Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 119-122 recibido 12-01-12 revisado 06-03-12 aceptado 18-03-12

Text Summarisation based on Human Language Technologies and its Applications

Generación de resúmenes basados en las Tecnologías del Lenguaje Humano y sus aplicaciones

Elena Lloret Pastor
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. de correos, 99
E-03080 Alicante
elloret@dlsi.ua.es

Resumen: Esta tesis doctoral se centra en la tarea de generación automática de resúmenes de textos, proponiendo la herramienta COMPENDIUM. Esta herramienta tiene en cuenta los principios cognitivos que explican cómo generan resúmenes los humanos, pero también aporta una componente computacional que permite su automatización. La evaluación de COMPENDIUM sobre diferentes géneros textuales y dominios, y su aplicación a las tareas de búsqueda de respuestas, minería de opiniones y clasificación de textos, demuestra que los resúmenes automáticos generados con COMPENDIUM son beneficiosos tanto para los usuarios, como para otras aplicaciones de basadas en Tecnologías del Lenguaje Humano.
Palabras clave: Generación de resúmenes, Tecnologías del Lenguaje Humano, Procesamiento del Lenguaje Natural

Abstract: The research work carried out in this thesis focuses on Text Summarisation, proposing and developing COMPENDIUM Text Summarisation tool. This tool takes into account the cognitive perspective, that provides insights of how humans summarise, as well as computational issues needed for its automation. For evaluating COMPENDIUM, we selected different corpora belonging to a wide range of domains and textual genres. Moreover, we also performed an extrinsic evaluation, applying COMPENDIUM to three Human Languages Technologies tasks: question answering, opinion minng and text classification. The results obtained show that the generated summaries are very appropriate both for individual users as well as for other Human Language Technologies applications.
Keywords: Text Summarisation, Human Language Technologies, Natural Language Processing

1. Introduction

In the current society, information plays a crucial role that brings competitive advantages to users, when it is managed correctly. However, due to the vast amount of available information, users cannot cope with it, and therefore new methods and approaches based on Human Language Technologies (HLT) are essential to process all the information in an effective and efficient manner. Text Summarisation (TS) is a research area in the context of HLT whose goal is to process, synthesise and present the information to users, avoiding the arduous task of having to read everything, as well as facilitating the process of guiding the user in what it is important in texts.

The main goal of this thesis is to analyse, develop and research into new techniques and approaches for Text Summarisation. Therefore, after analysing the main techniques and approaches in TS, as well as the existing evaluation methods, COMPENDIUM TS tool is proposed and developed. Our TS tool is based on a cognitive perspective (Van Dijk, 1980), (Van Dijk and Kintsch, 1983) that provides insights of how humans summarise, but in addition, it takes into account computational issues needed for its automation (Hovy, 2005).

2. Thesis Overview

This thesis is organised in several chapters, each of them focusing on a specific issue

ISSN 1135-5948

within the Text Summarisation (TS) research area.

- **Chapter 2 (Text Summarisation)** provides the state of the art in TS. This comprises the analysis of a wide range of methods and approaches for different summary types and scenarios, together with a review of existing corpora and the most relevant conferences with regard to this task.
- **Chapter 3 (Text Summarisation Evaluation)** contains the state of the art in the evaluation of summaries. It explains in detail the aspects concerning the evaluation of summaries, explaining the existing types for evaluating summaries, but focusing on the intrinsic assessment of summaries. The existing methods and tools for evaluating summaries, as well as their advantages and disadvantages are discussed. In addition, the use of crowdsourcing services in the context of TS evaluation is also explained.
- **Chapter 4 (COMPENDIUM Text Summarisation Tool)** provides the theoretical background behind the process of summarisation from a cognitive perspective. Taking this perspective as a basis, COMPENDIUM is proposed as a TS tool able to generate different types of summaries. A general overview of the suggested TS tool, as well as the stages involving COMPENDIUM are explained.
- **Chapter 5 (Evaluation and Experiments)** contains the evaluation environment, the experiments carried out, and the results obtained by COMPENDIUM. Firstly, the different methods and techniques COMPENDIUM employs are analysed. Then, the TS tool as a whole is evaluated in an intrinsic manner within different domains and types of summaries.
- **Chapter 6 (COMPENDIUM in Human Language Technology Applications)** analyses the influence of COMPENDIUM (i.e., text summaries) in other HLT applications, and serves as the extrinsic evaluation of the tool. Specifically, we applied COMPENDIUM to opinion mining, question answering and text classification.
- **Chatper 7 (Conclusion and Work in Progress)** draws the main conclusions of this research work and the main contributions of this thesis. It also addresses the research that is being conducted at the moment, as well as some issues that will be faced in the future. Finally, a list of the relevant publications is also provided.
- Finally, the **annexe "Síntesis en castellano"** provides a summary of the thesis in Spanish. This synthesis contains the main contributions and findings, as well as it explains the most relevant experiments carried out and the results obtained.

3. Main Contributions

With the background, reasoning and experiments conducted in this thesis, the proposed techniques employed in COMPENDIUM were shown appropriate for the generation of summaries. Furthermore, its assessment both intrinsically (i.e., with respect to the content and quality of the generated summaries) with different types of texts, as well as extrinsically, by integrating it into HLT applications, proved that the resulting summaries were good enough to be used on their own, as well as for being beneficial to other HLT applications.

Summing up all the research work carried out in this thesis, the main contributions were:

- **Analysis of the state of the art with respect to the approaches and methods for generating summaries**

  From the exhaustive analysis performed, we were able to come up with some insights concerning the directions of TS for the next years. This is related to the Web 2.0 and all the types of new textual genres that have appeared: reviews, forums, wikis, blogs, etc. It will be of crucial importance to study methods and approaches that are capable of providing summaries from such types of texts. Moreover, the large amount of information provided by people with different backgrounds, and in different languages will lead to the urgent need of carrying out research into abstractive summarisation, as well as multi-lingual and

cross-lingual summarisation techniques. This will allow TS approaches to manage the available information more efficiently, and enrich the resulting summaries with related information stated in different languages, formats, and with different intentions.

- **Analysis of the state of the art in the evaluation of summaries**

  Despite the large number of methods and tools for carrying out TS, at present, the evaluation of summaries is still a challenging issue and very difficult to tackle, even for humans. The inherent subjectivity associated to the TS process means that two humans can assess the same summary differently. The reason why this occurs is because each user has a specific interest or a different knowledge. In recent years, approaches have started to focus on the automatic evaluation of several aspects more related to the quality of the summary and not so much in its content.

- **Research into novel techniques and approaches for TS**

  In this thesis, we have proposed and study different novel methods for generating summaries. Concerning the redundancy problem, we proved that textual entailment is appropriate for identifying and discarding repeated information. Further on, we analysed the Code Quantity Principle (Givón, 1990) as a cognitive-based feature for detecting relevant information in documents, obtaining successful results once the topics of the documents have been identified in the first place. Finally, we analysed the use of word graphs to compress and fuse information, and we suggest a new type of summaries (abstractive-oriented) which combines extractive and abstractive information and goes beyond the simple selection of sentences.

- **Proposal and development of COMPENDIUM TS tool**

  In light of the aforementioned techniques that were shown appropriate for the generation of summaries, we developed COMPENDIUM TS tool, which relies on different stages and it is able to generate several types of summaries. The stages involved in COMPENDIUM can be grouped into two categories. On the one hand, there are a set of core stages, which constitute the central part of COMPENDIUM. These stages are: *surface linguistic analysis*; *redundancy detection*; *topic identification*; *relevance detection*; and *summary generation*, and they are mainly responsible for detecting and removing redundant information, determining the topic or topics of the document, and finally, identifying relevant information. On the other hand, there are several additional stages (*query similarity*; *subjective information detection*; and *information compression and fusion*), which specifically deal with a type of summary: query-focused; sentiment-based; and abstractive-oriented.

- **Evaluation of COMPENDIUM**

  In order to verify the appropriateness of the proposed TS method, the content and the quality of summaries generated with COMPENDIUM have to be evaluated. It has been shown that COMPENDIUM achieves competitive results with respect to the state of the art in TS. Furthermore, from the extensive evaluation carried out we can draw some interesting conclusion about its strong and weak points.

  Regarding its strengthens, COMPENDIUM generates different kinds of summaries, by employing the same core stages, and adding specific additional stages. In addition, the proposed techniques have been proven to work successful in a variety of domains and types of texts, such as newswire, blogs, image captions and medical research papers. In contrast, its main limitations involve two issues. On the one hand, multi-document summarisation has to be approached employing a different strategy, since it has been proved that the current one does not lead to very good results. On the other hand, it is crucial to incorporate semantic information in the process of TS, in order to be able to generalise and obtain new information, that can be later used for abstract generation. Finally, it is worth stressing upon the fact that, although at the moment, COMPENDIUM

is mono-lingual, it would be feasible to extent it to other languages, given that the language-specific resources involved in the process of TS are available for the target language.

- **Integration of COMPENDIUM in HLT applications: opinion mining, question answering and text classification**

  Once the summaries generated with COMPENDIUM have been evaluated, it is important to analyse to what extent it can be integrated into other HLT applications. In light of this, we showed the benefits of combining our TS tool with opinion mining, question answering and text classification.

  Regarding the first task, COMPENDIUM was used to improve the summaries generated only using an opinion mining tool, without taking into account any summarisation technique. In the second task, question answering, it was found that when using query-focused summaries generated with COMPENDIUM instead of search engine snippets in a Web-based question answering approach, its performance increased. Finally, concerning text classification, we employed summaries to filter noisy information from documents, and to be able to correctly predict a review's rating.

In conclusion, the research carried out in this thesis reveal the importance of taking into consideration different types of features and techniques derived from distinct perspectives (statistical, cognitive, linguistic)(Lloret et al., 2008), (Lloret and Palomar, 2009). The combination and integration of the proposed techniques leads to the implementation of a TS tool, COMPENDIUM, capable of producing a wide range of summaries automatically. The results obtained demonstrate that the tool is competitive and the generated summaries can be very useful for individuals as well as for applications.

4. Additional Information

Ph.D Thesis in Computer Science, specifically in the field of Computational Linguistics, written by Elena Lloret Pastor under the supervision of Dr. Manuel Palomar Sanz. The author was examined on June 20th, 2011 by a panel formed by Dr. Isidro Ramos (Univesidad Politécnica de Valencia), Dr. Paloma Moreda (Universidad de Alicante), Dr. Ruslan Mitkov (University of Wolverhampton), Dr. Pablo Gervás (Universidad Complutense de Madrid) and Dr. Rafael Muñoz (Universidad de Alicante). The grade obtained was Sobresaliente *Cum Laude por unanimidad*, with the European PhD award[1].

Acknowledgements

This research work was funded by the Spanish Government through the re- search program FPI (BES-2007-16268) associated to the project "TEXT-MESS: Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano" (TIN2006-1526-C06-01). Moreover, it was also partially funded by projects grants TIN2009-133991-C04 and PROMETEO/2009/199 from the Spanish and the Valencian Government, respectively.

References

Givón, Talmy, 1990. *Syntax: A functional-typological introduction, II.* John Benjamins.

Hovy, Eduard, 2005. *The Oxford Handbook of Computational Linguistics*, chapter Text Summarization, pages 583–598. Oxford University Press.

Lloret, Elena, Óscar Ferrández, Rafael Muñoz, and Manuel Palomar. 2008. Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. *Procesamiento del Lenguaje Natural*, (41):183–190.

Lloret, Elena and Manuel Palomar. 2009. A gradual combination of features for building automatic summarisation systems. In *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, pages 16–23.

Van Dijk, T. 1980. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Van Dijk, T. A. and W. Kintsch. 1983. *Strategies of Discourse Comprehension.* Academic Press, Inc., New York.

[1] the full text of the PhD is available at: http://gplsi.dlsi.ua.es/gplsi11/sites/default/files/elloret_tesis.pdf