Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 200–218

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Simplifying words in context. Experiments with two lexical resources in Spanish☆

Horacio Saggion[a,*], Stefan Bott[b], Luz Rello[c]

[a] *Universitat Pompeu Fabra, c/ Tanger 122, Barcelona 08018, Spain*
[b] *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, 70569 Stuttgart, Germany*
[c] *HCI Institute, School of Computer Science, Carnegie Mellon University, USA*

Received 18 November 2013; received in revised form 10 July 2014; accepted 6 February 2015
Available online 14 February 2015

Abstract

In this paper we study the effect of different lexical resources for selecting synonyms and strategies for word sense disambiguation in a lexical simplification system for the Spanish language. The resources used for the experiments are the Spanish EuroWordNet, the Spanish Open Thesaurus and a combination of both. As for the synonym selection strategies, we have used both local and global contexts for word sense disambiguation. We present a novel evaluation framework in lexical simplification that takes into account the level of ambiguity of the word to be simplified. The evaluation compares various instances of the lexical simplification system, a gold standard, and a baseline. The paper presents an in-depth qualitative error analysis of the results.

Keywords: Lexical simplification; Spanish; Text simplification; Evaluation

1. Introduction

Automatic text simplification is a technology to adapt the content of a text to the specific needs of particular individuals or target populations so that the text becomes more readable and understandable for them. The adapted text will most probably suffer from information loss and a too simplistic or boring style, which is not necessarily a bad thing if the original message can in the end be transmitted to the reader. Text simplification has also been suggested as a potential pre-processing step for making texts easier to handle by generic text processors such as parsers, or to be used in specific information access tasks such as information extraction. But our research is more related to the first objective of making texts more accessible to specific users. This is certainly more challenging than the second use of simplification because the output will necessarily be evaluated with the same yardstick that human written texts are evaluated with. The interest in automatic text simplification has grown in recent years and in spite of the many approaches and techniques proposed, there is still space for improvement. The growing interest in text simplification is evidenced by the number of languages which are targeted by researchers around the globe. Simplification systems and simplification studies do exist at least for English (Chandrasekar et al., 1996; Siddharthan, 2002; Carroll et al., 1998), Brazilian Portuguese

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +34 93 542 1119; fax: +34 93 542 2517.
E-mail addresses: horacio.saggion@upf.edu (H. Saggion), stefan.bott@upf.edu (S. Bott), luz.rello@upf.edu (L. Rello).

http://dx.doi.org/10.1016/j.csl.2015.02.001
0885-2308/

(Aluísio and Gasperin, 2010), Japanese (Inui et al., 2003), French (Seretan, 2012), Italian (Dell'Orletta et al., 2011; Barlacchi and Tonelli, 2013), and Basque (Aranzabe et al., 2012). Text simplification, as a general task, is similar to other NLP tasks, such as machine translation, paraphrasing, text summarization or sentence compression. The mixed nature of the task and the specific requirements of each aspect, however, make text simplification not fully comparable to the mentioned NLP problems. While text summarization, for example, tries to select the most relevant information from input texts, text simplification rather concentrates on the elimination of superfluous details, by either deleting them or re-phrasing them in a more general way. Text summarization techniques for simplification can be found in Drndarevic and Saggion (2012a) and Stajner et al. (2013).

Our research, concerned with simplification in the Spanish language (Saggion et al., 2011), has produced a text simplification system made up of components for reducing the syntactic complexity of sentences, deleting unnecessary information, rewriting numbers, normalizing reporting verbs, and substituting difficult words by their simpler synonyms (Bott et al., 2012; Drndarevic et al., 2013; Bott and Saggion, 2014). It is this last technology, lexical simplification, which is the focus of the present work. Lexical Simplification aims at replacing difficult words with easier synonyms, while preserving the meaning of the original text segments. Lexical simplification requires the solution of at least two problems: First, the finding of a set of synonymic candidates for a given word, generally relying on a dictionary or a lexical ontology and, second, replacing the target word by a synonym which is easier to read and understand in the given context. For the first task, lexical resources such as WordNet (Miller et al., 1990) can be used. For the second task, different strategies of word sense disambiguation (WSD) and simplicity computation are requited.

Even if there is a considerable number of approaches to lexical simplification in different languages, an estimation of how different lexical resources and WSD strategies impact the task have not yet been studied. There is also no previous work which addresses the question in how far the level of ambiguity of a word influences the degree of success in automatic lexical simplification. The goal of this paper is to address these gaps presenting LexSiS (Bott et al., 2012), a system for Spanish lexical simplification which is parametrized in a way it can use different lexical resources which provide word senses and lists of synonyms. Hence, the main contributions of this paper are:

- A description of a lexical simplification procedure for the Spanish language.
- A comparison of the performance of our lexical simplification system with two different lexical resources (Open Thesaurus and EuroWordNet), in addition to a combined version of the two.
- A comparison of two different strategies for word sense disambiguation, one which only considers the local context of a target word and another which assumes that each target word has only one meaning per text and takes all local contexts for a given target into account.
- An evaluation that assesses the performance of the system depending on different levels of the ambiguity of target words.
- A quantitative and qualitative analysis of the results.

The rest of the paper is organized as follows: In Section 2 we discuss the related work and the context in which our proposal has to be seen. In Section 3 we present a corpus study which provided insights in the human production of lexical simplifications and guided the development of our system. In Section 4 we describe our system, including the alternative resources it can work with and alternative strategies to perform word sense disambiguation. Section 5 explains the evaluation framework and presents the experimental results, while Section 6 draws some conclusions on the use of different lexical resources and disambiguation methods. In Section 7 we present an in-depth error analysis, which complements the quantitative analysis. Section 8 concludes the paper with a summary of the main results and an outlook on future work.

2. Related work on lexical simplification

Lexical simplification requires, at least, two things: a way of finding synonyms (or, in some cases, hyperonyms), and a way of measuring lexical *complexity* (or simplicity). Many approaches to lexical simplification (Carroll et al., 1998; Lal and Ruger, 2002; Burstein et al., 2007) used WordNet in order to find appropriate word substitutions. Bautista et al. (2011) instead use a dictionary of synonyms. As a measure of lexical simplicity most of the cited approaches (Carroll et al., 1998; Lal and Ruger, 2002; Burstein et al., 2007) have relied on word frequency, with the exception of Bautista et al. (2011), who use word length as a predictor for lexical simplicity. Since both word frequency and word length