Generating Synonyms Using Word Vectors and an Easy-to-Read Corpus

Linnea Fornander, Marc Friberg, Vida Johansson, Viktor Lind Hård, Pontus Ohlsson, Ida Palm

Linköping University
581 83 LINKÖPING
`linfo796@student.liu.se`, `marfr775@student.liu.se`, `vidjo788@student.liu.se`
`vikha664@student.liu.se`, `ponoh848@student.liu.se`, `idapa493@student.liu.se`

Abstract

With increasing amounts of information, the need for effective text simplification methods arises. One possible step in simplifying texts is synonym replacement. Previous work has evaluated several different methods of solving the problem, including the use of bilingual dictionaries and distributional similarity between words. Many of them have shown promising results. However, using only one of them has often been insufficient. This project approaches the problem by presenting two methods using a combination of bilingual dictionaries and vector representations of words. The results show that the methods used are capable of generating words that to a high rate are perceived as synonymous to the original word.

1. Introduction

As the amount of information increases, the need for simplified texts increases as well, with public authorities needing effective ways of simplifying texts in order to reach the entire public. This, in turn, causes the need to automate the process of simplifying texts. One part of the problem in this process is to develop a method for replacing words that are difficult to comprehend with easier ones. The goal of this project was to develop and evaluate two different methods that use word vectors, bilingual dictionaries and three corpora, including a corpus containing easy-to-read texts, to extract comprehensible synonyms to input words.

2. Theory

Part of what makes a text difficult is the words it consists of. Lexical simplification, to simplify a text by substituting words that are difficult to comprehend for easier ones, has shown to improve comprehensiblility for people with dyslexia and second language learners (Rello et al., 2013; Gardner and Hansen, 2007). However, replacing words that are difficult to comprehend, not altering the content, is a difficult task, as it requires synonyms to the original words.

Some methods used to extract semantically related words make use of distributional theory, which is based on the idea that words with similar meaning appear in similar contexts in texts (Harris, 1970). One group of models that has been developed using distributional theory is word2vec (Mikolov et al., 2013). Word2vec produces a set of vectors for each word, given a corpus, and calculates the cosine value between the word vectors. This can be used to point out and predict similarities between words. However, like other distributional models, word2vec does not differentiate between different distributional similarities, such as antonyms and synonyms (Deeplearning4j Development Team, 2016). To refine the results, additional methods can be used, such as bilingual dictionaries, parallel corpora, semantic mirroring, and crowdsourcing (Lin et al., 2003; Kann and Rosell, 2005; Dyvik, 2004; Priss and Old, 2005).

To evaluate the synonymy of words, several methods have been developed, whereof some are computational evaluation methods (Wu and Zhou, 2003; van der Plas and Tiedemann, 2006). Other methods use crowdsourcing, and make use of human layman knowledge (Buhrmester et al., 2011; Kann and Rosell, 2005; Mohammad and Turney, 2013).

3. Method

In this project, the line between words that are “difficult” and “easy” to comprehend was drawn with help from the corpus LäSBarT (Mühlenbock, 2013), which consists of easy-to-read texts. Included in the training data were also the Stockholm-Umeå-Corpus (SUC) (Källgren, 2007) and the Swedish Wikipedia Corpus (Denoyer and Gallinari, 2006), which resulted in a variety of texts in regular written Swedish.

One step in the methods was to train a word2vec model on the corpora. By computing the cosine value between words in the vector space, semantically related words to one specific word were found. Another step in the methods was to translate relevant words between Swedish and English, using a webcrawler and the online dictionary *bab.la*[1]. Both methods started with a selected input word and finished by returning a suggested synonym to the input word as output.

3.1 Find Closest Overlap method (FCO)

The hypothesis behind this method was that two semantically similar words, for which the translations found in a dictionary overlaps, are likely to be perceived as synonyms. The hypothesis was based on previous research, which showed one way to, based on this assumption, cancel out other semantically related words, such as antonyms (Lin et al., 2003). This method will be referred to as the FCO method.

3.2 Semantic Similarity in Translations method (SST)

The hypothesis behind this method was that translating a word to an other language and then translate the translations back would result in a list of synonyms candidates, whereof the most semantically similar to the input word is the best

---

[1] sv.bab.la

synonym. The hypothesis was based on previously used methods to extract synonyms (Kann and Rosell, 2005), aswell as the theory behind semantic mirroring (Dyvik, 2004). In this paper, this method will be referred to as the SST method.

3.3 Evaluation

The evaluation method used was an online survey. The participants were asked if the output word from the methods was synonymous to the input word, in questions of the format “Is the word X a synonym to the word Y?”. The answering selections were “Disagree”, “Doubtful”, “Sometimes”, “Totally agree”, and “I do not understand the word/words”, similar to the ones used by Kann and Rosell (2005). The survey comprised a total of 45 word pairs, whereof 15 was extracted by the FCO method and 15 by the SST method. The remaining 15, here referred to as the overlap, were pairs where both methods suggested the same synonym, given an input word. Data was gathered from 93 participants.

4. Results

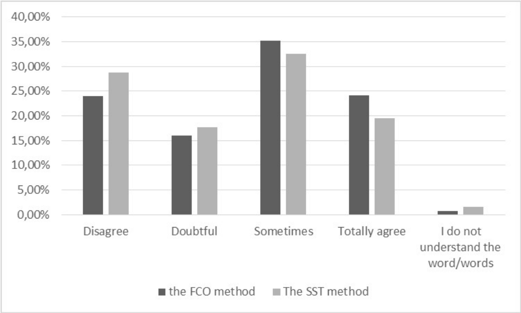

Figure 1: Average user grading per method.

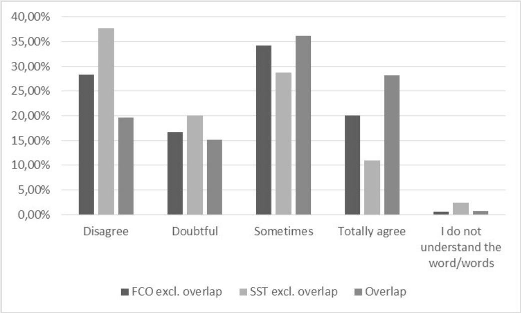

Figure 2: Average grading, separated from the overlap.

The results for the FCO and the SST methods are presented in Figure 1, showing small differences between them. The FCO method received slightly better scores of synonymy than the SST method, with 24.12% “Totally agree” and 35.20% “Sometimes”, compared to the 19.57% “Totally agree” and 32.51% “Sometimes” the SST method received. The percentage of “Disagree” is slightly lower for the FCO method than for the SST method, with 24.01% in contrast to 28.71%. The overlap received the better results than either method alone. The percentage of “Totally agree” was for the overlap 28.17%, compared to 20.07% for the FCO method and 10.97% for the SST method, as seen in Figure 2.

5. Discussion

The results showed that the answering alternative “Sometimes” was the most commonly used answer, which is deemed as a positive result since exact synonyms are rare and many words are only deemed synonymous in certain contexts.

When comparing the two methods, the FCO method proved to have slightly better results, with a higher percentage of the participants answering “Totally agree” or “Sometimes” for the suggested synonyms, than for the SST method, indicating that it might have higher precision. However, the FCO method did not generate potential synonyms for as many input words as the SST method, indicating that it might have lower recall. The methods also proved to achieve the highest rates of “Totally agree” and “Sometimes” when the output of both methods overlapped, indicating that a combination of them could receive further improved results of extracting synonymous words.

The evaluation method used served to grade at what rate two words were perceived as synonyms, something that could be expanded upon to create large sets of graded synonyms. Such an expansion could be utilized to help users pick from a list of synonyms based on their grade, making the decision easier and more well-reasoned.

The methods do not take the sentence that the input word appears in into consideration, which could mean that the potential synonyms generated are only good synonyms in some contexts. One use of our program could be to present lists of synonym candidates for words that might need replacement. That way, the writer or anyone correcting the text could handle erroneous suggestions.

Some words, generated by the methods, were the same as the input words, only differing in inflections. These types of output resulted in about 80% of the participants selecting negative answers, decreasing the total result of the methods. Word pairs consisting in words of different parts of speech may also have influenced the result. For one such pair, around 98% of the participants selected negative answers, with a majority of “Disagree”.

Based on the results, one suggestion of improvement would be to lemmatize the corpora and input words. That would prevent words of different inflectional form from the input word from being generated. It would also prevent translated words from returning without a value, due to inflections. Another suggestion is including the part of speech tags in the corpora, prior to training the word2vec model, and also to the input words. This would make it possible to skip synonym candidates of a different part of speech than the input word, supposedly giving a better synonym suggestion as output. Another advantage of using part of speech-tags is that it would make it possible to separate homonymous words that belong to different parts of speech.

6. Conclusions

This study has showed that the use of word vectors is an applicable method of extracting synonyms when used in combination with other methods, such as utilizing a bilingual dictionary. By also including an easy-to-read corpus our methods showed to be feasible when acquiring synonyms that are presumably easy to comprehend. The results showed that the methods often succeeded in generating semantically similar words to the original words, with participants deeming that the suggested words were totally or sometimes synonymous to the original words.

For improvement and future research some suggestions would be to lemmatize the corpora and the input words, along with including part of speech tags. Another suggestion is to use a larger corpus or set of corpora with easy-to-read texts, making it possible to find a wider range of synonym candidates to more input words.

The question of evaluating whether the assumption that the generated words are easy to comprehend holds, possibly by letting people with reading disabilities or second language learners evaluate the words, is left for further research.

Acknowledgements

We would like to thank our supervisors Evelina Rennes and Mikael Svensson. We would also like to thank Arne Jönsson, Marco Kuhlmann, and all the participants in our study.

References

M. Buhrmester, T. Kwang, and S.D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5.

Deeplearning4j Development Team. 2016. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0.

L. Denoyer and P. Gallinari. 2006. The Wikipedia XML Corpus. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 12–19. Springer Berlin Heidelberg, Berlin, Heidelberg.

H. Dyvik. 2004. Translations as Semantic Mirrors: From Parallel Corpus to Wordnet. *Language and Computers*, 49(1):311–326.

D. Gardner and E.C. Hansen. 2007. Effects of Lexical Simplification During Unaided Reading of English Informational Texts. *TESL Reporter*, 40(2):27–59.

Z. Harris. 1970. Distributional structure. In *Papers in structural and transformational Linguistics*, pages 775–794. Springer Science+Business Media B.V., Dordrecht.

G. Källgren. 2007. Documentation of the Stockholm UmeåCorpus.

V. Kann and M. Rosell. 2005. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th NODALIDA conference*, pages 105–110, Stockholm.

D. Lin, S. Zhao, L. Qin, and M. Zhou. 2003. Identifying synonyms among distributionally similar words. In *IJCAI'03 Proceedings of the 18th international joint conference on Artificial intelligence*, pages 1492–1493, San Fransisco. Morgan Kaufmann Publishers Inc.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

S.M. Mohammad and P.D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

K. Mühlenbock. 2013. *I see what you mean - Assessing readability for specific target groups*. Doctoral thesis, University of Gothenburg. Faculty of Arts.

U. Priss and L.J. Old, 2005. *Formal Concept Analysis: Third International Conference, ICFCA 2005, Lens, France, February 14-18, 2005. Proceedings*, chapter Conceptual Exploration of Semantic Mirrors, pages 21–32. Springer Berlin Heidelberg, Berlin, Heidelberg.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere, and Horacio Saggion. 2013. Frequent Words Improve Readability and Short Words Improve Understandability for People with Dyslexia. In *Human-Computer Interaction, INTERACT 2013 - 14th IFIP TC 13 International Conference*, pages 203–219, Cape Town.

L. van der Plas and J. Tiedemann. 2006. Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. In *Proceedings of the CO/ACL 2006 Main Conference Poster Sessions*, pages 866–873, Sydney.

H. Wu and M. Zhou. 2003. Optimizing Synonym Extraction Using Monolingual and Bilingual Resources. In *Paraphrase '03: Proceedings of the second international workshop on Paraphrasing*, volume 16, pages 72–79, Stroudsburg. Association for Computional Lingusitics.