

## Lexical Simplification

Jan De Belder · Koen Deschacht ·  
Marie-Francine Moens

### 1 Introduction

In this paper we present a generic approach to lexical simplification, that is easily portable to other languages than the ones studied here. Lexical simplification helps children, illiterate, foreign, and disabled people to read texts, by replacing difficult words with words that are easier to understand. Although syntactic simplification has received a lot of attention ([9], [2], [1], [3], [11]), the work on lexical simplification has been limited. The methods described in this paper are in the context of the EU-7 project PuppyIR, which focuses on information sources for children. In the next section we will briefly discuss previous work, in section 3 we define our method, and section 4 ends with conclusions and indications for future work.

### 2 Previous work

To the best of our knowledge the method proposed in [5], and used in [2] and [8], is the only available one for lexical simplification. This method was developed in the context of helping patients with aphasia. It simplifies a text on a word by word basis, by first generating a list of synonyms using WordNet, and then selecting the one with the highest Kucera-Francis frequency as obtained from the Oxford Psycholinguistic Database [10]. Word Sense Disambiguation (WSD) is not performed, as the author believes that less frequent words only have one specific meaning. This method also relies on the availability of WordNet and a psycholinguistic database, means that are not available for every language.

---

Department of Computer Science  
Katholieke Universiteit Leuven  
Celestijnenlaan 200A  
Leuven, Belgium  
E-mail: {jan.debelder, koen.deschacht, sien.moens}@cs.kuleuven.be

### 3 Our method

The method proposed in this paper is based on a similar idea as in [5], but is much more scalable to other languages and performs a form of WSD. An outline of the method is given in figure 1. Given a word, we first generate two sets of alternative words. One set is obtained from a dictionary with synonyms (or WordNet, if available), and the other set is generated by the Latent Words Language Model discussed in section 3.1. For each word in the intersection of these sets we generate a probability that it is a good replacement, as defined in section 3.2 by  $P_{simplification}$ .

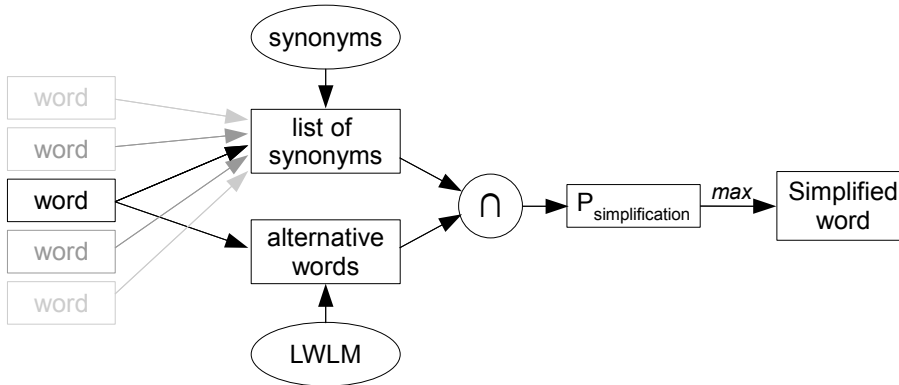


Fig. 1 Schematic representation of our method

#### 3.1 Latent Words Language Model

The Latent Words Language Model models both language in terms of consecutive words and the contextual meaning of the words as latent variables in a Bayesian network. In a training phase the model learns for every word a probabilistic set of synonyms and related words (i.e. the latent words) from a large, unlabeled training corpus. During the inference phase the model is applied to a previously unseen text and estimates for every word the synonyms for this word that are relevant in this particular context. The latent words help to solve the sparsity problem encountered with traditional n-gram models, leading to a higher quality language model, in terms of perplexity reduction on previously unseen texts [4].

A problem that occurs in the application targeted in this paper, is that the LWLM can also suggest antonyms as latent words, which is why we also rely on the list of synonyms. The added value of the LWLM can be found the fact that it offers a simple form of WSD. Given the context, it will only generate alternative words that fit in this context. This is motivated further by the observation that words tend to have one meaning in a specific context [12]. This also alleviates the problem observed in [8], where informal trials have shown that strange sounding text can be produced.

### 3.2 Modeling the easiness of words

The probability that a new word  $w$  is a good replacement for the original word  $w_{orig}$  in the text is modeled by the probability  $P_{simplification}$ , defined as follows:

$$P_{simplification}(w|w_{orig}) = P_{replace}(w|w_{orig}, context) \cdot P(easy|w) \quad (1)$$

The probability that a new word still fits the context is modeled through the LWLM, that tells us which replacements are more likely than others. The second factor of equation 1 estimates whether a word is easy or not. It can be instantiated in several ways, depending on the availability of resources. In the remainder of this section we give some possibilities:

**Psycholinguistic database:** This recourse was previously used in [6]. For research that focuses on children, using the age of acquisition rating rather than the Kucera-Francis frequency, offers a way to better simplify for a specific age. These metrics can be mapped to a [0,1] interval to obtain a probability.

**Unigram probability in easy text corpus:** If a large corpus of text, written in simple language, is available, unigram probabilities can be used. An example of such a corpus for English is the Simple English Wikipedia<sup>1</sup>. For Dutch we can rely on the news articles of Wablieft<sup>2</sup>, a newspaper written in simpler language, which is freely available for educational purposes.

**Number of syllables:** The average number of syllables and the average sentence length have been used to determine the reading difficulty of a text since the 1970's [7]. Hence, the number of syllables in a word should also give an indication of how difficult a word is.

## 4 Conclusion and future work

In this paper we have described a method for lexical simplification of text, that relies on a dictionary of synonyms and our Latent Words Language Model. We described several ways to calculate how 'easy' words in the generated list of alternative words are, taking into account the resources available for a language. Hence this method is easily portable to other languages.

Further research includes the evaluation of these method, which will be performed in a larger framework where we will also include syntactic simplifications. We will also investigate how to handle compound words, which are frequent in the Dutch language.

**Acknowledgements** This research is funded by the EU project *PuppyIR*<sup>3</sup> (EU FP7 231507), and the Dutch-Flemish NTU/STEVIN project *DAISY*<sup>4</sup> (ST 07 015).

## References

1. Candido Jr, A., Maziero, E., Gasperin, C., Pardo, T., Specia, L., Aluisio, S.: Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian

---

<sup>1</sup> <http://simple.wikipedia.org/>

<sup>2</sup> <http://bop.vgc.be/tijdschriften/wablieft/>

<sup>3</sup> <http://www.puppyir.eu>

<sup>4</sup> <http://www.cs.kuleuven.be/liir/projects/daisy/>

- Portuguese. In: Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 34–42. Association for Computational Linguistics (2009)
2. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In: Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, pp. 7–10. Citeseer (1998)
  3. Chandrasekar, R., Srinivas, B.: Automatic induction of rules for text simplification. *Knowledge Based Systems* **10**(3), 183–190 (1997)
  4. Deschacht, K., Moens, M.F.: The Latent Words Language Model. In: Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning (2009)
  5. Devlin, S.: Simplifying natural language for aphasic readers. Ph.D. thesis (1999)
  6. Devlin, S., Tait, J.: The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases* pp. 161–173 (1998)
  7. Kincaid, J., Fishburne, R., P, J., Rogers, R., Chissom, B.: Derivation of New Readability Formulas for Navy Enlisted Personnel. (1975)
  8. Lal, P., Ruger, S.: Extract-based summarization with simplification. In: DUC 2002: Workshop on Text Summarization, July 11–12, 2002, Philadelphia, PA, USA (2002)
  9. Petersen, S.: Natural language processing tools for reading level assessment and text simplification for bilingual education. Ph.D. thesis, University of Washington (2007)
  10. Quinlan, P.: The Oxford psycholinguistic database. Oxford University Press Oxford (1992)
  11. Siddharthan, A.: Syntactic simplification and text cohesion. *Research on Language & Computation* **4**(1), 77–109 (2006)
  12. Yarowsky, D.: One sense per collocation. In: Proceedings of the workshop on Human Language Technology, p. 271. Association for Computational Linguistics (1993)