Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 191-194 recibido 26-03-2015 revisado 25-04-2015 aceptado 13-05-2015

A Web-based Text Simplification System for English[*]

Un Sistema de simplificación de textos *on-line* ***para el inglés***

Daniel Ferrés
Universitat Pompeu Fabra
daniel.ferres@upf.edu

Montserrat Marimon
Universitat Pompeu Fabra
montserrat.marimon@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
horacio.saggion@upf.edu

Resumen: La simplificación textual consiste en reducir la complejidad léxica y sintáctica de documentos con el fin de mejorar su legibilidad y comprensibilidad. En este trabajo se presenta una demostración de un sistema *on-line* de simplificación léxica y sintáctica de textos en inglés. Nuestro sistema es modular y adaptable, lo que lo hace adecuado para diversos tipos de usuarios.
Palabras clave: Simplificación léxica, simplificación sintáctica, *demo on-line*

Abstract: Text Simplification is the task of reducing the lexical and syntactic complexity of documents in order to improve their readability and understandability. This paper presents a web-based demonstration of a text simplification system that performs state-of-the-art lexical and syntactic simplification of English texts. The core simplification technology used for this demonstration is highly customizable making it suitable for different types of users.
Keywords: Lexical simplification, syntactic simplification, web-based demo

1 Introduction

Text Simplification (Carroll et al., 1998; Siddharthan, 2006) is the task of reducing the lexical and syntactic complexity of textual documents in order to improve their readability and understandability. This paper presents a demonstration of a text simplification system that performs (sequentially) state-of-the-art lexical and syntactic simplification of documents in English. The lexical simplifier has been developed following current robust, corpus-based approaches (Biran and Brody, 2011; Bott et al., 2012). The syntactic simplifier has been built following a linguistically motivated approach implemented as transformation rules complemented with text generation techniques. The chosen approach, which uses typed dependencies as basic representation, is based on current arguments in favor of the use of such representations in order to produce correct output (Siddharthan, 2006; Siddharthan and Angrosh, 2014). These simplifiers have been built entirely in the Java programming language, with open source software and freely available lexical resources. The system is highly configurable and adaptable and the resources used can easily be changed to meet the needs of different target groups.

2 Lexical Simplifier

Lexical simplification aims at replacing difficult words with easier synonyms, while preserving the meaning of the original text segments (Carroll et al., 1998). Our lexical simplifier combines Word Sense Disambiguation (WSD) and Lexical Simplicity measures to simplify words in context. It is composed of the following processing phases (executed sequentially): Document Analysis, Complex Words Detection, WSD, Synonyms Ranking, and Language Realization. The Document Analysis phase uses default components from the GATE system (Cunningham et al., 2002) to perform tokenization, sentence splitting, part-of-speech (PoS) tagging, lemmatization, Named Entity Recognition and Classification, and co-reference resolution. In addition, only during syntactic simplification (see below), the MATE Tools dependency parser (Bohnet, 2010) adds dependency labels to sentence tokens.

2.1 Complex Word Detection

Complex word detection is carried out to identify target words to be substituted. The procedure identifies a word as complex when

[*] This work was funded by the ABLE-TO-INCLUDE project (European Commission Competitiveness and Innovation Framework Programme under Grant Agreement No. 621055) and project SKATER-UPF-TALN (TIN2012- 38584-C06-03) from Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain.

the frequency count of the word in a given psycholinguistic database is in a range determined by two threshold values (i.e. w is complex if $min \leq w_{frequency} \leq max$). The following psycholinguistic resources can be used separately: Age-of-acquisition norms (Kuperman, Stadthagen-Gonzalez, and Brysbaert, 2012)[1], Kucera-Francis[2] (Kucera and Francis, 1967) frequency counts (extracted from the Brown Corpus). For example, words such as "hand" and "sun" have 470 and 123 counts respectively in the Kucera-Francis, whereas less common words such as "manifest" and "gastronomy" have 9 and 1 counts.

2.2 Word Sense Disambiguation

Since words can have more than one meaning and in order to select an appropriate word replacement out of a list of "synonyms", a Word Sense Disambiguation (WSD) phase is applied. The WSD algorithm used is based on the Vector Space Model (Turney and Pantel, 2010) approach for lexical semantics which has been previously used in Lexical Simplification (Biran and Brody, 2011). This algorithm uses a word vectors model derived from a large text collection from which a word vector for each word in WordNet-3.1[3] is created by collecting co-occurring word lemmas of the word in N-window contexts (only nouns, verbs, adjectives, and adverbs). Then, a common vector is computed for each of the word senses of a given target word (lemma and PoS). These word sense vectors are created by adding the vectors of all words (e.g. synonyms, hypernyms) in each sense. When a complex word is detected, the WSD algorithm computes the cosine distance between the context vector computed from the words of the complex word context (at sentence or document level) and the word vectors of each sense from the model. The word sense selected is the one with the lowest cosine distance between its word vector in the model and the context vector of the complex word in the sentence or document to simplify. Two data structures were produced following this procedure: 1) one that contains 81,242 target words and 135,769 entries, 2) another version that uses only synonyms to create the word sense vectors and has 63,649 target words and 87,792 entries.

The Simple Wikipedia was used to extract the word vectors model: the plain text of its 99,943 documents was extracted using the WikiExtractor[4] tool and Freeling 3.1 (Padró and Stanilovsky, 2012) was used to extract the lemmas and PoS tags of each word, from a 11-word window (5 words to each side of the target word).

2.3 Synonyms Ranking

The Synonyms Ranking phase tries to rank synonyms by their lexical simplicity and finds the simplest and most appropriate synonym word for the given context. The simplicity measures implemented are two: 1) only the word frequency (used by default for the simplifier) is used to rank synonyms (i.e. more frequent is simpler) (Carroll et al., 1998) and 2) a metric which combines word length and word frequency proposed by (Bott et al. 2012). Frequency lists from the following corpora can be used to rank by lexical simplicity in our system: British National Corpus (BNC), Google Web 1T Corpus most frequent words[5], Simple English Wikipedia, English Wikipedia, American National Corpus, SUBTLEX-US[6], SUBTLEX-UK[7], Kucera-Francis, and Age-of-Aquisition norms.

2.4 Language Realization

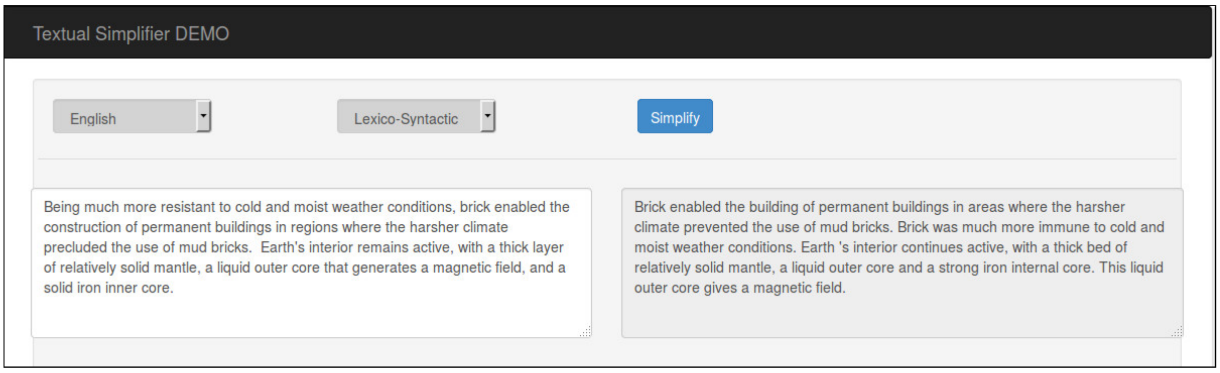
The Language Realization phase generates the correct inflected forms of the final selected synonym words. The SimpleNLG[8] (Gatt and Reiter, 2009) Java API is used to convert lemmas to their correct inflectional forms according to their context and PoS tag.

3 Syntactic Simplifier

Syntactic simplification aims at transforming long and complicated sentences into their more simpler equivalents. Similar to (Aluísio and Gasperin, 2010; Bott and Saggion, 2014), our Syntactic Simplifier is linguistically motivated. Linguistic phenomena that may complicate readability are identified and appropriate transformations to generate simpler paraphases are implemented. Our simplifier targets the following syntactic constructions: *Apposition*, *Relative Clauses*, *Coor-*

[1]http://crr.ugent.be/archives/806
[2]http://www.psych.rl.ac.uk/kf.wds
[3]http://wordnet.princeton.edu/

[4]http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
[5]http://norvig.com/ngrams/count_1w.txt
[6]http://expsy.ugent.be/subtlexus/
[7]http://crr.ugent.be/archives/1423
[8]http://code.google.com/p/simplenlg/



Being much more resistant to cold and moist weather conditions, brick enabled the construction of permanent buildings in regions where the harsher climate precluded the use of mud bricks. Earth's interior remains active, with a thick layer of relatively solid mantle, a liquid outer core that generates a magnetic field, and a solid iron inner core.

Brick enabled the building of permanent buildings in areas where the harsher climate prevented the use of mud bricks. Brick was much more immune to cold and moist weather conditions. Earth 's interior continues active, with a thick bed of relatively solid mantle, a liquid outer core and a strong iron internal core. This liquid outer core gives a magnetic field.

Figure 1: Screenshot of the web demonstration interface.

dination, *Coordinated Correlatives*, *Passive Constructions*, *Adverbial Clauses*, and *Subordinated Clauses*. After a process of document analysis which produces typed dependencies, the syntactic simplification is applied in two steps: 1) an analysis phase, that identifies the syntactic structures to be simplified and 2) a generation phase that produces correct simplified structures. The system recursively simplifies sentences until no more simplifications can be applied.

3.1 Syntactic Phenomena Identification and Analysis

Sentence analysis for simplification is implemented with GATE JAPE (Java Annotation Patterns Engine) grammars which detect and label the different kind of syntactic phenomena appearing in the sentences. For each of the above syntactic phenomena, a JAPE grammar contains several rules. Each rule contains a left-hand-side (LHS), which consists of an annotation pattern description, and the right-hand-side (RHS), which consists of annotation manipulation statements to produce rich simplification specific linguistic information useful for generation. These rules mainly rely on dependency information, which allows for a broad coverage of common syntactic phenomena. For example, the grammar for appositive phrases has a unique rule that identifies the apposition and its anchor using PoS and dependency labels. The LHS identifies the apposition's head by PoS and syntactic features (any common and proper noun and cardinal number which has the func(tion) appo(sition)). The RHS, first, finds out the head of the anchor (the token whose id unifies with the dependency of the apposition's head), and, then, it selects all the dependents of both the anchor's head and the apposition's head (and, recursively, the dependents of their dependents), and adds the annotations to the identified patterns. In addition, there are 17 rules for relative clauses (restrictive or non-restrictive). There are also 10 rules for coordination which deal with binary and three-conjunct coordination of sentences and VPs and 4 rules for coordinated correlatives which are distinguished by the endorsing item and the coordinator. Eight rules cover subordinated clauses expressing concession, cause, and time, both preceding and following the main clause, and 12 rules that deal with single adverbial clauses and up to three coordinated adverbial clauses, also preceding and following the modified clause. Finally, 14 rules cover passive constructions.

3.2 Sentence Generation

The generation phase uses the information provided by the analysis stage (i.e. precise annotations) to generate simple sentences. It applies a set of annotation manipulations which are specific for each phenomenon identified during analysis. These rules perform the common simplification operations, namely sentence splitting, reordering sentences, creation of new phrases, verbal tense adaption, personal pronouns transformation, capitalization and de-capitalization of some words, and word substitution.

4 Evaluation

We performed manual evaluation carried out by eight human judges, using the evaluation set used by Siddharthan and Angrosh (2014) from which we randomly selected 25 sentences. The judges assessed our system w.r.t. fluency, adequacy, and simplicity, with a 5 point rating scale and assigned a mean score of 3.98, 4.02, and 2.86, respectivelly.

5 Web Demonstration

A web demonstration of the system can be accessed and tested in the following web address: *193.145.50.158/simplifier*. The web interface is shown in Figure 1 with an example of 2 complex sentences being simplified. The visual interface has two textual areas: one of them enabled for entering the text to be simplifier (with a white color background) and another one active only to see the output of the textual simplifier (with a grey color background). There are two selection forms that allow to change the language of simplification (currently only English) and select the type of simplification. The following types of simplifications are allowed: 1) Lexical, 2) Syntactic, and 3) Lexico-Syntactic. An execution button (with the "Simplify" label) performs the delivery of the parameters and the textual input to a back-end that performs the simplification. In the example in Figure 1 the lexical simplifier replaced words such as "construction" by "building", "inner" by "interior", etc. The syntactic simplifier transformed sentences containing subordinate and relative clauses into simpler paraphrases. The back-end of the demonstration has the following configuration options selected for the lexical simplifier: 1) the complex word detector uses the *Age Of Acquisition* norms that are complex for an age of acquisition of 7 years-old or less, 2) the WSD phase uses word vectors derived from contexts of the Simple Wikipedia and a dictionary of target words and senses derived from Wordnet 3.1 (version with only synonyms), 3) the Synonyms Ranker uses frequencies extracted from the BNC corpus.

References

Aluísio, S.M. and C. Gasperin. 2010. Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proceedings of NAACL HLT 2010 YIWCALA*.

Biran, O. and N. Brody, S.and Elhadad. 2011. Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of NAACL HLT 2011*.

Bohnet, B. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING 2010*.

Bott, S., L. Rello, B. Drndarevic, and H. Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING 2012*.

Bott, Stefan and Horacio Saggion. 2014. Text simplification resources for Spanish. *Language Resources and Evaluation*, 48(1):93–120.

Carroll, J., G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI98 Workshop on Integrating AI and Assistive Technology*.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of ACL 2002*.

Gatt, A. and E. Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of ENLG 2009*.

Kucera, H. and W. N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Kuperman, V., H. Stadthagen-Gonzalez, and M. Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of LREC 2012*.

Siddharthan, A. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Siddharthan, A. and M. Angrosh. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of EACL 2014*.

Turney, P. D. and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188.