Automatic Text Simplification via Synonym Replacement

Robin Keskisärkkä, Arne Jönsson

Santa Anna IT Research Institute AB
Linköping, Sweden
robin.keskisarkka@liu.se, arnjo@ida.liu.se

Abstract

Automatic lexical simplification via synonym replacement in Swedish was investigated. Three different methods for choosing alternative synonyms were evaluated: (1) based on word frequency, (2) based on word length, and (3) based on level of synonymy. These three strategies were evaluated in terms of standardized readability metrics for Swedish, average word length, and proportion of long words, and in relation to the ratio of type A (severe) errors in relation to replacements.

1. Introduction

In this paper we present results from an investigation on whether a text can be successfully simplified using synonym replacement on the level of one-to-one word replacements. Theoretically, synonym replacements can affect established readability metrics in Swedish in different ways. The correlation between word length and text difficulty indicates that lexical simplification is likely to result in decreased word length overall, and a decrease in number of long words. Also, if words are replaced with simpler synonyms we can expect a smaller variation in terms of unique words, since multiple nuanced words may be replaced by the same word.

Studies within lexical simplification have historically investigated the properties of English mainly, and almost all rely in some way on the use of WordNet (Carroll et al., 1998; Lal and Rüger, 2002; Carroll et al., 1999). For Swedish there is no WordNet or system of similar magnitude or versatility. A few studies have used lexical simplification as a means of simplifying texts to improve automatic text summarization (Blake et al., 2007), and some have applied some type of lexical simplification coupled with syntactic simplification, but studies that focus on lexical simplification in its own right are rare. The studies that do exist tend to view lexical simplification as a simple task in which words are replaced with simpler synonyms, defining a *simpler* word as one that is more common than the original.

Words with identical meaning in all contexts are rare and any tool that replaces words automatically is therefore likely to affect the content of the text. This does not mean that automatic lexical simplification could not be useful, e.g. individuals with limited knowledge of economy may profit little by the distinction between the terms *income*, *salary*, *profit*, and *revenue*.

2. Method

We use the freely available SynLex in which level of synonymy between words is represented in the interval 3.0–5.0, where higher values indicate a greater level of synonymy between words. The lexicon was constructed by allowing Internet users of the Lexin translation service to rate the level of synonymy between Swedish words on a scale from one to five (Kann and Rosell, 2005). SynLex was combined with Parole's frequency list of the 100,000 most common Swedish words by summarizing the the frequencies of the different inflections of the words in the synonym dictionary. The final file contained synonym pairs in lemma form, level of synonymy between the words, and word frequency count for each word. The original synonym file contained a total of 37,969 synonym pairs. When adding frequency and excluding words with a word frequency of zero 23,836 pairs remained.

Readability was evaluated using LIX, OVIX, average word length, and proportion of long words. The texts were checked for errors manually, using a predefined manual. Inter-rater reliability, between two raters, was 91.3%.

Errors were clustered into two separate categories: *Type A errors* include replacements which change the semantic meaning of the sentence, introduce non-words, introduce co-reference errors within the sentence, or introduces a different word class (e.g. replaces a noun with an adjective). *Type B errors* consist of misspelled words, article or modifier errors, and erroneously inflected words.

Text were chosen from four different genres: newspaper articles from *Dagens nyheter* (DN), informative texts from Försakringskassan's homepage (FOKASS), articles from *Forskning och framsteg* (FOF), and academic text excerpts (ACADEMIC). Every genre consisted of four different documents which were of roughly the same size. The average text contained 54 sentences with an average of 19 words per sentence. In the experiments synonym replacement was performed on the texts using a one-to-one matching between all words in the original text and the available synonyms. A filter was used which allowed only open word classes to be replaced, i.e. replacements were only performed on words belonging to the word classes nouns, verbs, adjectives, and adverbs.

In the first two experiments the three conditions word frequency, word length, and level of synonymy are used to choose the best replacement alternative. The first strategy compares word frequencies and performs substitutions only if the alternative word's frequency is higher than that of the original, if more than one word meets this criteria the one with the highest word frequency is chosen. Similarly, word length replaces a word only if the alternative word is shorter, if more than one word meets the criteria the shortest one is chosen. The third strategy replaces every word

with the synonym that has the highest level of synonymy. In experiment two the inflection handler is introduced. The inflection handler allows synonym replacement to be performed based on lemmas, which increases the number of potential replacements. The inflection handler also functions as an extra filter for the replacements since only words that have an inflection form corresponding to that of the word being replaced are considered as alternatives. In the third experiment thresholds are introduced for the different strategies. The thresholds are increased incrementally and the errors are evaluated for every new threshold. Finally, in the fourth experiment word frequency and level of synonymy are combined and used with predefined thresholds.

3. Results

The results from experiment one showed that for all genres the replacement based on word frequency resulted in an improvement in terms of readability for every genre in all readability metrics. The error ratio was 0.52. Replacement based on word length also resulted in an improvement in terms of readability for every genre in all readability metrics. The average error ratio was 0.59. For all genres the replacement based on level of synonymy affected the readability metrics negatively for all metrics, except for the OVIX-value. The average error ratio was 0.50.

The results from experiment two showed that for all genres the replacement based on word frequency resulted in an improvement in terms of readability for every genre in all readability metrics. The error ratio was highest for FOF, 0.37, and lowest for FOKASS, 0.31. The average error ratio was 0.34. For all genres the replacement based on word length resulted in an improvement in terms of readability for every genre in all readability metrics. The error ratio was highest for FOKASS, 0.47, and lowest for FOF, 0.37. The average error ratio was 0.42. For all genres the replacement based on level of synonymy affected the readability metrics negatively for all genres in all readability metrics, accept for FOF where the OVIX-value decreased slightly. The error ratio was most highest for FOF, 0.46, and lowest for DN, 0.40. The average error ratio was 0.44.

Experiment three revealed no clear relationship between threshold and error ratio for any of the three replacements strategies. For some texts the error ratio decreased as the the threshold increased, while for others the opposite was true, and the ratio of errors remained relatively constant.

In experiment four we combined word frequency and level of synonymy. The frequency threshold was set to 2.0, meaning that only words with a frequency count of two times that of the original word were considered possible substitutions. The threshold for level of synonymy was set to 4.0. The experiment was run twice, prioritizing either word frequency (PrioFreq) or level of synonymy (PrioLevel) when more than one word qualified as an alternative. The same words are replaced in both cases, but the word chosen as the substitution may differ. In both runs the average error ratio was 0.27. PrioLevel performed significantly better than the frequency strategy in experiment two in terms of error ratio when looking at all texts and genre was not considered. Also, both PrioLevel and PrioFreq performed significantly better than the frequency strategy alone when looking at the genre DN specifically.

4. Discussion

The overall error ratio of replacing synonyms based on frequency is not significantly affected by the introduction of relative threshold frequencies for alternative words. As the frequency threshold is increased the new words are more likely to be familiar to the reader, but this does not significantly increase the likelihood that the replacement is a correct synonym in the particular context. Word length as a strategy for synonym replacement improves the text in terms of the readability metrics, but it is not clear whether it contributes to the actual readability of the text. Also, the combination of frequency and level of synonymy slightly improves the error ratio compared to frequency alone.

The study shows that the common view of automatic lexical simplification as a task of simply replacing words with more common synonyms results in a lot of erroneous replacements. The error ratio does not critically depend on level of synonymy, rather the overall error ratio remains roughly the same even when using words with the highest level of synonymy. The high error ratios at this level confirm that the concept of synonyms is highly dependent on the context. Handling word collocations and word disambiguation could greatly improve both the quality of the modified texts and substantially decrease the error ratio, but this is by no means a trivial task.

A simplified text can potentially be useful, even if it contains some errors, especially if the original text is too difficult to comprehend for the unassisted reader. It would be interesting to study the sensitivity of readers to typical erroneous replacements, and the effects simplification has on comprehension. Future studies should also aim at replacing only those words which are regarded difficult to a particular reader, rather than trying to simplify all words.

5. References

Catherine Blake, Julia Kampov, Andreas K Orphanides, David West, and Cory Lown. 2007. Unc-ch at duc 2007: Query expansion, lexical simplification and sentence selection strategies for multi-document summarization. *Proceedings of Document Understanding Conference (DUC) Workshop 2007*.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, volume 1, pages 7–10. Citeseer.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 269–270.

Viggo Kann and Magnus Rosell. 2005. Free construction of a free swedishdictionary of synonyms. In *NoDaLiDa 2005*, pages 1–6. QC 20100806.

Patha Lal and Stefan Rüger. 2002. Extract-based summarization with simplification. In *Proceedings of the ACL*.