Similarity-Based Alignment of Monolingual Corpora for Text Simplification Purposes

Sarah Albertsson, Evelina Rennes, Arne Jönsson
SICS East Swedish ICT AB, Linköping, Sweden
Department of Computer and Information Science,
Linköping University, Linköping, Sweden
`s.hantosialbertsson@gmail.com` `evelina.rennes@liu.se`
`arne.jonsson@liu.se`

Abstract

Comparable or parallel corpora are beneficial for many NLP tasks. The automatic collection of corpora enables large-scale resources, even for less-resourced languages, which in turn can be useful for deducing rules and patterns for text rewriting algorithms, a subtask of automatic text simplification. We present two methods for the alignment of Swedish easy-to-read text segments to text segments from a reference corpus. The first method (M1) was originally developed for the task of text reuse detection, measuring sentence similarity by a modified version of a TF-IDF vector space model. A second method (M2), also accounting for part-of-speech tags, was developed, and the methods were compared. For evaluation, a crowdsourcing platform was built for human judgement data collection, and preliminary results showed that cosine similarity relates better to human ranks than the Dice coefficient. We also saw a tendency that including syntactic context to the TF-IDF vector space model is beneficial for this kind of paraphrase alignment task.

1 Introduction

Automatic text simplification is defined as the process of reducing text complexity, while maintaining most of the content (Chandrasekar and Srinivas, 1997; Carroll et al., 1998). While the first approaches handled the task of automatically simplifying texts by the application of hand-crafted rules, cf. Rennes and Jönsson (2015) for a recent example for Swedish, data-driven methods have gained momentum in text simplification, as in other areas within natural language processing.

By using large corpora of aligned monolingual material, it is possible to automatically extract patterns or rules but since such induction requires large amounts of aligned material, hand-crafted systems are often the only practically performable alternative for languages without such resources. To automatically collect comparable corpora would result in large-scale resources useful for deducing rules and patterns for text rewriting algorithms, particularly beneficial for less-resourced languages with sparse linguistic resources.

In this study, we hypothesized that we could extend the work of Sanchez-Perez et al. (2014), originally developed for test reuse detection, to detect paraphrased segments from two corpora; one of them containing only easy-to-read material, and the other representing the full spectra of Swedish texts. We replicated the algorithm proposed by Sanchez-Perez et al. (2014), and contrasted this method to a method that included part-of-speech tags as context. An ongoing crowdsourcing evalutation is presented, in terms of evaluation design and preliminary results.

2 Related Work

The creation of aligned comparable monolingual corpora has been suggested as a step for several tasks within the field of natural language processing, such as paraphrasing (Barzilay and Elhadad, 2003; Dolan et al., 2004), automatic text summarisation (Knight and Marcu, 2000; Jin, 2002), terminology extraction (Hazem and Morin, 2016), and automatic text simplification (Bott and Saggion, 2011; Coster and Kauchak, 2011; Klerke and Søgaard, 2012).

According to Nelken and Shieber (2006), the alignment task of monolingual corpora differs from the multi-lingual counterpart. While the latter exhibits conformity between source and target documents, aligned monolingual material is characterised by being similar in content, rather than linguistically, making many established methods developed for multi-lingual alignment less successful.

Since the aim of monolingual text alignment is to find similar text fragments, it forms an important subtask of applications such as text reuse detection. The method described in this paper is inspired by the approach used by Sanchez-Perez et al. (2014), originally developed for the task of text reuse detection, measuring sentence similarity by a modified version of a TF-IDF vector space model. Other approaches to monolingual text alignment have measured sentence similarity by TF-IDF. Nelken and Shieber (2006) used this similarity score, treating each individual sentence as a document, in order to estimate the probability that two sentences are aligned using logistic regression, achieving higher accuracy than previous systems. This was confirmed by Zhu et al. (2010) who showed that when comparing the TF-IDF approach by Nelken and Shieber (2006) to other similarity measures for monolingual text alignment, the former outperformed the other two measures (word overlap and maximum edit distance).

For the task of text simplification, easy-to-read material aligned to their original counterparts would be beneficial. One way to ensure that a text is easy-to-read is by applying readability metrics. For Swedish, the standard readability metric is the *Readability Index*, LIX (Björnsson, 1968), but recently, its dominance has been questioned (Mühlenbock and Johansson Kokkinakis, 2009; Heimann Mühlenbock, 2013). Another measure, often said to complement LIX for Swedish, is the *Word Variation Index*, OVIX (Hultman and Westman, 1977). The readability metrics considered in this study are further defined in section 3.3.

An automatic evaluation of paraphrases requires annotated data. Crowdsourcing enables a cheap and effective way of collecting multiple annotations from multiple annotators for one task. For annotation tasks, Snow et al. (2008) showed that crowdsourced annotations are similar to traditional annotations made by experts. The field of natural language processing has used crowdsourcing for various tasks within the NLP area, often by the use of Amazon's Mechanical Turk (Callison-burch et al., 2006). For Swedish, this is a resource that is lacking and it can be a tedious task to collect human annotated data.

3 Method

This section will describe the text resources used in this study, as well as an overview of the replicated algorithm and the experimental design of the evaluation.

3.1 Corpora

LÄSBART (Mühlenbock, 2008) is a Swedish corpus containing a collection of easy-to-read material of a total of 1.1 million tokens. Four genres are represented in the corpus; easy-to-read news texts, fiction, community information, and children's fiction.

STOCKHOLM-UMEÅ CORPUS (SUC) (Källgren et al., 2006) is a corpus of one million words of published Swedish texts written in the 1990's. The corpus is balanced according to genres and annotated with part-of-speech tags, morphological features, lemmas, and some structural and functionally interpreted tags.

Although both LÄSBART and SUC were previously tagged, the corpora were annotated again for this study, in order to create a more uniformly annotated amount of text. For this, we used STAGGER (Östling, 2013), a part-of-speech tagger based on the averaged perceptron.

3.2 Algorithm

The alignment algorithm followed the procedure described in Sanchez-Perez et al. (2014), whose original purpose was to detect text reuse. By the use of a TF-IDF vector space model, the similarity between text fragments was calculated, with a slight modification: each sentence was considered a document, and the full collection of sentences in the original document was considered the document collection. Thus, rather than an inverse document frequency measure, an inverse sentence frequency was calculated. Each pair of text fragments was given a similarity score (cosine measure and Dice coefficient), and if the score

exceeded a certain threshold, originally 0.33, the text fragments were considered similar, and were thus aligned. The alignment was performed in two iterations, where the first iteration, followed the procedure given by Sanchez-Perez et al. (2014) with vectors based on lemmatised words, and the second iteration included part-of-speech tags as context information, in addition to the lemmatised lower cased words. The replicated method will henceforth be known as **M1**, and **M2** denotes the method using part-of-speech tags as context. By introducing the part-of-speech tags, we reasoned that we would improve the precision for disambiguating words and enable synonyms higher probability (Turney et al., 2010).

This algorithm was used to align one text segment originating from SUC with text segments originating from the LäSBarT corpus. The aim of this procedure was to construct the monolingual corpus consisting of reference segments (**RS**) aligned with easy-to-read segments (**ES**). The threshold of a cosine of 0.33, presented by Sanchez-Perez et al. (2014), was used as a minimum value for aligning a candidate segment. No maximum value was used, and a cosine of 1.0 was thus an admissible paraphrase candidate.

3.3 Features

To assess the algorithms' ability to produce more readable paraphrases, a number of readability measures can be used, see for instance Falkenjack et al. (2013). In this study we limit ourselves to using only a variety of commonly used readability measures, for Swedish.

N-gram overlap. An evaluation of two texts' similarity can be aided by the measure of shared *n*-grams between text pairs. In this evaluation, each paired segment, one RS paired to one ES, was treated as a collection of *n*-grams, with *n* ranging from 1 to 4. The proportion of intersecting *n*-grams in ES and RS was computed, and divided by the total number of *n*-grams in RS, resulting in a value ranging from 0 to 1. This value represents the overall mean *n*-gram overlap between the aligned segments. A value of 1 indicates that the ES is an exact copy of, or contained within, the RS. We used this measure to compare the overall *n*-gram overlaps between the two methods.

LIX, *readability index* (Björnsson, 1968), Equation 1. Ratio of words longer than 6 characters coupled with average sentence length. By computing this measure we were able to examine how the values differ from the original corpora's LIX, and gain a better understanding of the subset in comparison with the full set. These values were computed for M1 and M2 separately.

$$LIX = \frac{n(w)}{n(s)} + \left(\frac{n(words > 6\,chars)}{n(w)} \times 100\right) \tag{1}$$

where $n(s)$ denotes the number of sentences and $n(w)$ the number of words.

OVIX, *word variation index*, Equation 2. Originally developed by Hultman and Westman (1977) and related to type-token ratio. Logarithms are used to cancel out type-token ratio problems with varying text length. In this paper, OVIX was computed by treating the collection of aligned cluster sentences, originating from SUC, and the aligned text segments, originating from LäSBarT, as corpora. As for LIX, this measure is taken into account for evaluating the subsets of the two corpora.

$$OVIX = \frac{log(n(w))}{log\left(2 - \frac{log(n(uw))}{log(n(w))}\right)} \tag{2}$$

where $n(w)$ denotes the number of words and $n(uw)$ the number of unique words.

Length. Measures of long documents and words have been used in readability studies (Feng, 2010) and as a baseline for evaluating new features (Pitler and Nenkova, 2008). In this paper we computed the average word length as the average characters per word, the average number of long words per segment and the average number of words per segment. As for OVIX and LIX, the measures were computed by treating each subset of the corpora, containing only RS or only ES, as its own document.

Cosine similarity, Equation 3, calculates the cosine angle between two non-zero *n*-dimensional vectors, as the dot product of two vectors normalised by the product of the vector lengths. The cosine similarity measure ranges between -1 and 1, where a value closer to 1 indicates a high similarity between vectors. For the purpose of aligning paraphrases, we assumed that a cosine value of 1 is unwanted since a paraphrase is defined by being syntactically different at the same time as being semantically equivalent.

$$cos(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \qquad (3)$$

The **Dice coefficient**, Equation 4, is defined as the product of 2 times the number of features in common by the sum of the length of RS and the length of ES. For M1, using lemmatised lower case words as features and for M2 by also including the part-of-speech tags as features. This value ranges from 0 to 1 and represents the similarity of two segments, where a value closer to 1 means a higher similarity between the segments. As for the cosine similarity measure, we assumed that a Dice value of 1 is undesirable for this specific paraphrasing task.

$$Dice(A, B) = \frac{2\,|A \cap B|}{|A| + |B|} \qquad (4)$$

3.4 Crowdsourcing Evaluation

In this study, crowdsourcing was used to assess the two methods. The aligned items used for evaluation were chosen by only considering the RS that were paired with at least one ES at every cosine value (0.40, 0.50, 0.60, 0.70, 0.80), rounded to 2 decimals. When multiple sentences with the same cosine value were encountered, one was randomly chosen. The aim of this heuristic was to be able to better assess the cosine threshold values.

Typically, the recruitment process is managed by an employer, such as Amazon's Mechanical Turk. Since our evaluation task treated texts in Swedish, we constructed our own platform to host the tasks and collect the annotated data. For this project, the recruitment was made by public postings on Facebook and e-mails to current graduate students at a Swedish University. The recruitment post stated the aim of the tasks and contained a link to our web page. The web page presented the RS randomly to annotators followed by the (randomly ordered) aligned ES. Similarity judgements were made by rating the pair on a scale 0-4, corresponding to the following categories, as proposed in the Cross-Level Semantic Similarity Task of Semeval 2014 (Nakov and Zesch, 2014):

4. *The two items have very similar meanings and the most important ideas, concepts, or actions in the larger text are represented in the smaller text.*

3. *The two items share many of the same important ideas, concepts, or actions, but those expressed in the smaller text are similar but not identical to the most important in the larger text.*

2. *The two items have dissimilar meaning, but the shared concepts, ideas, and actions in the smaller text are related (but not similar) to those of the large text.*

1. *The two items describe dissimilar concepts, ideas and actions, but might be likely to be found together in a longer document on the same topic.*

0. *The two items do not mean the same thing and are not on the same topic.*

All categories were translated into Swedish. Even though Antoine et al. (2014) presented results that favour choosing 3 ordinal categories over 5, we believe that the nuances introduced by letting the crowdworkers annotate 5 categories will result in a richer understanding on how to choose the threshold value. The participants were not trained for the task, i.e. no example task was shown prior to the test. Studies have shown that training can render more precise results for individual workers (Le et al., 2010). As we had no experience on how the ranked categories relate to paraphrases, we chose to omit a training phase for the crowdworkers. There was no time restriction for the individual tasks, nor for the crowdsourcing test as a whole. When every aligned text item had been judged, the annotator was able to proceed. This process was repeated until reaching an end, where all aligned items had been given a ranking, or until leaving the annotation web page by choice.

4 Results

In this section, the preliminary results describing the performance of the alignment algorithm are presented and compared. As presented in Table 1, M2 resulted in higher arithmetic mean value for cosine as well as for Dice. The method produced 18,115 more aligned groups than did M1. As a result of more aligned groups, M2 contained about eight million more aligned text segments than M1. As for *n*-gram overlap, M1 had higher unigram overlap (0.31) than M2 (0.18), indicating that for M1, the ES is more alike the RS only considering the words, than of M2. The alignment clusters for both M1 and M2 were normally distributed with less occurrences towards the extreme values, both for cosine and Dice.

	M1	**M2**
Arithmetic mean Dice	0.47	0.50
Arithmetic mean cosine	0.49	0.50
Std deviation, Dice	0.13	0.13
Std deviation, cosine	0.09	0.10
Total number of aligned cluster	113,993	132,108
Total number of aligned text segments	9,294,015	17,422,338
Unigram overlap	0.31	0.18
Bigram overlap	0.05	0.04
Trigram overlap	0.01	0.01
Quadrigram overlap	0.0	0.00

Table 1: Comparison of M1 and M2 regarding descriptive features.

4.1 Alignment without context (M1)

This section presents results containing the shallow features measured for both sides of the alignment of the replicated method originally presented by Sanchez-Perez et al. (2014).

M1	SUC	LäSBarT	RS_{total}	ES_{total}	RS_{subset}	ES_{subset}
Tot. no of tokens	1,048,657	1,142,666	5,378,071	344,951,473	181	1,308
% of long words	26.55	18.03	0.09	0.14	0.05	0.02
No. of unique tokens	106,853	49,776	17,362	26,889	30	179
LIX	41.97	27.46	18.4	24.22	-*	-*
OVIX	90.9	68.83	49.22	50.04	31.93	29.33
No. of sentences	68,038	121,212	617,167	33,247,883	19	95
Avg. sentence length	15.41	9.43	10.37	8.71	6.20	6.77
Avg. word length	5.21	4.58	4.10	3.93	3.9	3.33
M2						
Tot. no. of tokens	1,048,657	1,142,666	2,455,941	419,768,883	252	1,377
% of long words	26.55	18.03	0.09	0.15	0.007	0.05
No. of unique tokens	106,853	49,776	15,369	21,871	50	221
LIX	41.97	27.46	24.93	16.67	-*	-*
OVIX	90.9	68.83	49.66	49.23	21.55	32.03
No. of sentences	68,038	121,212	314,399	41,658,621	25	125
Avg. sentence length	15.40	9.43	7.80	10.07	4.30	6.25
Avg. word length	5.21	4.58	3.50	3.88	3.30	3.80

* LIX is only applicable for documents rather than unique sentences

Table 2: Feature values of corpora, alignments and subsets for M1 and M2.

In Table 2, the shallow features of the original corpora are presented, as well as the corresponding values of the aligned total (RS_{total} and ES_{total}), and the subset later evaluated by crowdsourcing (RS_{subset}

and ES$_{subset}$). The readability metric, LIX, features a lower value for the RS as well as for the ES. The OVIX value for the RS compared to SUC is lower, as well as the OVIX value for the ES, when compared to LäSBarT.

The greater number of tokens and sentences is due to the fact that the total number of sentences contains multiple copies as a result of segments which are considered candidates in different alignment setups. The average sentence length and the average word length comparing the three; full corpora, aligned total and evaluated subset, show a converging tendency for RS and ES with respect to the readability metrics as well as for the length features.

4.2 Alignment with context (M2)

The descriptive results for M2 are found in the second section of Table 2, as well as the shallow features of the full original corpora, the aligned total (RS$_{total}$ and ES$_{total}$), and the evaluated subset (RS$_{subset}$ and ES$_{subset}$). Both readability metrics for RS are lower than for the corresponding original corpus. For LIX, this also applies for ES and LäSBarT. This converging tendency is also noticeable for M2, where the RS segments might be the most easy-to-read segments in the original corpora, and vice versa for ES. As with M1, M2 resulted in an occurrence distribution similar to M1, where a lesser frequency of aligned pairs is present, as a result of an extreme value.

4.3 Evaluation

Rank	RS	ES	Cosine	Dice	Method
0	Jag vet inte, sa jag. *I don't know, I said*	Jo, jag vet. *Yes, I know.*	0.50	0.57	M1
1	Ja, det är du, säger Oscar. *Yes, you are, says Oscar.*	– Ja, det säger alla! *– Yes, everybody says that!*	0.60	0.60	M1
2	Vad är det med dig? *What's up with you?*	– Vad är det som har hänt? *– What has happened?*	0.40	0.43	M1
3	Vad är det? *What is it?*	–Vad skulle det vara? *– What would that be?*	0.80	0.86	M1
4	Jag vet inte, sa jag. *I don't know, I said.*	Jag vet inte. *I don't know.*	0.70	0.55	M1
0	Majsan rycker på axlarna. *Majsan shrugs.*	Tanne ryckte tag i henne. *Tanne grabbed her.*	0.40	0.36	M2
1	Vad menar du? *What do you mean?*	Vad hade han gjort dig? *What did he do to you?*	0.60	0.60	M2
2	undrar jag. *I wonder*	undrade hon *she wondered*	0.80	0.67	M2
3	Varför då? *Why?*	Jaså, varför det? *Oh, really? Why?*	0.50	0.44	M2
4	Det visste jag. *I knew that.*	Jag visste det, fan jag visste det. *I knew it, damn, i knew it.*	0.80	0.80	M2

Table 3: Examples sentences per rank category for sentences aligned by M1 and M2.

The heuristic rendered 220 aligned items to evaluate by crowdsourcing, each containing one RS and one ES from each cosine value. 95 of the aligned items were from M1 and the remaining 125 were from M2. Table 3 presents some examples from the alignment clusters which have been evaluated for each method. The alignments are presented with a typical rank that annotators have been giving them. The examples illustrate how the aligned sentences are loosely coupled by the words present in the sentences. The preliminary results of the mapping of cosine similarity and participant ranking is presented in Figure 1, with cosine values divided into intervals. There is a tendency of M2 scoring consistently higher than M1.

The preliminary results of the mapping of Dice similarity and participant ranking is presented in Figure 2, with Dice values divided into intervals. The results of the first and last intervals clearly differ from the remaining intervals due to skewed data, and can in this context be considered outliers. From interval 0.31–0.40 to interval 0.71–0.80, the Dice similarity seems to stabilise for values over 0.5. These prelim-
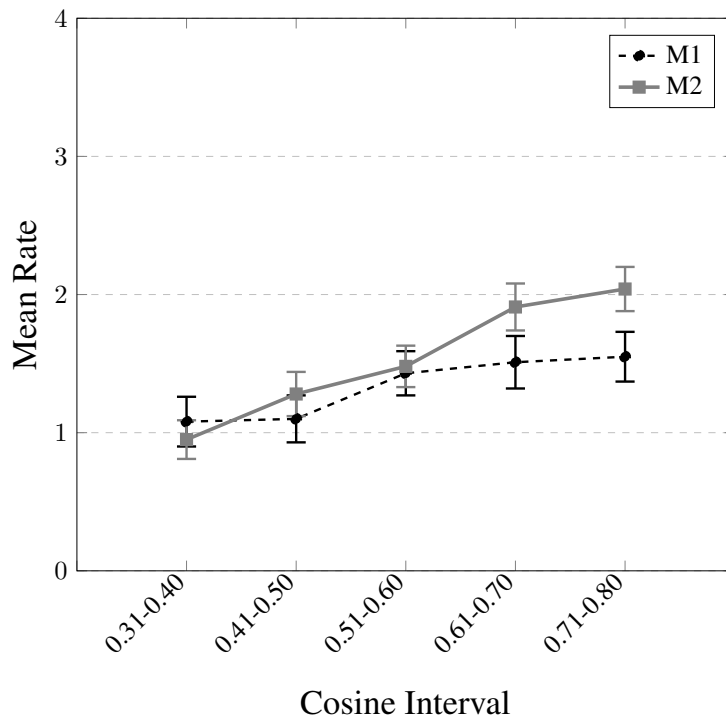
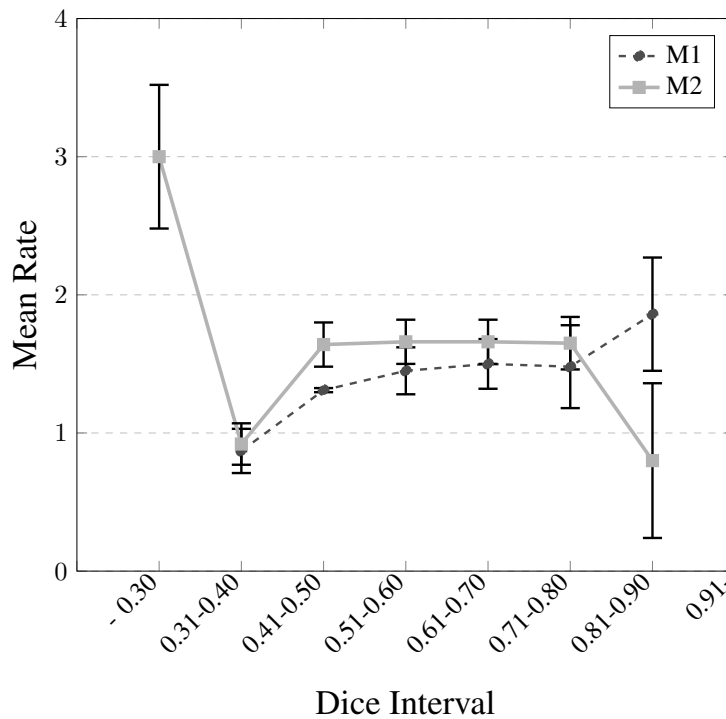Figure 1: Mean rate for M1 and M2 over different cosine intervals

†

Figure 2: Mean rate for M1 and M2 over different Dice intervals

inary results propose that M2 provides a more stable relationship between the human ranked categories and the similarity measures.

5   Discussion

As the sentences in Table 3 imply, the easy-to-read segments seem syntactically alike and faithful to the reference segments on a word level. This is not always desirable for the task of extracting paraphrases. Extracting paraphrases that are dissimilar by their word usage, but semantically equivalent, enables us to achieve data that better fit the definition of paraphrases. This will be addressed in future work.

As seen in Figues 1 and 2, M2 seems to relate better to human ranking than M1. We will further explore this in a more thorough statistical analysis of a larger amount of data when available.

It is still not clear whether or not there is a limit as to how large either side of the segment should be, i.e. that a larger chunk of ES could be aligned with a smaller chunk of RS, or vice versa. The algorithm allows the aligned chunks to be expanded beyond sentence boundaries, meaning that the lengths of the resulting segments are only limited by the length of the entire document.

We presented the relationships between the annotated rankings and the cosine and Dice groups. These show trends which we believe can be used to accept a cosine higher than 0.33 as the threshold for deciding which segments to treat as candidate paraphrases.

In designing our annotator task, we chose to let annotators be new to the tasks in the sense that the crowdworkers had no intuition about how we would rate a pair of text segments. Thus, they were not primed to any systematic approach for solving the tasks, contrary to previous studies (Le et al., 2010). This has probable explanation effects on the variance in the annotated data and the possible disagreement the annotators seem to display. But it is possible that these results are only preliminary and that the gathering of more data would result in higher agreement across the categories. It would be interesting to study the alignments that people had more difficulty to agree upon, since that analysis itself could have an effect on the way we understand features of readability in text or intra class differences in peoples' ability to read and understand text segments. It might prove to be necessary to divide crowdworkers into groups based on some training task, measuring reading ability. For a language such as Swedish, there is no available platform to invite human annotators, nor is it possible to give crowdworkers any payment. There are multiple ways in which NLP can be helped by the work of human annotators, and we perceive this as an opportunity to develop and distribute a platform that could be used as portal for NLP scientists for collecting data based on peoples language skills.

6   Conclusions and future work

Data collection is ongoing, but preliminary conclusions of this study are that 1) the Dice coefficient does not seem to correspond well with human ranking, 2) the method with part-of-speech tags included (M2) provides a more stable relationship between the human ranked categories and the similarity measures and, 3) the framework developed for this study proved to be an effective tool for collecting data when crowdsourcing human rankings for a NLP task.

In this study we assumed that the corpus data would entail that the ES was easy-to-read based on the typology of the data. Future work will try to validate the readability of the aligned ES to assure that an ES is easier than its corresponding RS. At the time of writing, the data collection is ongoing, and final results from the evaluation, including a thorough statistical analysis, will be presented in future work. From the evaluation, a cosine threshold value will be estimated. The final goal is an aligned monolingual corpus, from which it is possible to deduce patterns and operations for the purpose of automatic text simplification.

Acknowledgements

References

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation.

Regina Barzilay and Noemie Elhadad. 2003. Sentence Alignment for Monolingual Comparable Corpora. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Carl Hugo Björnsson. 1968. *Läsbarhet[Readability]*. Liber, Stockholm.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. *Proceeding HLT-NAACL '06 Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24.

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.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic Induction of Rules for Text Simplification. *Knowledge-Based Systems*, 10(3):183–190.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A New Text Simplification Task. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, page Article No. 350.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway*, NEALT Proceedings Series 16.

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York.

Amir Hazem and Emmanuel Morin. 2016. Improving bilingual terminology extraction from comparable corpora via multiple word-space models. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Katarina Heimann Mühlenbock. 2013. *I see what you mean. Assessing readability for specific target groups*. Dissertation, Språkbanken, Dept of Swedish, University of Gothenburg.

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

Hongyan Jin. 2002. Using hidden markov modeling to decompose human-written summaries. *Computational Linguistics*, 28(4):527–543.

Gunnel Källgren, Sofia Gustafson-Capková, and Britt Hartmann. 2006. Manual of the stockholm umeå corpus version 2.0.

Sigrid Klerke and Anders Søgaard. 2012. Dsim, a danish parallel corpus for text simplification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),Istanbul, Turkey*. European Language Resources Association (ELRA).

Kevin Knight and Daniel Marcu. 2000. Statistics-Based Summarization - Step One: Sentence Compression. *AAAI-00 - 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages 703–710.

John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 21–26.

Katarina Mühlenbock and Sofie Johansson Kokkinakis. 2009. LIX 68 revisited - An extended readability measure. In Michaela Mahlberg, Victorina González-Díaz, and Catherine Smith, editors, *Proceedings of the Corpus Linguistics Conference CL2009*, Liverpool, UK, July 20-23.

Katarina Mühlenbock. 2008. Readable, Legible or Plain Words – Presentation of an easy-to-read Swedish corpus. In Anju Saxena and Åke Viberg, editors, *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume 8 of *Acta Universitatis Upsaliensis*, pages 327–329, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Preslav Nakov and Torsten Zesch, editors. 2014. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Rani Nelken and Stuart M Shieber. 2006. Towards Robust Context-Sensitive Sentence Alignment for Monolingual Corpora. *Eacl*, pages 161–168.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, HI, October.

Evelina Rennes and Arne Jönsson. 2015. A tool for automatic simplification of swedish texts,. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa-2015), Vilnius, Lithuania,*.

Miguel A. Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh. 2014. The winning approach to text alignment for text reuse detection at PAN 2014: Notebook for PAN at CLEF 2014. *CEUR Workshop Proceedings*, 1180:1004–1011.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. *23rd International Conference on Computational Linguistics*, (August):1353–1361.