DSim, a Danish Parallel Corpus for Text Simplification

Sigrid Klerke, Anders Søgaard

Center for Language Technology
University of Copenhagen
sigridklerke@gmail.com, soegaard@hum.ku.dk

Abstract

We present DSim, a new sentence aligned Danish monolingual parallel corpus extracted from 3701 pairs of news telegrams and corresponding professionally simplified short news articles. The corpus is intended for building automatic text simplification for adult readers. We compare DSim to different examples of monolingual parallel corpora, and we argue that this corpus is a promising basis for future development of automatic data-driven text simplification systems in Danish. The corpus contains both the collection of paired articles and a sentence aligned bitext, and we show that sentence alignment using simple tf*idf weighted cosine similarity scoring is on line with state–of–the–art when evaluated against a hand-aligned sample. The alignment results are compared to state of the art for English sentence alignment. We finally compare the source and simplified sides of the corpus in terms of lexical and syntactic characteristics and readability, and find that the one–to–many sentence aligned corpus is representative of the sentence simplifications observed in the unaligned collection of article pairs.

Keywords: Simplification, Monolingual, Parallel corpus, Sentence alignment, Tf*idf

1. Introduction

Monolingual parallel corpora are used for building and evaluating data-driven systems for paraphrasing, text generation, summarizing and text simplification. In this paper we focus on text simplification.

Rule-based automatic text simplification has been investigated previously, but recently the field has experienced success experimenting with tools inspired from statistical machine translation. For exploring this approach further, access to aligned monolingual corpora of normal and simplified text is a necessary prerequisite.

We understand simple text as a type of text with a limited vocabulary of high frequency words and a simple and predictable syntax. These notions depend to a great extent on the target reader population. For instance, how frequent a word is seen and how predictable specific syntactical structures are to a reader depends on whether the reader is a child starting to learn to read, an adult second language learner or an elderly aphasic patient. It is thecharacteristics of the original data that constrain which possible target reader groups a data-driven system might be able to address. The aim of this work is to describe the simplifications present in our corpus.

Automated text simplification is useful both as an intermediate step of complex natural language processing systems and on its own. As part of a system, simplification has been applied to facilitate parsing, translation and generation of text for question generation and answering, summarizing and dialog systems. As an end product, simplification aids people with reading disabilities and language learners and facilitates rewriting of instructional and other texts where readability puts direct constraints on the final text such as security guidelines.

2. Other Monolingual Parallel Corpora

A number of text types are candidate resources for building monolingual parallel corpora. The most prevalent resource for simplification at the moment is the English Wikipedia paired with Simple-English Wikipedia (Coster and Kauchak, 2011). This is the underlying corpus for all state-of-the-art automatic simplification systems (Woodsend and Lapata, 2011; Yatskar et al., 2010; Zhu et al., 2010). But unlike the main Wikipedia, the simple version is only available in English.

In the Britannica corpus, Barzilay and Lee (2003) pairs encyclopedic resources covering comparable content but intended for different readers, e.g. adults and children. It is possible, however, that relying on such data could lead to simplifications perceived as childish due to an audience specific language use.

Another source of texts with a systematic difference in complexity was included in a study by Marsi and Krahmer (2007). Among other resources, they exploited the parallel text hidden in sub-texted news television production where the subtitles are a simplified version of the teleprompter manuscript. By trying to capture as much of the spoken prompt as possible in a restricted space while being optimized for fast reading, subtitlers produce a simplification resource which is inherently closely aligned. However, its special synchronicity suggests that this resource may be optimized for deletions and light syntax revisions while avoiding lexical substitutions.

Not all monolingual corpora are suited for simplification purposes because the aligned texts do not differ systematically in readability. For text generation Dolan et al. (2004) collected a large monolingual parallel corpus of news from Internet news sources, and Marsi and Krahmer (2007) aligned telegrams from two Dutch news agencies. The lack of a simplification relation also holds for retranslations of books, as was also used in Marsi and Krahmer (2007), and for the different gospels of the Bible used for sentence alignment evaluation in Nelken and Shieber (2006).

All of these examples are pre-aligned to a varying degree, ranging from a topical alignment, as in Wikipedia and other encyclopedias, to paragraph alignments, as in the combination of TV-prompt and subtitles. Below, we describe the DSim corpus in terms of origin, alignment quality and observed simplifications.

3. Development of the DSim Corpus

The Danish parallel corpus for simplification consists of a collection of news telegrams in Danish and hand-simplified versions of the same texts. The simple news service is non-commercial, and while it was first launched on the website ligetil.dk in August 2008, it is now published at dr.dk/ligetil, as part of the Danish Broadcasting Corporation's news and educational services. The resource is intended for use by reading impaired adults and adult learners of Danish. The simple news is rewritten by hand from ordinary news telegrams and has been published on weekdays since 2008. The simplifications are done by journalists specifically trained for the task.

At the lexical level, strategies reported by the journalists include substituting rare words with more common words and minimizing the use of long words, either by replacing them with shorter, more frequent alternatives or by providing a description instead. A third frequently used strategy is to hyphenate long compound words[1], a practice that is generally considered to be incorrect, but has a measurable effect on reading-ease for impaired readers, by facilitating a morphological analysis strategy (Elbro and Arnbak, 1996).

A few general sentence-level simplification strategies were reported by the journalists. These mainly result in splitting or pruning syntactic structures into predictable units and simplifying the flow of information by limiting the amount of new content introduced in each sentence. This may lead to repeating pieces of information where deemed necessary. Ordering of information is also considered, as texts presenting events in accordance with their chronological order are in general easier to comprehend for a struggling reader.

The strategy of reducing the average sentence length is evident in Figure 1 which depicts the number of sentences of a given length in the original and the simplified articles. For all sentence lengths up to 15 words, the count has more than doubled in the process of simplifying the articles.

At the formatting level, the remaining decisions of the journalist involve reformatting with more paragraphs, adding background facts and linking to related video material. Addition of background material, which helps minimizing the need for making difficult inferences, together with some of the syntactic and lexical simplification strategies frequently result in the simple articles being slightly longer than their original source telegram.

4. Sentence Alignment

In this section we present the alignment procedure and evaluation before describing the resulting aligned corpus statistically.

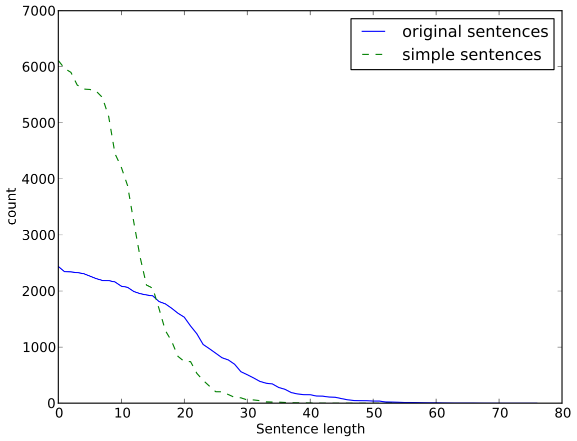

Figure 1: Frequency of each sentence length in the source articles and in their simplified counterparts.

4.1. Method

No link between the simplified and the original article was kept for the first two years and only loosely kept for the past year. Therefore, recovering the pairs was based mainly on publication date and subsequent coarse content mapping yielding 3701 article pairs.

We aligned sentences in each article pair by calculating the cosine similarity between vector representations of the source and the simplified target sentences and aligning all matches with a similarity score above a fixed threshold. Each vector of terms was weighted by the term frequency inverse document frequency (tf*idf) metric.

Tf*idf is a metric from Information Retrieval, which is used for weighting *local* term frequencies in a document by the *global* term frequencies in a collection of documents. Following Nelken and Shieber (2006) one article is taken to represent a document collection, and each sentence plays the role of a document of the collection. The definition of the function for assigning weights to each term $w_s(t)$ is given in Equation (1) where N is the number of sentences in the source article, $\mathrm{TF}_s(t)$ is the frequency count of the term in the sentence, and $\mathrm{DF}(t)$ is the number of sentences in which t occurs.[2]

$$w_s(t) = \mathrm{TF}_s(t) \cdot \log\left(\frac{N}{\mathrm{DF}(t)}\right) \tag{1}$$

For measuring the similarity of two sentences $\mathbf{s_1}, \mathbf{s_2}$ by their vector space cosine similarity, we use Equation (2).

$$sim(\mathbf{s_1}, \mathbf{s_2}) = \frac{\mathbf{s_1} \cdot \mathbf{s_2}}{|\mathbf{s_1}||\mathbf{s_2}|} = \frac{\sum_{i=1}^{t} w_{s_1}(t)\, w_{s_2}(t)}{\sqrt{\sum_{i=1}^{t} w_{s_1}(t)^2}\sqrt{\sum_{i=1}^{t} w_{s_2}(t)^2}} \tag{2}$$

[1] An example of a word that is not replaced but hyphenated is *udenrigs-minister* (minister of foreign affairs) which is normally written *udenrigsminister*

[2] This definition differs from the definition in (Nelken and Shieber, 2006) by counting term frequency rather than using a binary indicator, and by only considering the source article as the collection to weigh term frequencies against.

Nelken and Shieber (2006) proposed to use this method combined with a step of logistic regression for determining the threshold level for when to accept a match. Nelken and Shieber (2006), Barzilay and Lee (2003) and Coster and Kauchak (2011) all use different global algorithms to constrain the possibilities of long distance alignments and one-to-many alignments to differing degrees. This introduces a sensitivity to the relative ordering of sentences. Nelken and Shieber (2006) shows that this sensitivity is beneficial because it reduces the number of incorrect long-distance alignments. Given our collection of short one-topic articles, the first step is sufficient, as shown in the following.

The term set we use for calculating the weights and the cosine distance of a potential sentence pair in a document pair, is the intersection of the term sets of both documents. Also, n-grams up to a sequence of 5 tokens, appearing in both documents, are added to the term set. This serves to boost the similarity score of short sentences aligning only to a part of a source sentence and enables us to rediscover cases of sentence splitting. 5-grams were chosen as maximum, rather than shorter or longer sequences, because at this level the size of the intersected term sets stagnated, meaning that we rarely see units of 5 or more tokens appear without changes between source and target. In order to be able to represent even very short sentences reliably, we removed only punctuation and four top-frequent stopwords.

4.2. Alignment Results

For evaluation we hand-aligned a sample of 16 document pairs containing 585 sentences. Of these 230 were source sentences and 355 target sentences. The alignments were created by correcting automatically identified alignments rather than discovering them from scratch, following the guidelines in Marsi and Krahmer (2007). Nelken and Shieber (2006) also notes the difficulty for readers in detecting all good candidate alignments.

Allowing one-to-many relations yielded a total of 313 aligned sentence pairs in the sample. Of these 26 sentence pairs, or 8.3% of the pairs, had targets identical to the source, 70 sentences, or 11.2% of the 585 sentences, were not aligned.

Precision and recall of the alignment when evaluated against the hand-aligned sample is presented in Figure 2. Performance of the tf*idf at two different similarity threshold levels along with the portion of unaligned sentences for each level and for the hand-aligned sample can be seen in Table 1. To our knowledge no research has been done in sentence aligning Danish. Therefore, our comparison is to the results reported for the Britannica corpus by Barzilay and Lee (2003) and Nelken and Shieber (2006), shown in Table 2. Both of these aproaches take further steps to control sentence alignment by constraining or penalizing distant and multiple alignments. As the articles on the source and target side of the Danish data are very similar to each other, and the paired articles are short (10-30 sentences), it is possible to allow many-to-many alignments with no restrictions on the relative distance or ordering of aligned sentences without risking too many false positives. Thus it is sufficient to align sentences using the tf*idf weighted cosine similarity to get a high precision of 90.8%, even at a recall of 84.7% at a similarity score threshold of .35. Adjusting the threshold to .5 in order to obtain a precision of 94.9% drops recall to 70.9%. These figures are still higher than the reported performance on the one-third of the aligned Britannica corpus with a lexical overlap in the range 40–70%.

	Unaligned	Precision / Recall
Hand-align	11.2%	—
Tf*idf .35	19.0%	90.8% / 84.7%
Tf*idf .5	35.2%	94.9% / 70.9%

Table 1: Percent unaligned sentences for different alignment thresholds and precision/recall compared to hand-aligned sample.

	Precision / Recall
Cosine sim.*	57.9% / 55.8%
Tf*idf**	77.0% / 55.8%
Nelken & Shieber**	83.1% / 55.8%
Barzilay & Elhadad*	(85%) / (73%)

Table 2: * From Barzilay and Lee (2003), bracketed result is on the third of the aligned paragraphs with a lexical overlap of 40–70%. ** From Nelken and Shieber (2006). The Britannica corpus is not directly comparable as described in the text.

5. Characteristics of DSim

We chose to build our aligned corpus from the alignments with a threshold value of .35, keeping each aligned source sentence once and aligning it to the longest possible unbroken sequence of target sentences. This is a more restrictive alignment than free one–to–many alignment. We chose the restrictive alignment under the assumption that it would yield the most meaningful, fluent target strings compared to one–to–one or many–to–many. We compare the source side and the simplified target side in terms of vocabulary

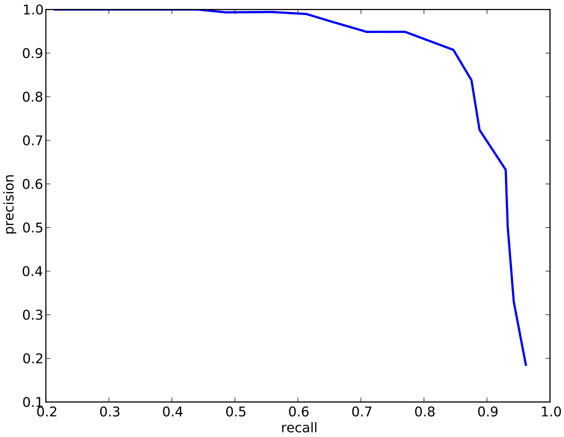

Figure 2: Sentence alignment precision and recall.

size, average sentence length, proportion of long words (i.e. longer than 6 letters) and LIX[3]. LIX is a standard measure of readability commonly used for classifying texts, originally introduced by (Bjornsson, 1983). A LIX-score of 30 is classified as easy text and 40 is classified as average. In Danish, typical books for children score 25, while fiction and factual prose score 35 and 45, respectively. The corpus statistics, shown in Table 3, reveal differences of average sentence length of nearly 6 words per sentence and a 31% reduction of vocabulary size. This is naturally reflected in the LIX drop in complexity from 45, a level of factual prose, to 37, which is a level above fiction but below average readability. These differences of relevance to readability scoring are comparable to the differences which were present in the corpus before it was aligned (indicated in brackets in Table 3). However, the restrictive alignment cuts the total number of sentences on the simplified side by 24%, while only cutting 14% on the source side.

	Source	Simplified
Vocabulary	52,273	36,154
Sentences	48,186 (55,708)	62,365 (82,200)
Sentence length / SD	17.3 / 9.3 (17.1 / 9.4)	11.1 / 5.3 (10.8 / 5.3)
%–Long words	27.5% (28.8%)	25.4% (26.4%)
LIX	45 (46)	37 (37)

Table 3: Statistics for the sentence aligned DSim corpus. Bracketed results are statistics for the unaligned corpus. %–Long words is the proportion of words with more than 6 letters.

6. Conclusion and Future Work

We have presented a simplification corpus for Danish and shown that sentence aligning using simple tf*idf weighted similarity alone pair the state of the art on English sentence alignment (Nelken and Shieber, 2006) with regards to precision and outperforms it on recall on this closely aligned Danish parallel corpus. In addition, we have shown that the aligned corpus is a good model of the entire hand-simplified corpus with regards to readability as measured by sentence length and amount of long words.

The DSim corpus presented here will be the basis for investigating automated text simplification in Danish with emphasis on exploring robustness and precision of existing state-of-the-art strategies when applied to a different language and a closely aligned corpus.

7. References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 16–23. Association for Computational Linguistics.

C. H. Bjornsson. 1983. Readability of Newspapers in 11 Languages. *Reading Research Quarterly*, 18(4):480.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 665–669. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. ACL.

C. Elbro and E. Arnbak. 1996. The role of morpheme recognition and morphological awareness in dyslexia. *Annals of dyslexia*, 46(1):209–240.

Erwin Marsi and Emiel Krahmer. 2007. Annotating a parallel monolingual treebank with semantic similarity relations. In *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories*, pages 85–96.

Rani Nelken and Stuart M Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06).*

Kristian Woodsend and Mirella Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (2011)*, pages 409–420.

Mark Yatskar, Bo Pang, C. Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and I. Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*, pages 1353–1361.

[3] LIX (Læsbarhedsindex) is the sum of the average sentence length, S, and the percentage of words longer than 6 letters, W; $LIX = S + W$