Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 341-342 recibido 03-05-2011 aceptado 24-05-2011

Text Simplification in Simplext: Making Texts more Accessible

Simplificación de textos en Simplext: haciendo textos más accesibles

Horacio Saggion*
Universitat Pompeu Fabra, Grupo TALN, horacio.saggion@upf.edu

Elena Gómez-Martínez, Esteban Etayo
Technosite-ONCE Foundation, megomez@technosite.es, eetayo@technosite.es

Alberto Anula
Universidad Autónoma de Madrid, alberto.anula@uam.es

Lorena Bourg
Ariadna Servicios Informáticos, lbourg@asi-soft.com

Resumen: El proyecto Simplext propone el desarrollo de un sistema ubicuo para la simplificación automática de textos. Simplext se basa en principios de Fácil Lectura y en la aplicación de técnicas robustas de procesamiento de lenguaje natural.
Palabras clave: Simplificación de textos; Alineación de oraciones; Generación de textos

Abstract: The Simplext project aims at producing an ubiquitous text simplification system for Spanish. The automatic simplification system is being developed following the easy-to-read principles and applying robust Natural Language Processing techiques.
Keywords: Text Simplification; Sentence Alignment; Text-to-Text Generation

1 Introduction

Text simplification is the process of transforming a text into an equivalent which is more understandable. This simplification is beneficial for many groups of readers, such as language learners, elderly persons and people with other special reading and comprehension necessities. Simplext[1] ("Un sistema automático de simplificación de textos") is a project to develop an ubiquitous text simplification solution for Spanish. In addition to its scientific interest and the fact of being the first application of text simplification to Spanish, Simplext has also an important social function since it aims at developing a solution to make text accessible to people with a cognitive impairment. This paper will give an overview of the work involved in the Simplext project.

2 Simplification Methodology

Manual simplification methodology in Simplext follows work by Anula (2007) proven to contribute to the reduction of complexity in written language. Two types of simplifications are considered here: at the level of vocabulary, simplification is based on principles such as *frequency of use* (e.g., frequent terms are preferred) and *lexical density* (e.g., repetition). However, the control of low frequency does not assure, by itself, the improvement of comprehensibility, since for instance most lexical words are polysemic. At the syntactic level, there are many elements than can and must be simplified to obtain simpler and more comprehensible discourses for people with reading difficulties. Among many others, we could mention the length of discursive segments, the abundance of subordinate structures, the use of impersonal or passive sentences, etc. From the linguistic point of view, the main aim of the project is to identify those phenomena and categories which hinder the cognitive processing

* Horacio Saggion is grateful to a fellowship from Programa Ramón y Cajal, Ministerio de Ciencia e Innovación, Spain.

[1]Simplext is led by Technosite and partially funded by the Ministry of Industry, Tourism and Trade of the Government of Spain, by means of the National Plan of Scientific Research, Development and Technological Innovation (I+D+i) (Avanza Competitiveness, with file number TSI-020302-2010-84).

 © 2011 Sociedad Española Para el Procesamiento del Lenguaje Natural

that takes place while and after reading, and which could undergo a linguistic simplification that allows us to create a formal variable accessible to the people with difficulties in reading comprehension, thus facilitating the cognitive processes to the reader.

We aim, with this methodology, at the development of a corpus of original texts and their adaptations based on the aforementioned principles to carry out research and development into text simplification in Spanish. The Simplext corpus will consist of a set of around 200 short news articles provided by one of the partners in the consortium (Servimedia) and adaptations of these articles to our target user groups.

2.1 Natural Language Processing in Simplext

There are various problems we are addressing in the project from the NLP viewpoint. First, the Simplext corpus will be aligned at the sentence level in two steps: an automatic aligment algorithm will be applied to identify relations between original and simplified sentences, and then a bi-text editor will be used to correct the alignments proposed automatically to make sure that in the final version of the corpus these are correct. The alignment algorithm has already been proposed and results are reported in (Bott and Saggion, 2011b). Second, we are developing text analysis pipelines to carry out linguistic processing of the corpus and the new documents to be simplied in application time. We base this work on various free-tools such as Freeling (Atserias et al., 2006) and GATE to implement syntactic and semantic analysis. Third, we are analysing the corpus in order to identify potential simplification operations which could be implemented (e.g. systematic operations instead of idiosyncratic ones). We have detected a set of simplification operations including: change, delete, insert, split, etc. and we are studing how to implement them (Bott and Saggion, 2011a) in a decision module and text-to-text generation component.

3 Simplext Architecture

The Simplext architecture is an event-driven and service-oriented architecture that further develops the idea of delivering the simplification service. The rationale of such approach is that a person with an adapted device (e.g., mobile phone) should be able to receive easy-to-read digital contents, such as RSS or digital press. The simplification web service is the software component that will allow to publish the simplification engine using SaaS (*Software as a Service*) as software delivery model. To develop this services layer, which will permit deliver the above mentioned functionalities, an architecture based on SOA and REST (*Representational State Transfer*) is planned. The simplification engine capabilities will be accessed using standard identifiers (URI), where each request may be made independently, atomic, stateless.

4 Conclusion and Outlook

Although the project is still in its first year, we have already collected part of corpus and developed and evaluated a robust sentence aligment algorithm for text simplification. Additionally, we have carried out a study of text simplification operations that will inform the development of the simplification solution. We are currently working on the development of the simplification solution which, given the reduced size of the corpus, will be a hybrid system combining machine learning with hand-crafted rules. Both intrinsic and extrinsic system evaluations are planned.

References

Anula, A. 2007. Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia, Man-Ki, Jy-Eun, y Macías (eds.)*, pages 45–61, Seúl, República de Corea.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library . In *Proceedings of LREC)*, Genoa, Italy. ELRA.

Bott, S. and H. Saggion. 2011a. Spanish text simplification: An exploratory study. *Revista de Procesamiento de Lenguaje Natural*.

Bott, S. and H. Saggion. 2011b. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the ACL 2011 Workshop on Monolingual Text-to-Text Generation*, Portland, Oregon, USA, June. ACL, ACL.