Thesis Overview:

Intelligent automatic generation of text summaries with Soft Computing techniques

Augusto Villa Monte
National University of La Plata (Argentina), and University of Castilla-La Mancha (Spain)
March 2019
Advisors: Laura Lanzarini, and José A. Olivas
Co-tutorship PhD Thesis in Computer Science, and in Advanced Information Technologies[1]
{avillamonte,laural}@lidi.info.unlp.edu.ar, and JoseAngel.Olivas@uclm.es

Nowadays, the Internet is the chosen medium for disseminating information that is then used to solve a wide range of problems. However, as the amount of data stored grows, its administration becomes more difficult and users begin to suffer from the so-called information overload. Many are the sectors that, affected by this phenomenon, do not find a solution to the problem.

The use, availability and development of technology in recent decades have facilitated the collection of information and allowed the generation of large data repositories. In recent years, repositories of text documents, such as the Web for example, have gained more attention.

Given the exponential growth in the volume of textual information, it became essential to have automatic tools that, based on the original information, differentiate what is essential from what is not. Not all information has the same level of relevance. Not only in terms of content, but also in terms of interests.

To obtain text summaries automatically can constitute the solution to this problem, especially in those areas of science, such as medicine, in which research and dissemination of information are fundamental for its development.

This thesis develops two different strategies to build automatic summaries of texts using *Soft Computing* techniques. The first uses a *Particle Swarm Optimization* technique that, from the vectorial representation of the texts, constructs an extractive summary combining adequately several punctuation metrics. The second strategy is related to the study of causality inspired with the management of uncertainty by the *Fuzzy Logic*. Here, the analysis of the texts is carried out through the construction of a graph by means of which the most important causal relationships are obtained together with the temporal restrictions that affect their interpretation. Both strategies fundamentally imply the classification of the information and reduce the volume of the text considering the recipient of the summary constructed in each case.

The emphasis of this thesis lays on the combination of approaches. On the one hand, identifying the criteria that the user uses when selecting the relevant parts of a document. On the other hand, constructing a graph as from textual patterns useful in decision making. In order to carry out the case of study, several medical documents were obtained from the Internet, an area where a mobile application was developed to prevent common errors in the administration of time-dependent drugs.

This thesis is divided into five chapters and two appendices:

- Chapter 1 presents an introduction to the topic of the thesis. This initial chapter defines the main objectives, and contributions. In addition, the scientific publications supporting this work are detailed.
- Chapter 2 begins with a general description of Artificial Intelligence, Computational Linguistic, Natural Language Processing, and Text Mining. Then, a study and analysis of the text pre-processing tasks and alternatives of representation of documents are presented. The two most important types of automatic summaries are introduced, emphasising theirs differences. Towards the end of the chapter, the most commonly used scoring metrics for extractive summaries, as well as the document representation used in the next chapter, are detailed.
- Chapter 3 describes an alternative method that permits generating extractive summaries of documents by properly weighting punctuation characteristics of the sentences. The main aspect of the proposed method is the identification of those characteristics that are closest to the criteria used by the human when summarizing. This chapter begins introducing the optimization techniques, with emphasis on Particle Swarm Optimization used for the development of the proposed method.

[1] Full text available at http://sedici.unlp.edu.ar/handle/10915/74098

- Chapter 4 develops the concepts of causality and temporality, describing the reasons why it is interesting to study them. In this chapter, specific content will be extracted from medical documents. A mechanism will be provided for the extraction and representation of causal sentences affected by temporary restrictions, whose main benefited area will be medicine.
- Chapter 5 summarizes the conclusions and discusses future lines of work.
- Appendix 1 describes how the technique developed in Chapter 3 can be applied to obtain classification rules, and Appendix 2 details the process of the corpus construction used in that chapter.

The main scientific publications that support this thesis are the following:

- Augusto Villa Monte, Laura Lanzarini, Luis Rojas Flores, and José A. Olivas. Document summarization using a scoring-based representation. In *2016 XLII Latin American Computing Conference*, pages 1–7, 2016
- Cristina Puente, Augusto Villa Monte, Laura Lanzarini, Alejandro Sobrino, and José A. Olivas. Evaluation of causal sentences in automated summaries. In *2017 IEEE International Conference on Fuzzy Systems*, pages 1–6, 2017
- Laura Lanzarini, Augusto Villa Monte, Aurelio F. Bariviera, and Patricia Jimbo Santana. Simplifying credit scoring rules using LVQ+PSO. *Kybernetes*, 46(1):8–16, 2017
- Augusto Villa Monte, Laura Lanzarini, Aurelio F. Bariviera, and José A. Olivas. Obtaining and evaluation of extractive summaries from stored text documents. In *Proceedings of the Third Conference on Business Analytics in Finance and Industry*, pages 65–66, 2018
- Cristina Puente, Alejandro Sobrino, Augusto Villa Monte, and José A. Olivas. Alert system for timely medication administration. In *Proceedings of the 2018 International Conference on Artificial Intelligence (ICAI'18), located at 2018 World Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE'18)*, pages 387–392. CSREA Press, 2018
- Augusto Villa Monte, Julieta Corvi, Laura Lanzarini, Cristina Puente, Alfredo Simon Cuevas, and José A. Olivas. Text pre-processing tool to increase the exactness of experimental results in summarization solutions. In *Proceedings of the XXIV Argentine Congress of Computer Science*, 2018
- Cristina Puente, Alejandro Sobrino, José A. Olivas, and Augusto Villa Monte. Designing a system to extract and interpret timed causal sentences in medical reports. *Journal of Experimental & Theoretical Artificial Intelligence*, 31(1):1–13, 2019

This thesis has been developed following the lines of research that the Institute of Research in Computer Science LIDI (III-LIDI, Argentina) and the Soft Managemement of Internet and Learning (SMILe, Spain) research group carried out collaboratively. It had the external support of Professors PhD Cristina Puente (Comillas Pontifical University), PhD Aurelio F. Bariviera (Rovira i Virgili University), and PhD Alejandro Sobrino (University of Santiago de Compostela). It was presented by Augusto Villa Monte, in the framework of his co-tutorship thesis, as requirement to obtain the PhD degree in Computer Science by the National University of La Plata (UNLP, Argentina) and in Advanced Information Technologies by the University of Castilla-La Mancha (UCLM, Spain).

Citation: A. Villa Monte. "Thesis Overview: *Intelligent automatic generation of text summaries with Soft Computing techniques"*, Journal of Computer Science & Technology, vol. 19, no. 1, pp. 91–92, 2019.

DOI: 10.24215/16666038.19.e09