Automatic Extraction of Synonyms from an Easy-to-Read Corpus

Vida Johansson, Evelina Rennes

SICS East Swedish ICT AB, Linköping, Sweden
Department of Computer and Information Science, Linköping University, Linköping, Sweden
vidjo788@student.liu.se, evelina.rennes@liu.se

Abstract

This paper presents two novel methods aimed at extracting comprehensible synonyms to regular words, as a step towards automatic text simplification. The methods are inspired by previous work on synonym extraction and make use of bilingual dictionaries, as well as distributional semantic similarity between words. Human layman knowledge was used to evaluate the synonym candidates extracted by the methods, which proved to be perceived as synonymous to a high degree. Both methods showed promising results and qualities, and have potential to be used to further improve automatic text simplification.

1. Introduction

The need for simplified texts increases with the amount of information, and public authorities need effective ways of simplifying texts in order increase digital inclusion. This, in turn, motivates the need to automate the process of simplifying texts. One part of the problem in this process is to develop a method for substituting words that are difficult to comprehend with easier synonyms, a process known as lexical simplification.

While the difficulty of a text depends partly on the words it consists of, the difficulty of a word depends mainly on how familiar it is to the reader (Anderson and Freebody, 1979). Lexical simplification has shown to improve both the ability to read and understand texts for people with dyslexia (Rello et al., 2013) and for second language learners (Gardner and Hansen, 2007). However, to lexically simplify a text is a difficult task, as the substitutions need to preserve the original words' semantic meaning and grammatical form. Words may also be synonymous only in specific contexts, which further increases the difficulty of the task.

Some methods used to extract semantically related words are based on the distributional hypothesis. That is, words that appear in similar contexts often have similar meanings (Harris, 1970). This group of models are known as distributional semantic models (DSMs). While the original DSMs built vectors based on values derived from event frequencies, another type of DSMs has later been developed, tackling vector estimation as a supervised task, aiming to predict a term given a context or a context given a term (Baroni et al., 2014). One group of such DSMs is word2vec, with its two approaches CBOW and skip-gram (Mikolov et al., 2013b; Mikolov et al., 2013a). Like other DSMs, word2vec models are not perfect when it comes to differentiating between distributional similarities (e.g. antonyms and synonyms). To refine the models, additional methods, such as bilingual dictionaries, parallell corpora, semantic mirroring, and crowdsourcing, can be used (Lin et al., 2003; Kann and Rosell, 2005).

To evaluate the synonymy of word pairs, several methods have been developed. Some are computational evaluation methods (van der Plas and Tiedemann, 2006; Wu and Zhou, 2003), while other use crowdsourcing to make use of human layman knowledge (Kann and Rosell, 2005).

Several studies have presented methods that can be used to extract synonyms, or to choose a synonym to a word given a few candidates (as in the TOEFL test). Previous work on lexical simplification for Swedish (Keskisärkkä and Jönsson, 2013) and for Swedish medical texts (Abrahamsson et al., 2014) present methods for synonym replacement. However, to the authors' best knowledge, none of them have been aimed at extracting comprehensible synonyms using DSMs and bilingual resources.

In this paper, we present results from using two novel methods to automatically extract Swedish synonyms from a corpus of easy-to-read texts.

2. Method

The two methods made use of DSMs that were trained using the word2vec toolkit[1]. Only the CBOW approach of Mikolov et al. (2013a) was used, as CBOW models have shown to achieve the best results in previous studies and to be robust regarding parameter settings (Baroni et al., 2014; Mandera et al., 2017). Semantic similarity between words was measured using the cosine value of their vectors. The training data consisted of the three corpora LäSBarT (Mühlenbock, 2013), the Stockholm-Umeå-Corpus (SUC) (Ejerhed et al., 2006), and the Swedish Wikipedia Corpus (Denoyer and Gallinari, 2006). All words contained in LäSBarT were, in this study, considered easy to comprehend, as the corpus consists only of easy-to-read texts. The other two corpora represent regular written Swedish. The freely available Swedish-English version of the online dictionary *bab.la*[2] was used in both methods. For more details on the methods, see Fornander et al. (2016).

Figure 1 depicts an overview of **Method 1**. The method aimed to find the most semantically similar, comprehensible, word to the input word that also shared at least one English translation with the input word. Method 1 was inspired by Lin et al. (2003), who showed that one way to handle semantic relations other than synonymy is to compare the translations of words. Words with overlapping

[1] https://code.google.com/p/word2vec/
[2] http://bab.la/

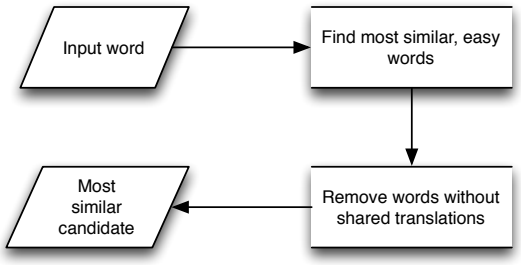

Figure 1: A simplified flowchart over Method 1.

translations are likely to be synonyms. Method 1 started by finding the 40 most semantically similar words to the input word that also occur in the LäSBarT corpus (leaving only comprehensible words). They were then translated to English, and words that did not share any translation with the input word were filtered out. Finally, the most semantically similar word of the remaining words was selected.

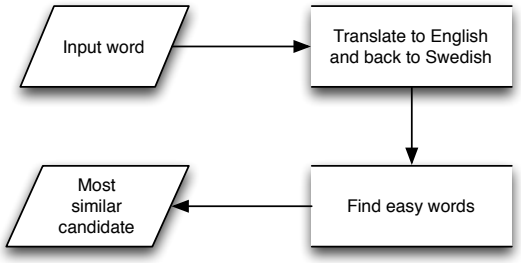

Figure 2: A simplified flowchart over Method 2.

An overview of **Method 2** is displayed in Figure 2. The method aimed to find the most semantically similar, comprehensible, word to the input word out of a list of synonym candidates. Method 2 took inspiration from the methods used to extract synonyms by Kann and Rosell (2005). By translating the input word to English, and all of the English translations of the input word back to Swedish, synonym candidates were gathered. Candidates not occurring in the LäSBarT corpus were discarded, again leaving only comprehensible words. Candidates that were the same word as the input word were discarded as well. Finally, the most semantically similar word to the input word out of the remaining candidates was selected.

The word pairs extracted by Method 1 and 2 were evaluated using an online survey. The question posed to the participants was: "Is the word X a synonym to the word Y?" and the possible answers were "I do not understand the word/words" (0) , "Disagree" (1), "Doubtful" (2), "Sometimes" (3), and "Totally agree" (4), similar to the ones used by Kann and Rosell (2005). The survey comprised 45 input words, paired with synonym candidates extracted using the methods. 30 candidates were extracted using Method 1 and 30 using Method 2. However, to 15 of the input words used in the evaluation, both methods proposed the same synonym candidate. That is, there was an intersection of the methods, which gives a total of 45 unique word pairs (and thus 45 questions). Data were gathered from 99 participants.

3. Results

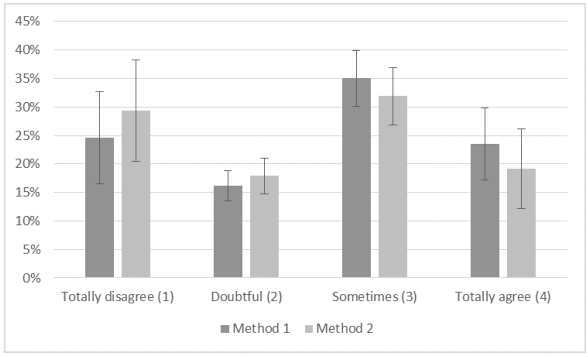

Figure 3: Frequencies of grades for word pairs extracted by Method 1 and Method 2, with 95% confidence intervals. Frequencies for word pairs contained in the intersection of the methods are included in the results.

Figure 3 shows frequencies of grades, ranging from "Disagree" (1) to "Totally agree" (4), for word pairs extracted by Method 1 and Method 2. Word pairs contained in the intersection of the methods are included in the results for both methods in the chart. Participants answered "Disagree" (1) and "Doubtful" (2) more frequently for word pairs extracted by Method 2 than for word pairs extracted by Method 1. "Sometimes" (3) and "Totally agree" (4) were more frequent for word pairs extracted by Method 1 than Method 2. The grade "Sometimes" (3) was most common for both methods.

On average, Method 1 (M = 2.58, SE = .04) performed better than Method 2 (M = 2.41, SE = .04), $t(98) = 10.90, p < .001, r = .90$.

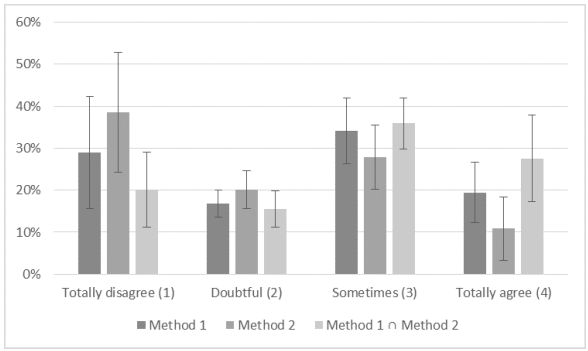

Figure 4: Frequencies of grades of word pairs extracted by Method 1 and Method 2, with 95% confidence intervals. Frequencies for word pairs contained in the intersection of the methods are shown separately.

In some cases, the same synonym candidate was generated by both methods. To better compare the methods, only word pairs containing input words to which both methods suggested a synonym candidate were evaluated. In Figure 4, the intersection is excluded from both Method 1 and Method 2 and is shown separately. The most common

grade was "Sometimes" (3) for Method 1 and the intersection, but "Disagree" (1) for Method 2.

In order to further compare the methods, the intersection was treated as a separate method in a repeated measures ANOVA. The statistical analysis determined that the methods differed significantly in performance, $F(2, 196) = 197.41, p < .001$. Post hoc tests using the Bonferroni correction, presented in Table 1, revealed that Method 1 performed significantly better than Method 2, and that the intersection between Method 1 and Method 2 performed significantly better than both Method 1 and Method 2.

Table 1: Pairwise comparisons from the Bonferroni post-hoc test.

Pairwise	Comparisons	*M*. Diff.	Sig.
Method 1	Method 2	.166	.000
	Method 1 ∩ Method 2	-.136	.000
Method 2	Method 1	-.166	.000
	Method 1 ∩ Method 2	-.302	.000
Method 1 ∩ Method 2	Method 1	.136	.000
	Method 2	.302	.000

The grade "I do not understand the word/words" (0) was excluded from the figures and the data analysis, as it made up a very small part of the grades (0.61-2.42%) and did not contribute to the understanding of to what degree word pairs are perceived as synonymous.

4. Discussion

The goal of this paper was to describe and evaluate two different methods to extract comprehensible synonyms. The results showed that "Sometimes" (3) was the most frequent grade for both methods, as shown in Figure 3. Considering that words are often synonymous in some contexts and not all contexts, this should be considered a good result.

There was a positive significant difference between the frequencies of grades when word pairs extracted by Method 1 were compared to those extracted by Method 2. There was, however, a big overlap between the synonyms produced by the two methods, and this intersection proved to score significantly higher than both Method 1 and Method 2. This indicates that it could be beneficial to combine methods for synonym extraction in order to achieve good synonyms.

Method 1 was more often the method that determined whether an input word was to be discarded or not. The intersection comprises words generated by both methods, but the input words in the other word pairs might have been were better suited for Method 1 than Method 2. Thus, it is possible that the methods extract synonym candidates equally well, but that they are better suited for different input words.

As previously mentioned, only input words to which both methods suggested a synonym candidate were evaluated. One of the first steps in Method 1 was to select the 40 most semantically similar words to the input word, which in many cases resulted in few words that occur in the LäSBarT corpus and hence decreased the method's ability to fulfill its process. In cases where Method 1 could not find a synonym candidate, words that had been extracted by Method 2 were discarded. This indicates that the process inspired by Kann and Rosell (2005), to first translate the input word in order to find synonym candidates, was better suited when a large number of words are to be filtered out. On the other hand, the results showed that word pairs extracted by Method 2 received significantly lower grades than those extracted by either Method 1 or both methods (the intersection). This indicates that the process inspired by Lin et al. (2003), in which semantically similar words were found before they were translated, resulted in better synonym candidates.

To be able to find synonym candidates to a larger number of input words is a valuable characteristic of Method 2. It could be used to extract and present synonym suggestions to users, who can filter out any erroneous suggestions themselves. It could also be used as a back-up method for when the primary method does not manage to find a synonym candidate. However, synonym suggestions to a larger number of input words are not helpful unless they are good enough.

If a threshold for cosine values would be set, Method 2 would likely benefit from it the most. Fewer words would be accepted as synonym candidates by the method, but fewer erroneous synonym suggestions would also be made. This would affect the selection of word pairs to be included in the evaluation, which, in turn, could have an effect on the results. Assuming that fewer erroneous synonym suggestions would be made, lower grades become less frequent and the results for the method would be better.

To examine this assumption, it would be interesting to make some improvements of the methods and evaluate them again, but on a larger number of input words and without the criterion that both methods have to find candidates to the same input words. Possibly, Method 2 would still find synonym candidates to more input words than Method 1, due to its process of first translating and then filtering, but many of the erroneous synonym candidates could be filtered out. If so, Method 2 is the most useful method, with the ability to find synonym candidates to more words and a performance equal to or better than that of Method 1.

Additional to the inclusion of a threshold cosine value, other changes might also improve the methods' results further. A larger set of corpora, especially containing easy-to-read texts, as training data would probably result in better vector representations of words and fewer discarded input words. Furthermore, the word pairs that received the most "Disagree" (1) grades were words that differed in part of speech or grammatical form, which probably caused confusion among the participants. Lemmatization of the training and input data is an easy way to exclude such word pairs in the future. With these small changes made, many of the erroneous synonym suggestions could easily be avoided. This would, most likely, improve the results for both methods.

One possible application for the methods is to give suggestions on comprehensible synonyms. It would be helpful not only for readers, but also for text producers who want to be able to reach a wider public. With the suggested improvements, a combination of the methods could be used

for automatic lexical simplification. This would be an important step in automatic text simplification, as it makes the process of simplifying texts more effective and, thus, increases the amount of information that can be made easy to read.

5. Conclusions

In this paper, the development and evaluation of two novel methods to extract comprehensible synonyms were presented. The two methods were inspired by previous work on synonym extraction and made use of distributional semantic models and a bilingual dictionary. The methods were evaluated using an online survey, in which the perceived synonymy of word pairs, extracted by the methods, was graded from "Disagree" (1) to "Totally agree" (4). The results were promising and showed, for example, that the most common grade was "Sometimes" (3) for both methods, indicating that the methods found useful synonyms. Method 1 generated synonyms that were rated significantly higher than Method 2, and the synonyms generated by both methods had, not surprisingly, significantly higher rating than both Method 1 and Method 2, indicating that word pairs generated by a combination of methods are perceived as more synonymous. The results could easily be further improved with some small changes of the methods.

Future research on the methods include evaluation of the methods with the suggested improvements. Also, it is left to examine to what degree the extracted synonyms are perceived as more comprehensible.

Acknowledgements

This work would not have been possible if it were not for Linnea Fornander, Marc Friberg, Viktor Lind-Håård, and Pontus Ohlsson. This research was sponsored by The Internet Foundation in Sweden.

References

E. Abrahamsson, T. Forni, M. Skeppstedt and M. Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, pages 57–65.

R.C. Anderson and P Freebody. 1979. Vocabulary Knowledge. *Technical Report No. 136*.

M. Baroni, G. Dinu, and G. Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

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

E. Ejerhed, G. Källgren, and B. Brodda. 2006. Stockholm Umeå Corpus version 2.0.

L. Fornander, M. Friberg, V. Johansson, V. Lindh-Håård, P. Ohlsson, and I. Palm. 2016. Generating Synonyms Using Word Vectors and an Easy-to-Read Corpus.

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.

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.

R. Keskisärkkä and A. Jönsson. 2013. Investigations of Synonym Replacement for Swedish. *Northern European Journal of Language Technology*, 3(3):41–59.

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.

P. Mandera, E. Keuleers, and M. Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

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

L. Rello, R. Baeza-Yates, L. Dempere, and H. 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 *Proceedings of the second international workshop on Paraphrasing-volume 16*, pages 72–79, Sapporo. Association for Computional Lingusitics.