Chapter 8

Conclusions

Rehuya, al principio, formas y temas demasiado corrientes: son los más difíciles. Pues se necesita una fuerza muy grande y muy madura para poder dar de sí algo propio ahí donde existe ya multitud de buenos y, en parte, brillantes legados. Por esto, líbrese de los motivos de índole general. Recurra a los que cada día le ofrece su propia vida. Describa sus tristezas y sus anhelos, sus pensamientos fugaces y su fe en algo bello; y dígalo todo con íntima, callada y humilde sinceridad.

(R.M. Rilke, Cartas a un joven poeta)

With the aim of solving new real user needs, the objective of this thesis is to provide a flexible multitask summarizer architecture that should allow the integration of different Natural Language Processing techniques and tools. Given a new summarization challenge, new components have to be easily integrated allowing to re-use components when possible. So, once studied several Summarization models as well as different existing automatic systems, this main goal has been achieved by designing the Flexible Eclectic Multitask summarizer, a modular architecture to facilitate the re-usability of code. Moreover, in order to show the validity of our design, several instantiations of FEMsum have been implemented to solve different summarization tasks using state-of-art techniques. Each task is supposed to face specific user information needs, including: to process one single document vs. to process several documents; to take into account the information provided by a user query vs. to produce generic text-driven summaries; to assume a previous user knowledge giving only just-the-new information vs. the production of background summaries; and to produce summaries of different lengths (ranging from 10 to 250 words). This general framework has to be able to deal with different kinds of digital documents such as news reports or scientific presentations. This input can be of different media (e.g. text, speech), languages (e.g. English, Spanish) or domains (e.g. journalistic, scientific).

Some FEMsum instantiations have been evaluated when summarizing documents from dif-

ferent domain, genre, media and language. For each instantiation, Table 8.1 details the input characteristics, as well as the specific information needs.

System	Domain	Language	Media	Unit	Genre	Audience	Content	Length
	News	Spanish	Text	SDS	Journ	BackG	Generic	10%
LCsum	News	Catalan	Text	SDS	Journ	BackG	Generic	10%
(PreSeg)	Scientific	English	Speech	SDS	ScSpeech	BackG	Generic	10,30,70w
LCsum								
(DiscStr)	News	Spanish	Text	SDS	Journ	BackG	Generic	10%
LCsum	Scientific	English	Speech	SDS	ScSpeech	BackG	Generic	10,30,70w
(PostSeg)	Scientific	English	Speech	SDS	ScPres	BackG	keywordQ	100w
LCsum								
(MDS)	News	English	Text	MDS	Journ	BackG	complexQ	250w
MLsum								
(HL)	News	English	Text	SDS	Journ	BackG	Generic	10w
MLum								
(MDS,Q)	News	English	Text	MDS	Journ	BackG	complexQ	250w
QAsum								
(TALP_QA)	News	English	Text	MDS	Journ	BackG	complexQ	250w
LEXsum	News	English	Text	MDS	Journ	BackG	complexQ	250w
(JIRS)	Scientific	English	Multi	MDS	ScPres	BackG	keywordQ	100w
SEMsum	News	English	Text	MDS	Journ	BackG	complexQ	250w
(JIRS)	Scientific	English	Multi	MDS	ScPres	BackG	keywordQ	100w
SEMsum								
(JIRS,Rerank)	News	English	Text	MDS	Journ	BackG	complexQ	250w
SEMsum								
(JIRS,Update)	News	English	Text	MDS	Journ	JNews	complexQ	100w

Table 8.1: Details of several tested approaches: input, purpose and output aspects.

The first column of Table 8.1 specifies the domain: *News* or *Scientific*. The second column indicates the language of the documents to be summarized: *Catalan*, *Spanish* or *English*. The media of the input appears in the third column: *Text*, *Speech* or *Multi*, if different sort of documents are processed. In Next column, *SDS* denotes that the approach has been evaluated with one document as input and *MDS* when the input unit is a set of documents. The fifth column of Table 8.1 shows that three different genres of documents have been processed: journalistic (*Journ*) genre; and two different genres in the scientific domain: *ScSpeech*, when the input is a spontaneous speech transcript from an oral scientific presentation transcript and *ScPres*, when the input consists of a set of different kinds of documents related with a scientific presentation (transcript, slides, author notes or scientific papers). The audience, in column six, is related with the production of background (*BackG*) summaries vs the production of just-the-new information

(*JNews*). Moreover, the summary content can be *Generic* or *Query-driven*, being the query a real-world complex question (*compQ*) or a list of keywords (*keyQ*). The last column reflects that the evaluated summary length can be expressed in compression rate percentage (10 %), or in number of words (10, 30, 70, 100, 250w).

In order to analyze the performance of the developed prototypes different evaluation frameworks have been created to allow the study of their portability to different tasks. The experiments carried out show the following facts:

First, with respect to Single Document Summarization, shallow text-oriented techniques, previously evaluated for textual Spanish and Catalan documents, are robust enough to identify relevant information and basic units of content in spontaneous speech. These techniques obtain similar performance when processing manual and automatic transcripts.

Second, two approaches were evaluated in the same conditions when dealing with two different query-focused Multi-Document Summarization tasks. One approach is based on the use of lexical features, while the other adds semantic information. Their performance is analyzed in regard to two different genres: written news, obtaining state-of-art results, and scientific oral presentations, showing that the use of different types of documents outperforms the use of transcripts. It was observed that approaches using only lexical information achieve similar performance in both scenarios. The fact of adding semantic information significantly increases the performance when dealing with written news articles. However, there is room for improvement when adding semantic information when dealing with different sorts of scientific documents, most likely because the language processing tools were trained on a different domain. Moreover, oral presentation documents are less structured than formal text and less edited.

The main contributions of the reported research are outlined in the next section and in Section 8.2 some possible extensions of the work described in this thesis are presented.

8.1 Contributions

The main contributions of this thesis are related with the FEMsum framework and their specific instantiations. In addition, some relevant contributions are devoted with respect to the automatic evaluation of different summarizers or evolving prototypes of FEMsum instantiations.

FEMsum framework

- Regarding to the input, although most of the FEMsum instantiations have been evaluated for English, they can be considered partially language independent. All the presented approaches can at least summarize English and Spanish documents. Moreover, the performance of some instantiations have been evaluated when summarizing well-written or

ill-formed text from different genres and domains.

- The Linguistic Processing FEMsum component is the most dependent to the type of input. This fact facilitates re-using the rest of components when dealing with different languages or different sort of documents. Furthermore, some of them are re-used by several approaches.
- FEMsum approaches can produce different sorts of summaries depending on the summarization task (SDS vs MDS) or the user information need (very-short or short text-driven vs query-driven summaries). In the query-driven summaries the user need can be expressed by a complex NLP query or by a list of keywords.

FEMsum instantiations

- The *LCsum* has been evaluated when processing different sort of documents (journalistic well-written text or scientific ill-formed spontaneous speech transcripts) in different languages (Catalan, Spanish or English). With respect to well-written text, we show that the collaborative integration of discursive information yields an improvement on the lexical chain text representation. Moreover, text-oriented techniques are robust enough to identify relevant information and basic units of content in spontaneous speech.
- The annotations created during DUC conferences provide a valuable source of information for training automatically text summarization systems using Machine Learning techniques. A first DT sentence classifier was trained using DUC 2002 SDS data to instantiate a Content Extractor component for a Headline Extractor approach. In a similar direction, different possibilities for applying query-focused MDS DUC data in training SVMs to be used as Relevant Information Detection FEMsum component have been explored. The experiments have provided some insights on which can be the best way to exploit the annotations.
- Due to the fact that oral communication is harder to process than written text, we propose to use a multi-document summarizer capable of handling documents from different media types. In fact, combining documents from different media can help counteract not only the difficulties in the processing of oral communication, but also those errors introduced by automatic speech recognizers.
- Several FEMsum approaches were manually evaluated by NIST assessors in DUC 2006. Using semantic information to detect summary content and avoiding redundancy, *SEMsum* ranked among the best participants, obtaining an acceptable performance in content responsiveness and a good performance in non-redundancy. An improved version of the prototype evaluated in DUC 2006 was automatically evaluated in DUC 2005 data, and it was ranked among the best DUC 2005 participant systems. The *SEMsum* approach was

also evaluated in DUC 2007 producing acceptable summaries and ranking among the 10 best state-of-art systems.

Automatic evaluation

- Several summary reference corpus have been created to evaluate the FEMsum instantiations when dealing with different tasks. The HERMES corpus was created to evaluate the performance of summarizing news agency documents in different languages, Spanish and Catalan. In the framework of the CHIL project, two different kind of corpus were created. One to deal with generic summaries of scientific oral presentation transcripts and the other to deal with query-focused summarization of different sort of documents related with an oral presentation.
- To take advantage of the big quantity of human-made judgements on summary quality available for a big number of automatic extractive summaries produced by DUC assessors, we propose two methodologies for re-using this information to automatically evaluate new summarizers. The first methodology is based in unigram overlap between manual scored summaries and the summary to be scored. It was applied to DUC 2003 Task1 participating systems. In contrast, the second methodology, AutoPan, does not take into account the scores manually assigned to a summary, but the Summary Content Units manually annotated for each cluster of documents to be summarized. In this case the measure used to score new summaries is the one proposed in the Pyramid method. AutoPan obtained a good correlation with human Pyramid scores and other automatic metrics in DUC 2005 data.

8.2 Future Work

With the aim of obtaining a generic summarizer architecture, many Automatic Summarization aspects have been addressed in this thesis, most of them could be extended. Some of the extensions proposed in this section refer to the application of FEMsum to new tasks, while some others are related with the improvement of the actual instantiations. Moreover, another future line is to extend the automatic evaluation methods.

Applying FEMsum to new tasks

- Most of the FEMsum instantiations could be adapted to new languages, domains or genres. Furthermore other ill-formed input could be summarized, such as mails, OCR output, as well as automatically translated documents.

- Crosslingual summarization is another new task were FEMsum can be applied. The first step consists in processing documents in English and Spanish or English and any other language previously automatically translated to English.

Improving FEMsum component instantiations

- *Linguistic Processor*: To deal with ill-formed data, to improve the specific modules or to retrain the actual generic NLP tools, if corpora available.
- *Query Processor*: In query-focussed approaches, to add new rules to improve the way of extending the query in order to obtain better results in the *Relevant Information Detection*.
- *Relevant Information Detection*: With respect to the ML approaches, extending the feature set that characterizes a sentence, such as including new features relating sentences with semantic overlapping measures or including features from the adjacent sentences. In the SVMs based approach, automating the selection of the seed negatives would make unnecessary any human involvement in training.
- *Content Extractor (Lexical Chainer)*: To exploit other EuroWordNet relations, adding new lexical chain candidate members such as verb, and study the possibility of relating different kinds of chains. For instance taking into account the NE classification to relate common noun lexical chains with NE lexical chains or syntactic relations to relate chains of verbs with chains of nouns.
- *Content Extractor (Semantic Content Extractor)*: To exploit new lexico-semantic proximity measures between sentences to be used in the *Similarity Matrix Generator* component. Additionally, other graph-based algorithms could be instantiated and evaluated as *Candidate Selector*.
- *Summary Composer*: To refine subtasks to improve the final quality of extract based summaries: reordering, sentence compression, anaphora resolution or repetition avoidance.

Enhancing automatic evaluation methods

- To add deeper linguistic information to both proposed methods. By now, only word-form similarity has been considered when matching fragments of summaries or SCUs and fragments of summaries to be evaluated. However, breaking down each sentence into a set of minimal semantic units, such as the Basic Elements, may improve the quality of the alignment as well as the quality of the resulting scores or pyramid annotations.
- To use QARLA to try to combine different automatic evaluation measures, based in: n-gram overlap, SCUs, and lexico-semantic similarities.