Auxiliary Techniques to Help Readers Understand Texts

Daniel Holmer, Arne Jönsson
Department of Computer and Information Science
Linköping University, Linköping, Sweden
holmer.daniel@gmail.com, arne.jonsson@liu.se

Abstract

We explore three auxiliary techniques for automatic text adaptation (ATA)—epithets for nouns, explanations for keywords, and syllabification—to aid reading for individuals with reading difficulties. In an initial evaluation, we conducted a study with individuals possessing average reading skills. Results indicate that while all three techniques demonstrate high accuracy, their usefulness varies. Epithets were found to be less beneficial, possibly due to the introduction of excessive information, although they may assist certain populations, such as individuals with intellectual disabilities. Keyword explanations were generally helpful and accurate, though occasional inaccuracies arose with rare or domain-specific terms. The effectiveness of syllabification was found to be contingent on the specific words being processed. These findings suggest that while ATA techniques can improve reading accessibility, their varying impacts highlight the need for tailored approaches based on the reader's needs.

1 Introduction

Text adaptation normally includes lexical simplification, syntactic simplification and various forms of text summarisation. Other important techniques to make texts easier to read include font type and size, line width and line spacing. But there are other means to make texts easier to read that are in between the former language technology techniques that rewrite a text and the latter more surface oriented, requiring no linguistic processing. In this paper we present three such techniques: the use of epithets to help understand certain nouns, explaining central keywords in a text instead of simplifying them, and splitting words into syllables.

We select the three techniques based on the guidelines developed by the Swedish Agency for Accessible Media (MTM, 2021) and various studies, c.f. Kearns and Whaley (2019). These guidelines suggest different ways to write texts in an accessible way, for instance regarding linguistic constructions and to select simple and short words. However, in some instances there are no suitable simple and short synonyms to a word. Difficult words should therefore be given an explanation (MTM, 2021). Two of our techniques aim to provide this in different ways; the first by providing short, descriptive epithets to give more context to a word, and the second by providing a clarifying explanation to certain keywords.

2 Methods

In all studies we will use three Swedish texts on minority languages in Sweden, one on Yiddish, one on Finnish and one on Swedish Sign Language. The texts are to be used in an extensive study with readers having reading problems and are part of education material provided by The Institute for Language and Folklore (Isof)[1], where texts on Sweden's minority languages and Sign language are covered. In our selection of the three texts, we ran an analysis of six different text complexity metrics on all texts, and selected the three texts that had the most similar complexity according to metrics about different aspects of the texts. We use LIX (Björnsson, 1968) as a surface metric regarding sentence and word length, OVIX (Hultman and Westman, 1977) for idea density, three syntactic metrics (AVG_DEP_DISTANCE_DEPENDENT, AVG_SENTENCE_DEPTH, and NOMINAL_RATIO) (Falkenjack, 2018), and the cohesion metric ADJACENT ANAPHORS[2].

[1] https://www.isof.se/nationella-minoritetssprak/laromedel/laromedel-fran-isof

[2] Index 38 from the Coh-Metrix documentation found at https://web.archive.org/web/20230130040543/http://cohmetrix.memphis.edu/cohmetrixhome/documentation_indices.html

2.1 Epithets

Epithets are descriptive terms accompanying the name of a person, place, or thing. For epithets we use a pipeline of two BERT-models. The first is fine-tuned for named entity recognition (NER)[3]. The model is trained to identify different types of entities, for instance persons, locations and organisations. For each such identified entity, we add a [MASK] token in the position before the entity. We then feed the whole sentence to a second BERT-model (Malmsten et al., 2020)[4], which is tasked to predict the [MASK]-token. In essence, this mimics the MLM pre-training step described in (Devlin et al., 2019). We add an additional post processing step that cross references the predicted epithet token to a list of manually curated epithets, to make sure that added tokens are a theoretically valid epithet. A typical epithet to, for instance, the word Sweden is "the country" producing "the country Sweden".

2.2 Keywords

To extract keywords we use a system based on YAKE! (Campos et al., 2020), a custom n-gram extractor, and KeyBERT (Grootendorst, 2020).

We use YAKE! and the n-gram extractor to find possible keyword candidates. These candidates are then fed to KeyBERT, which ranks the most relevant keywords from the candidate list. KeyBERT follows an approach where it uses embeddings from a BERT-model in two steps. First, it works on the word level, where an embedding for each candidate keyword is created. Second, it creates embeddings on the document level. To select the most important keywords, the cosine-similarities between all the candidate and document embeddings are calculated, and the keywords with the highest similarity to the documents are considered to be the most relevant.

While it is possible to let KeyBERT treat the entire text as keyword candidates, we opted for the pre-processing approach where YAKE! and the n-gram extractor provide a limited selection of candidates. The reason for this is two-fold; we want to have greater control over what words are possible for selection (we select proper nouns, adjectives, and nouns as valid candidates for the n-gram extractor), and due to limited hardware we want to avoid creating BERT-embeddings for every individual word in the whole text. To further alleviate the computational need of KeyBERT, we use a distilled version of the Swedish SBERT model (Rekathati, 2021) from KBLab[5].

The identified keywords are then given an explanation by prompting the LLM GPT4-TURBO-PREVIEW[6] from OpenAI. We use a zero-shot prompt where the model is instructed, in Swedish, to explain the given word in a simple way, and avoid using difficult words:

> *Provide an explanation in no more than one sentence for this word:* ***{word}****. The explanation should be easy to understand and not contain long or difficult words. Use words that are easy to understand.*

where **{word}** is the given keyword to be explained.

We keep the hyper-parameters at their default values in the CHAT-COMPLETIONS interface from OpenAI.

A typical example is *Sign Language – Sign language is a language where hands, facial expressions, and body movements are used to communicate instead of speaking with the voice.*

2.3 Syllabification

The syllabification technique used in this research is based on morphological rather than phonetic principles. We use the compound analysis of the Sparv pipeline[7] where tokens and their POS tags are looked up in the SALDO lexicon (Borin et al., 2013) and enriched with compound information.

The compound analysis includes identifying candidate words according to criteria such as having a prefix in the SALDO lexicon, being compound, having a suffix with certain properties such as being noun, verb or adjective, etc. The candidates are then ranked based on criteria such as number of compounds and a statistical model [8] (Borin et al., 2016).

[3] https://huggingface.co/KBLab/bert-base-swedish-lowermix-reallysimple-ner

[4] https://huggingface.co/KBLab/bert-base-swedish-cased

[5] https://huggingface.co/KBLab/sentence-bert-swedish-cased

[6] All further references to GPT-4 are the state of GPT4-TURBO-PREVIEW as accessed from OpenAI:s API in February 2024.

[7] https://spraakbanken.gu.se/sparv/#/user-manual/available-analyses?id=compound-analysis-with-saldo

[8] Details in Swedish on https://spraakbanken.gu.se/faq/hur-fungerar-sparvs-sammansattningsanalys

The three texts contain a total of 1788 words of which 254 achieved a syllabification from the SALDO lexicon (200 unique instances).

The results from the syllabification often affects compound words, such as, *väst•finska* (eng. west•Finnish), but there are also instances of syllabification's inside words, like *ar•bete* (eng. work).

3 Evaluation

The three techniques have been assessed by readers with average reading skills, 10 students and teachers at Linköping University, in a survey comprising 10 random instances of each technique. The survey uses a five grade Likert scale from *Helt enig* (eng. Strongly agree), grade 5 to *Helt oenig* (eng. Strongly disagree), grade 1.

For the ten epithets and the ten keyword explanations two questions were asked: *The epithet|word explanation is correct* and *The epithet|word explanation facilitates understanding*. For the syllabifications the first question was instead formulated as if the syllabification is good, as there is not always an obvious correct syllabification and the second was formulated as *The syllabification facilitates reading* as syllabification is more an aid for reading.

Table 1 shows the results from the survey. We present both results interpreting the Likert scale as an interval scale, mean and standard deviation, as well as an ordinal scale, median. As can be seen from Table 1 all techniques perform well, median 5.0. Looking at the mean we see that there are some deviations and when we further study the various items we can identify some interesting patterns.

Looking at the epithets we see that some of them are considered less correct and helpful. For instance, words with the epithet "state" (e.g. the state of Sweden) as opposed to the epithet "country" (e.g. the country Sweden) are considered less correct and also considered less helpful.

Overall epithets facilitate understanding the least, median only 3.0.

Word explanations are also correct and much more useful. Interestingly, one word explanation is wrong, the explanation about the Swedish so called *Judereglementet* (eng. 'The Jewish Regulations') explains rules of the sport Judo. This is also observed by more or less all participants and considered both not correct and not to facilitate understanding. If we remove this item we get a mean of 4.6 (stdev 0.903) for explanations and 4.5 (stdev. 0.946) for facilitating understanding, which clearly shows that word explanations are both correct and facilitates understanding.

Syllabifications, finally, are also considered correct but does not facilitate reading as much. Here we see differences between the two types of syllabifications that the technique provides, one that more or less divides Swedish compounds into their parts, such as *tecken•språket* (eng. the sign language) and the other that divides words into syllables, such as *as•kan* (eng. the ashes). The latter is regarded less correct and much less useful. This also seems to depend on the length of the word, short words such as *askan* are considered both not correct and not to facilitate reading whereas slightly longer words, such as *ar•bete* (eng. work) are considered less incorrect and not as bad when it comes to facilitate reading.

4 Discussion

All techniques perform well and we can conclude that it is possible to accurately add epithets to nouns, explain keywords, and perform syllabification on words. Their estimated usefulness, for readers with average reading skills, varies, however. Estimated usefulness does not necessarily mean that texts with these features would have been helpful for the readers in the study. Rather, they can see potential gains from using these features.

Epithets clearly do not always facilitate understanding even if they are correct. This may not be surprising for readers with average reading skills, where the epithet can be seen as adding a word that is not necessary to understand the word. We believe that the same is true for people with dyslexia, but, for instance, for people with intellectual disabilities we believe that it may be useful, c.f. Nilsson et al. (2021).

Recently, different types of LLMs have shown great results on many NLP-tasks, and in particular generation tasks. In the case of word explanations, it is clear that an LLM in the like of GPT-4 can provide more helpful explanations than previous techniques. However, in the task of generating an epithet before identified keywords, the advantages of such LLMs are not as obvious. In our experiments, we also generated epithets using LLMs by prompting GPT-4 to adhere to the experimental settings of our BERT-based system. The gains of the more environmentally expensive GPT-4 model are slight, for instance, our BERT-based system

	Median	Mean	Standard deviation
The epithet is correct	5.0	4.15	1.445
The epithet facilitates understanding	3.0	2.99	1.507
The word explanation is correct	5.0	4.31	1.309
The word explanation facilitates understanding	5.0	4.25	1.268
The syllabification is good	5.0	4.27	1.318
The syllabification facilitates reading	4.0	3.67	1.537

Table 1: Results from the evaluation

sometimes would put the epithet "state" instead of "country" in conjunction to countries. When prompting GPT-4, it consistently delivers "country". Otherwise, the results are shown to be nearly identical. It is however possible that the advantages of an LLM-based solution would be more evident where even more complex words would have to be associated with an epithet, or in a setting where for example phrases of epithets would be allowed (i.e. "the *Nordic country* Sweden", or "the *American city* New York").

Word explanations are considered more helpful. When they are correct they also facilitate understanding. However, there is a risk that they are wrong. As expected, large language models (in this case GPT-4), struggle to provide a feasible explanation for rare and highly domain specific words. For instance, this is demonstrated in the earlier example, where the term *Judereglementet* (eng. 'The Jewish Regulations') resulted in an explanation of the rules of the sport Judo. It is reasonable to believe that the term was not all that common in the model's training data, and therefore these kinds of hallucinations might appear.

Syllabification is the technique that most depends on which word that is being processed. Many of the words were considered less useful, and sometimes not even considered correct by readers with average reading skills. We believe, however, that syllabification of these words may help people with dyslexia, c.f. Vellutino et al. (2004); Hyönä and Olson (1995), not only compounds but also words that are not compounds; unless they are very short. In our further studies we will not syllabify words shorter that 6 characters, c.f. Björnsson (1968). For our three texts, we are then left with 215 (out of 254) words to be syllabified.

5 Summary

In this paper, we have presented results from an investigation of three techniques that could be used in conjunction with text summarization and text simplification to facilitate reading for different target groups. The three techniques are the addition of epithets for nouns, explanations of keywords, and splitting words into syllables.

The techniques were evaluated through a survey completed by individuals with average reading skills. Although the evaluation was limited in scope and the results are indicative, they demonstrate that all three techniques generally perform well in their respective tasks. However, the usefulness for individuals with average reading skills varies. The effectiveness of epithets largely depends on the specific epithet used; some are helpful, while many do not enhance understanding. In contrast, word explanations are consistently perceived as beneficial, while the effectiveness of syllabification also depends significantly on the specific words being syllabified. Future studies will explore how these techniques are received by readers with different reading difficulties.

Acknowledgments

This research is part of the project Text Adaptation for Increased Reading Comprehension, funded by The Swedish Research Council.

References

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm, Sweden.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Johan Falkenjack. 2018. *Towards a model of general text complexity for Swedish*. Licentiate thesis, Linköping University Electronic Press.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Tor G. Hultman and Margareta Westman. 1977. *Gymnasistsvenska*. LiberLäromedel, Lund.

Jukka Hyönä and Richard K Olson. 1995. Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6):1430.

Devin M. Kearns and Victoria M. Whaley. 2019. Helping students with dyslexia read long words: Using syllables and morphemes. *Teaching Exceptional Children*, 51(3).

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden – making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

MTM. 2021. Att skriva lättläst. https://www.mtm.se/var-verksamhet/lattlast/att-skriva-lattlast/. Accessed: 2021-10-05.

Karin Nilsson, Henrik Danielsson, Åsa Elwér, David Messer, Lucy Henry, and Stefan Samuelsson. 2021. Investigating reading comprehension in adolescents with intellectual disabilites: Evaluating the simple view of reading. *Journal of Cognition*, 4(1):1–16.

Faton Rekathati. 2021. The KBLab blog: Introducing a Swedish sentence transformer. https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/.

Frank R Vellutino, Jack M Fletcher, Margaret J Snowling, and Donna M Scanlon. 2004. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, 45(1):2–40.