1st Int. Workshop on Spoken Language Translation, IWSLT'04*, pags. 85–90, October 2004.

[Gis04b] A. de Gispert, J.B. Mariño, and J.M. Crego, "Phrase-based alignment combining corpus cooccurrences and linguistic knowledge", *Proc. of the 1st Int. Workshop on Spoken Language Translation, IWSLT'04*, pags. 107–114, October 2004.

[Gis05a] A. de Gispert, "Phrase linguistic classification and generalization for improving statistical machine translation", *Proc. of the ACL Student Research Workshop*, pags. 67–72, June 2005.

[Gis05b] A. de Gispert, J.B. Mariño, and J.M. Crego, "Clasificación y generalización de formas verbales en sistemas de traducción estocástica", *Procesamiento del Lenguaje Natural, SEPLN'05*, pags. 69–76, September 2005.

[Gis05c] A. de Gispert, J.B. Mariño, and J.M. Crego, "Improving statistical machine translation by classifying and generalizing inflected verb forms", *Proc. of the 9th European Conference on Speech Communication and Technology, Interspeech'05*, pags. 107–114, September 2005.

[Gis06a] A. de Gispert, D. Gupta, M. Popovic, P. Lambert, J.B. Mariño, M. Federico, H. Ney, and R. Banchs, "Improving statistical word alignments with morpho-syntactic transformations", *Proceedings of 5th International Conference on Natural Language Processing, FinTAL'06*, pags. 368–379, August 2006.

[Gis06b] A. de Gispert, and J.B. Mariño, "Catalan-english statistical machine translation without parallel corpus: Bridging through spanish", *Proceedings of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages, SALTMIL'06*, pags. 65–68, May 2006.

[Gis06c] A. de Gispert, and J.B. Mariño, "Linguistic knowledge in statistical phrase-based word alignment", *Natural Language Engineering*, Vol. 12, nº 1, pags. 91–108, March 2006.

[Gis06d] A. de Gispert, and J.B. Mariño, "Linguistic tuple segmentation in ngram-based statistical machine translation", *Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'06*, pags. 1149–1152, September 2006.

[Gis06e] A. de Gispert, and J.B. Mariño, "Segmentación lingüística de tuplas para el modelado de la traducción estocástica mediante n-gramas", *Procesamiento del Lenguaje Natural, SEPLN'06*, pags. 241–248, September 2006.

[Gol05] Sh. Goldwater, and D. McClosky, "Improving statistical mt through morphological analysis", *Proc. of the Human Language Technology Conference and the Conference*

*on Empirical Methods in Natural Language Processing, HLT/EMNLP'05*, pags. 676–683, Association for Computational Linguistics, Vancouver, Canada, October 2005.

[GV02] I. García Varea, F.J. Och, H. Ney, and F. Casacuberta, "Improving word alignment quality using morpho-syntactic information", *Proc. of the 19th Int. Conf. on Computational Linguistics, COLING'02*, pags. 1051–1057, August 2002.

[GV03] I. García Varea, *Traducción automática estadística: modelos de traducción basados en máxima entropía y algoritmos de búsqueda*, PhD Thesis in Informatics, Dep. de Sistemes Informàtics i Computació, Universitat Politècnica de València, 2003.

[Hab06] N. Habash, and F. Sadat, "Arabic preprocessing schemes for statistical machine translation", *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pags. 49–52, Association for Computational Linguistics, New York City, USA, June 2006.

[Hew05] S. Hewavitharana, B. Zhao, A.S. Hildebrand, M. Eck, Ch. Hori, S. Vogel, and A. Waibel, "The cmu statistical machine translation system for IWSLT 2005", *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pags. 63–70, October 2005.

[Hut86] J.W. Hutchins, *Machine translation: past, present and future*, Ellis Horwood, Chichester, England, 1986.

[Hut92] W.J. Hutchins, and H.L. Somers, "An introduction to machine translation", 1992.

[Itt05] A. Ittycheriah, and S. Roukos, "A maximum entropy word aligner for arabic-english machine translation", *Proc. of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, HLT/EMNLP'05*, pags. 89–96, Association for Computational Linguistics, Vancouver, Canada, October 2005.

[Kan04] Stephan Kanthak, and Hermann Ney, "Fsa: An efficient and flexible c++ toolkit for finite state automata using on-demand computation", *42nd Annual Meeting of the Association for Computational Linguistics*, pags. 510–517, July 2004.

[Kir05] K. Kirchhoff, and M. Yang, "Improved language modeling for statistical machine translation", *Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pags. 125–128, June 2005.

[Kir06] K. Kirchhoff, M. Yang, and K. Duh, "Statistical machine translation of parliamentary proceedings using morpho-syntactic knowledge", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 57–62, Barcelona, Spain, June 2006.

[Kni99] K. Knight, "Decoding complexity in word replacement translation models", *Computational Linguistics*, Vol. 26, nº 2, pags. 607–615, 1999.

[Koe03] Ph. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-based translation", *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, May 2003.

[Koe04] Ph. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models", *Proc. of the 6th Conf. of the Association for Machine Translation in the Americas*, pags. 115–124, October 2004.

[Koe05a] Ph. Koehn, "Europarl: A parallel corpus for statistical machine translation", *Proc. of the MT Summit X*, pags. 79–86, September 2005.

[Koe05b] Ph. Koehn, and C. Monz, "Shared task: Statistical Machine Translation between European Languages", *Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pags. 119–124, June 2005.

[Koe06] Ph. Koehn, and C. Monz, "Manual and automatic evaluation of machine translation between european languages", *Proceedings of the Workshop on Statistical Machine Translation*, pags. 102–21, Association for Computational Linguistics, New York City, June 2006.

[Kuh06] R. Kuhn, G. Foster, S. Larkin, and N. Ueffing, "Portage phrase-based system for chinese-to-english translation", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 75–80, Barcelona, Spain, June 2006.

[Kum04] S. Kumar, and W. Byrne, "Minimum bayes-risk decoding for statistical machine translation", *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pags. 169–176, May 2004.

[Lam04] P. Lambert, and N. Castell, "Alignment of parallel corpora exploiting asymmetrically aligned phrases", *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, pags. 26–29, May 2004.

[Lam05] P. Lambert, A. de Gispert, R. Banchs, and J.B. Mariño, "Guidelines for word alignment evaluation and manual alignment", *Language Resources and Evaluation*, Vol. 39, nº 4, pags. 267–285, December 2005.

[Lee04] Y.S. Lee, "Morphological analysis for statistical machine translation", Daniel Marcu Susan Dumais, Salim Roukos (eds.), *HLT-NAACL 2004: Short Papers*, pags. 57–60, Association for Computational Linguistics, Boston, Massachusetts, USA, May 2004.

[Lee06] Y.S. Lee, Y. Al-Onaizan, K. Papineni, and S. Roukos, "Ibm spoken language translation system", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 13–18, Barcelona, Spain, June 2006.

[Lin04a] Chin-Yew Lin, "ROUGE: a package for automatic evaluation of summaries", *ACL 2004 Workshop: Text Summarization Branches Out*, Barcelona, Spain, July 2004.

[Lin04b] Chin-Yew Lin, and F.J. Och, "ORANGE: a method for evaluating automatic evaluation metrics for machine translation", *Proc. of the 20th Int. Conf. on Computational Linguistics, COLING'04*, pags. 501–507, August 2004.

[Liu05] Y. Liu, Q. Liu, and Sh. Lin, "Log-linear models for word alignment", *43rd Annual Meeting of the Association for Computational Linguistics*, pags. 459–466, June 2005.

[Mar02] D. Marcu, and W. Wong, "A phrase-based, joint probability model for statistical machine translation", *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, pags. 133–139, July 2002.

[Mar05] J. Martin, R. Mihalcea, and T. Pedersen, "Word alignment for languages with scarce resources", *Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pags. 65–74, June 2005.

[Mat05] E. Matusov, S. Kanthak, and H. Ney, "Efficient statistical machine translation with constrained reordering", *Proc. of 10th Annual Conference of the European Association for Machine Translation, EAMT'05*, pags. 181–188, May 2005.

[Mat06] E. Matusov, R. Zens, D. Vilar, A. Mauser, M. Popovic, S. Hasan, and H. Ney, "The rwth machine translation system", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 31–36, Barcelona, Spain, June 2006.

[Mel00] D. Melamed, "Models of translational equivalence among words", *Computational Linguistics*, Vol. 26, nº 2, pags. 221–249, 2000.

[Mel01] D. Melamed, *Empirical Methods for Exploiting Parallel Text*, MIT Press, Cambridge, MA, 2001.

[Mel04] D. Melamed, "Statistical machine translation by parsing", *42nd Annual Meeting of the Association for Computational Linguistics*, pags. 653–661, July 2004.

[Mih03] R. Mihalcea, and T. Pedersen, "An evaluation exercise for word alignment", Rada Mihalcea, Ted Pedersen (eds.), *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pags. 1–10, Association for Computational Linguistics, Edmonton, Canada, May 2003.

[Mil91] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi, "Five papers on wordnet", *Special Issue of International Journal of Lexicography*, Vol. 3, nº 4, pags. 235–312, 1991.

[Mn99] J.B. Mariño, and A. Nogueiras, "Top-down bottom-up hybrid clustering algorithm for acoustic-phonetic modeling of speech", *Proc. of the 6th European Conference on Speech Communication and Technology, Eurospeech'99*, pags. 1343–1346, September 1999.

[Mn00] J.B. Mariño, A. Nogueiras, P. Pachès, and A. Bonafonte, "The demiphone: an efficient contextual subword unit for continuous speech recognition", *Speech Communication*, Vol. 32, nº 3, pags. 187–197, October 2000.

[Mn06a] J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussà, "N-gram based machine translation", *Accepted for publication in Computational Linguistics*, 2006.

[Mn06b] J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R Fonollosa, M.R. Costa-jussà, and M. Khalilov, "Upc's bilingual n-gram translation system", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 43–48, Barcelona, Spain, June 2006.

[Moo06] R.C. Moore, W. Yih, and A. Bode, "Improved discriminative bilingual word alignment", *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 513–520, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Nel65] J.A. Nelder, and R. Mead, "A simplex method for function minimization", *The Computer Journal*, Vol. 7, pags. 308–313, 1965.

[Ney05] H. Ney, V. Steinbiss, R. Zens, E. Matusov, J. Gonzalez, Y.S. Lee, S. Roukos, M. Federico, M. Kolss, and R. Banchs, "Deliverable d5: Slt progress report", Tech. rep., TC-STAR, http://www.tc-star.org/documents/deliverable/Deliv_D5_Total_21May05.pdf, May 2005.

[Ney06] H. Ney, "Overview of the SLT evaluation: Spoken language translation", *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006.

[Nie00] S. Nießen, and H. Ney, "Improving SMT quality with morpho-syntactic analysis", *Proc. of the 18th Int. Conf. on Computational Linguistics, COLING'00*, pags. 1081–1085, July 2000.

[Nie04] S. Nießen, and H. Ney, "Statistical machine translation with scarce resources using morpho-syntactic information", *Computational Linguistics*, Vol. 30, nº 2, pags. 181–204, June 2004.

[Och99a] F.J. Och, "An efficient method for determining bilingual word classes", *9th Conf. of the European Chapter of the Association for Computational Linguistics*, pags. 71–76, June 1999.

[Och99b] F.J. Och, Ch. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation", *Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pags. 20–28, June 1999.

[Och00a] F.J. Och, and H. Ney, "A comparison of aligmnent models for statistical machine translation", *Proc. of the 18th Int. Conf. on Computational Linguistics, COLING'00*, pags. 1086–1090, July 2000.

[Och00b] F.J. Och, and H. Ney, "Improved statistical alignment models", *38th Annual Meeting of the Association for Computational Linguistics*, pags. 440–447, October 2000.

[Och02] F.J. Och, and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation", *40th Annual Meeting of the Association for Computational Linguistics*, pags. 295–302, July 2002.

[Och03a] F.J. Och, "Giza++ software. http://www-i6.informatik.rwth-aachen.de/~och/ software/giza++.html", Tech. rep., RWTH Aachen University, 2003.

[Och03b] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev, "Syntax for statistical machine translation", Tech. Rep. Summer Workshop Final Report, Johns Hopkins University, Baltimore, USA, 2003.

[Och03c] F.J. Och, and H. Ney, "A systematic comparison of various statistical alignment models", *Computational Linguistics*, Vol. 29, nº 1, pags. 19–51, March 2003.

[Och04a] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev, "A smorgasbord of features for statistical machine translation", *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pags. 161–168, May 2004.

[Och04b] F.J. Och, and H. Ney, "The alignment template approach to statistical machine translation", *Computational Linguistics*, Vol. 30, nº 4, pags. 417–449, December 2004.

[Olt06] M. Olteanu, Ch. Davis, I. Volosen, and D. Moldovan, "Phramer - an open source statistical phrase-based translator", *Proceedings on the Workshop on Statistical Machine Translation*, pags. 146–149, Association for Computational Linguistics, New York City, June 2006.

[Ort05] D. Ortiz, I. García-Varea, and F. Casacuberta, "Thot: a toolkit to train phrase-based statistical translation models", *Proc. of the MT Summit X*, pags. 141–148, September 2005.

[Pap98] K.A. Papineni, S. Roukos, and R.T. Ward, "Maximum likelihood and discriminative training of direct translation models", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pags. 189–192, May 1998.

[Pap01] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation", Tech. Rep. RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, 2001.

[Pat06] A. Patry, F. Gotti, and Ph. Langlais, "Mood at work: Ramses versus pharaoh", *Proceedings on the Workshop on Statistical Machine Translation*, pags. 126–129, Association for Computational Linguistics, New York City, June 2006.

[Pop04a] M. Popovic, and H. Ney, "Improving word alignment quality using morpho-syntactic information", *Proc. of the 20th Int. Conf. on Computational Linguistics, COLING'04*, pags. 310–314, August 2004.

[Pop04b] M. Popovic, and H. Ney, "Towards the use of word stems and suffixes for statistical machine translation", *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, pags. 1585–1588, May 2004.

[Pop06a] M. Popovic, A. de Gispert, D. Gupta, P. Lambert, H. Ney, J.B. Mariño, M. Federico, and R. Banchs, "Morpho-syntactic information for automatic error analysis of statistical machine translation output", *Proceedings of the Workshop on Statistical Machine Translation*, pags. 1–6, Association for Computational Linguistics, New York City, June 2006.

[Pop06b] M. Popovic, and H. Ney, "Error analysis of verb inflections in spanish translation output", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 99–103, Barcelona, Spain, June 2006.

[Pop06c] M. Popovic, and H. Ney, "Pos-based word reorderings for statistical machine translation", *5th Int. Conf. on Language Resources and Evaluation, LREC'06*, pags. 1278–1283, May 2006.

[Prz06] M. Przybocki, G. Sanders, and A. Le, "Edit distance: A metric for machine translation evaluation", *5th Int. Conf. on Language Resources and Evaluation, LREC'06*, pags. 2038–2043, May 2006.

[Qua05] V.H. Quan, M. Federico, and M. Cettolo, "Integrated n-best re-ranking for spoken language translation", *Proc. of the 9th European Conference on Speech Communication and Technology, Interspeech'05*, September 2005.

[Qui05] Ch. Quirk, A. Menezes, and C. Cherry, "Dependency treelet translation: Syntactically informed phrasal SMT", *43rd Annual Meeting of the Association for Computational Linguistics*, pags. 271–279, June 2005.

[Sch99] Robert E. Schapire, and Yoram Singer, "Improved boosting using confidence-rated predictions", *Machine Learning*, Vol. 37, nº 3, pags. 297–336, 1999.

[Sha49a] C.E. Shannon, "Communication theory of secrecy systems", *The Bell System Technical Journal*, Vol. 28, pags. 656–715, 1949.

[Sha49b] C.E. Shannon, and W. Weaver, *The mathematical theory of communication*, University of Illinois Press, Urbana, IL, 1949.

[Sha51] C.E. Shannon, "Prediction and entropy of printed english", *The Bell System Technical Journal*, Vol. 30, pags. 50–64, 1951.

[She04] L. Shen, A. Sarkar, and F.J. Och, "Discriminative reranking for machine translation", Daniel Marcu Susan Dumais, Salim Roukos (eds.), *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pags. 177–184, Association for Computational Linguistics, Boston, Massachusetts, USA, May 2004.

[Sno05] M. Snover, B. Dorr, R. Schwartz, J. Makhoul, L. Micciula, and R. Weischedel, "A study of translation error rate with targeted human annotation", Tech. Rep. LAMP-TR-126,CS-TR-4755,UMIACS-TR-2005-58, University of Maryland, College Park and BBN Technologies, July 2005.

[Sto02] A. Stolcke, "Srilm - an extensible language modeling toolkit", *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, pags. 901–904, September 2002.

[Tal06] D. Talbot, and M. Osborne, "Modelling lexical redundancy for machine translation", *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 969–976, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Til00] C. Tillmann, and H. Ney, "Word re-ordering and dp-based search in statistical machine translation", *Proc. of the 18th Int. Conf. on Computational Linguistics, COLING'00*, pags. 850–856, July 2000.

[Til03] C. Tillmann, and H. Ney, "Word reordering and a dynamic programming beam search algorithm for statistical machine translation", *Computational Linguistics*, Vol. 29, nº 1, pags. 97–133, March 2003.

[Tou02] K. Toutanova, H. Tolga Ilhan, and C.D. Manning, "Extensions to hmm-based statistical word alignment models", *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, pags. 87–94, July 2002.

[Tur03] J.P. Turian, L. Shen, and D. Melamed, "Evaluation of machine translation and its evaluation", *Proc. of the MT Summit IX*, September 2003.

[Uef03] N. Ueffing, and H. Ney, "Using pos information for smt into morphologically rich languages", *10th Conf. of the European Chapter of the Association for Computational Linguistics*, pags. 347–354, April 2003.

[Vid97] E. Vidal, "Finite-state speech-to-speech translation", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pags. 111–114, April 1997.

[Vil05] D. Vilar, E. Matusov, S. Hasan, R. Zens, and H. Ney, "Statistical machine translation of european parliamentary speeches", *Proc. of the MT Summit X*, pags. 259–266, September 2005.

[Vil06] D. Vilar, J. Xu, L.F. D'Haro, and H. Ney, "Error analysis of statistical machine translation output", *5th Int. Conf. on Language Resources and Evaluation, LREC'06*, pags. 697–702, May 2006.

[Vog96] S. Vogel, H. Ney, and C. Tillmann, "Hmm-based word alignment in statistical translation", *Proc. of the 16th Int. Conf. on Computational Linguistics, COLING'96*, pags. 836–841, August 1996.

[Vog00] S. Vogel, and H. Ney, "Translation with cascaded finite state transducers", *38th Annual Meeting of the Association for Computational Linguistics*, pags. 23–30, October 2000.

[Vog03] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao, and A. Waibel, "The cmu statistical translation system", *Proc. of the MT Summit IX*, September 2003.

[Wah00] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag, Berlin, 2000.

[Wan97] Y. Wang, and A. Waibel, "Decoding algorithm in statistical machine translation", *35th Annual Meeting of the Association for Computational Linguistics*, pags. 366–372, Association for Computational Linguistics, July 1997.

[Wea55] W. Weaver, "Translation", W.N. Locke, A.D. Booth (eds.), *Machine Translation of Languages*, pags. 15–23, MIT Press, Cambridge, MA, 1955.

[Wu97] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora", *Computational Linguistics*, Vol. 23, nº 3, pags. 377–403, September 1997.

[Yam01] K. Yamada, and K. Knight, "A syntax-based statistical translation model", *39th Annual Meeting of the Association for Computational Linguistics*, pags. 523–530, July 2001.

[Zen02] R. Zens, F.J. Och, and H. Ney, "Phrase-based statistical machine translation", M. Jarke, J. Koehler, G. Lakemeyer (eds.), *KI - 2002: Advances in artificial intelligence*, Vol. LNAI 2479, pags. 18–32, Springer Verlag, September 2002.

[Zen04] R. Zens, F.J. Och, and H. Ney, "Improvements in phrase-based statistical machine translation", *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pags. 257–264, May 2004.

[Zen05] R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney, "The RWTH phrase-based statistical machine translation system", *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pags. 148–154, October 2005.

[Zol06] A. Zollmann, A. Venugopal, and S. Vogel, "Bridging the inflection morphology gap for arabic statistical machine translation", *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pags. 201–204, Association for Computational Linguistics, New York City, USA, June 2006.