Can automatic cohesion measures improve the readability of summaries by reordering their sentences?

Alfred Sjöqvist, Daniel Tufvesson, Isak Wanström, Karin Stendahl,
Linus Tullstedt, Ludvig Rammus & Sofia Davidsson

Abstract

Extractive text summarization models often generate summaries with fragmented and incohesive sentences. This paper investigates the application of various methods of reordering the sentences in the generated summaries to improve their readability. We developed an application, CohSort, that reorders sentences to maximize the cohesion between sentences in the text according to an aggregate of the LSA-indexes and the L2 readability index in Coh-Metrix. By conducting self-report surveys asking participants to compare original summaries with reordered ones, we found that these reordered summaries were less readable than the originally ordered. This suggests that simply maximizing sentence-to-sentence cohesion for a text does not make it more readable. Further, it suggests that the used cohesion measures do not fully capture readability as a phenomenon.

1. Introduction

In an increasingly information-driven world, the ability to access and understand written information is pivotal. However, this process may pose significant challenges for individuals with reading difficulties, such as dyslexia. This reality underscores the urgent need for innovative solutions to enhance the readability and accessibility of written information.

The Text Adaptation for Increased Reading Comprehension (TextAD) research project at Linköping University, through its iteration ElsaSum (Andersson, 2022), is aiming to develop such a tool, optimizing the readability of the extracted summaries. In this context, readability refers to the ease with which a reader can understand a written text. It is influenced by various factors including the complexity of the language, the structure of the text, the reader's prior knowledge, and, crucially, the cohesion of the text (Graesser et al., 2004; Linderholm et al., 2000). Cohesion is an objective property inherent to the text itself. It involves explicit linguistic features that guide the reader towards constructing coherent mental representations of the content. These features include words, phrases, and sentences that interconnect to create a unified, cohesive whole. Given its objectivity and inherent text property, cohesion is particularly relevant for computational approaches to text adaptation.

1.1 Purpose

The research presented in this paper aims to explore how reordering the sentences of summaries generated by ElsaSum can improve the readability of the text. In this work we present a new application, CohSort, which reorders sentences of a summary in order to maximize the sentence-to-sentence cohesion. To reorder sentences, CohSort uses Latent Semantic Analysis (LSA) with Sentence-BERT (SBERT), and an implementation of the L2 Reading Index. Both are measures within the Coh-Metrix framework (Crossley, Greenfield, and McNamara, 2008) a set of computational indexes for measuring text cohesion and difficulty. Coh-Metrix provides a detailed analysis that goes beyond just surface-level linguistic features. It delves into cognitive reading processes such as understanding, decoding, and syntactic parsing.

The results are assessed through an online survey, where participants compare the readability of summaries generated by CohSort to those of ElsaSum. This is subsequently analyzed and evaluated through statistical tests. A technical evaluation of the application is carried out as well.

2. CohSort

We introduce CohSort and its implementation in this section. CohSort takes a summarized Swedish text as input and produces an output text, where the sentences have been reordered to

maximize their sentence-to-sentence cohesion. This order is determined from a cohesion score consisting of the average of a set of Coh-Metrix indexes. These indexes are the L2 Reading Index, the LSA Adjacent Sentence indexes, and the LSA Givenness indexes. CohSort finds the order with the highest score by permuting the order of sentences using simulated annealing.

2.1. L2 Reading Index

The Coh-Metrix L2 Reading Index is a readability formula that focuses on the cognitive and psycholinguistic dimensions of reading. The formula takes into account both the content and the complexity of a text, utilizing a formula that combines three Coh-Metrix indexes: Content Word Overlap, Sentence Syntax Similarity, and CELEX Word Frequency (Crossley et al., 2008; McNamara et al., 2014, p. 80–81).

Content Word Overlap denotes the ratio of shared content words between adjacent sentences. This index uses the POS-tag of each parsed word to identify content words.

Sentence Syntax Similarity illustrates the average syntactic similarity between neighboring sentences. This was calculated through parsing constituency trees using Benepar. For tree analysis, we used the NetworkX library for Python, which offers tools for constructing and analyzing graphs.

For our study we excluded the CELEX Word Frequency index since it is only affected by the presence of certain words in a text, and not the order of sentences in the text.

2.2. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method for analyzing the semantics of written text using word embeddings, vector representations of words computed based on statistical co-occurrence with other words in the training data (Landauer, 2013). This makes the vector space of the embeddings semantically meaningful, where embeddings that are located close to each other often are semantically similar.

For our study we used the LSA indexes in Coh-Metrix that operate on sentences (McNamara et al., 2014, p. 66–67). We used the LSA Adjacent Sentence indexes (LSASS1, LSASS1d) which measure how similar adjacent sentences are to each other, where higher similarity means higher cohesion. We also used the LSA Givenness indexes (LSAGN, LSAGNd) which measure how much new information is contained in each sentence in relation to previous sentences. Less new information means higher cohesion.

2.3. SBERT for Sentence Embeddings

LSA typically utilizes Singular Value Decomposition (SVD) to compute embeddings. In CohSort, Sentence-BERT (SBERT), a modification of BERT, was employed to compute sentence embeddings. SBERT, through its use of siamese and triplet network structures, can produce such embeddings while reducing computational effort and time. The application of SBERT is particularly beneficial for tasks involving comparison of large-scale semantics similarity, clustering, and information retrieval (Reimers and Gurevych, 2019).

The implementation of LSA-indexes in this study involved using a Swedish SBERT model for sentence embeddings, with embeddings computed beforehand and used multiple times in the LSA-indexes. All sentences were analyzed in their original word form, without lemmatization.

3. Study Design

To evaluate if reordering sentences using CohSort improves the readability, we conducted both a technical evaluation using computational measures, as well as an online survey letting human participants evaluate generated and reordered summaries.

3.1. Summaries

As a basis for the summaries, 15 articles from the Swedish newspaper Dagens Nyheter (DN), were selected and retrieved from the Mediearkivet database. These articles ranged between 300 and 400 words and covered neutral topics to avoid strong emotions of participants.

These articles were summarized with ElsaSum down to eight sentences. The summaries were then reordered with CohSort. This resulted in two sets of summaries, one of ElsaSum summaries and one of CohSort summaries.

3.2. Technical evaluation

The technical evaluation was conducted using SAPIS (Fahlborg and Rennes, 2016), looking

at the differences of the top 8 most frequently changed metrics between the ElsaSum and CohSort summaries. All of the 15 retrieved articles were utilized for summaries in this process.

3.3. Survey

To determine if people find that reordering sentences improves readability, we conducted an online survey, where the participants compared ElsaSum summaries with CohSort summaries.

The survey was answered by 22 participants, 18 of which were university students, three with a university degree, and one with a completed highschool education. Their average age was 24.5 years old, with 41% identifying as female and 59% as male. These participants were selected using convenience sampling.

The survey consisted of summaries from three of the retrieved articles; we call these Article 1, 2, and 3. The summaries were arranged in corresponding pairs: the ElsaSum summary with its corresponding CohSort summary. Participants were asked to rate the summaries based on perceived coherency, ease of reading, and ease of understanding the information. Each summary had the following questions:

(a) How coherent was the text?

(b) How easy was the text to read?

(c) How easy was it to understand the information in the text?

The questions were assessed using a six-point scale. This was done to each summary independently before the participants got to compare them directly, displaying each summary pair simultaneously. In the direct comparison, the following questions were asked:

(d) Which text was the easiest to read?

(e) Which text was the most coherent?

(f) Which text was the easiest to understand?

These questions were assessed by asking the participants to choose between text one, two, or that they performed equally.

3.4. Survey Analysis

A statistical analysis was performed on the data collected from the survey, where each question represented certain properties we wanted to measure. Questions (a) and (e) were intended to represent coherency; (b) and (d) ease of reading the text; (c) and (f) ease of understanding the information in the text. The mean value of these properties combined were calculated, and further analyzed using paired samples t-test. Skewness and kurtosis was measured for understanding deviations from normal distribution.

4. Results

When comparing participants' perceived coherency, reading ease, and information ease of ElsaSum and CohSort summaries, significant differences were found between the mean values of the ElsaSum and CohSort summaries of Article 1 (ElsaSum: $M = 4.21$; CohSort: $M = 3.35$), as shown by a paired t-test, $t(20) = 2.82$, $p = 0.011$, and Article 2 (ElsaSum: $M = 4.52$; CohSort: $M = 4.00$), as shown by a paired t-test, $t(20) = 2.11$, $p = 0.048$. However, no significant difference was found for summaries of Article 3 (ElsaSum: $M = 4.78$; CohSort: $M = 4.32$) as shown by a paired t-test, $t(20) = 1.65$, $p = 0.114$. These differences are illustrated by Figure 1, 2, and 3.

Minimal skewness was present in regards to the summaries of Article 1 by both ElsaSum and CohSort, with no relevant deviation from normal distribution (mesokurtic). For the Article 2 summaries, ElsaSum coherency was skewed (-1.13) and leptokurtic (1.73), with CohSort being less skewed (-0.686) and mesokurtic (-0.339). For the Article 3 summaries, ElsaSum was notably skewed in relation to information ease (-1.46) and platykurtic (2.25), with CohSort being less skewed (-0.995) and mesokurtic (0.458), showing no major deviation from normal distribution.

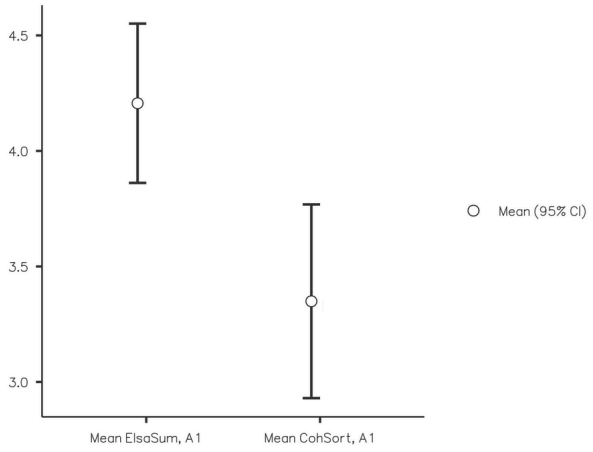

Figure 1: Mean readability for summaries of Article 1. Significant difference.

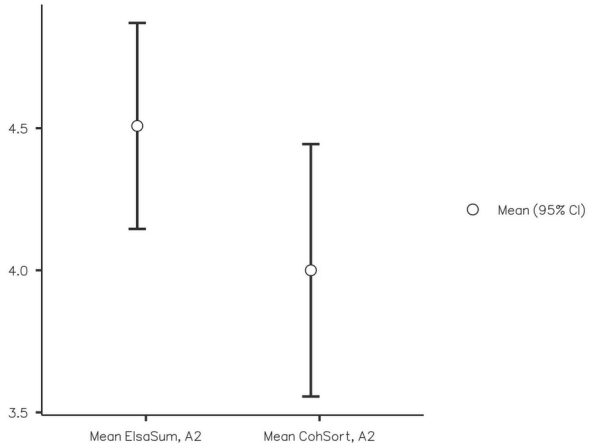

Figure 2: Mean readability for summaries of Article 2. Significant difference.

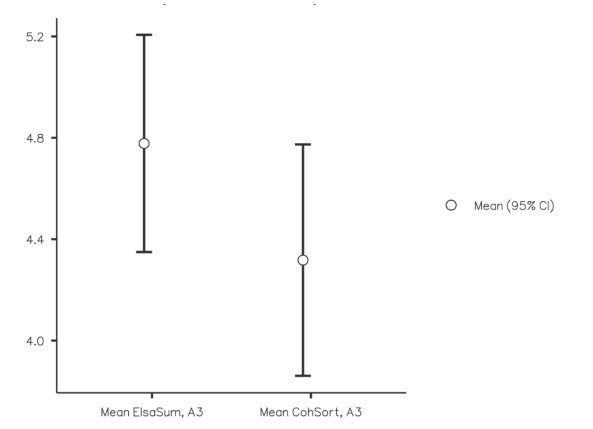

Figure 3: Mean readability for summaries of Article 3. No significant difference.

The technical analysis revealed a total of 68 metrics that had changed between the ElsaSum summaries and the CohSort summaries. Among these, the top 8 (most frequently changed) metrics saw an average increase of 25.779% for CohSort as compared to ElsaSum. The largest difference was observed in the metric for measuring the ratio of content words in adjacent sentences, with a percentage increase of 54.59% for CohSort. On the other hand, the metric measuring the ratio of content words globally in the summary showed the smallest change, with a percentage difference of 0.97%. This was the only percentage difference among the top 8 frequently changed metrics which represented a negative change from the ElsaSum summary and the CohSort summary. The percentage changes are illustrated in Table 1.

Changed metric	**Change**
Coh-Metrix, LSA - Adjacent sentences: Average	1.784%
Coh-Metrix, Cohesion - Adjacent sentences, Content Words: Standard deviation	27.407%
Coh-Metrix,Cohesion - Global, Anaphors	2.380%
Coh-Metrix, LSA - Adjacent Sentences: Standard deviation	28.725%
Coh-Metrix, Cohesion - Adjacent Sentences, Content Words: Ratio	54.597%
Coh-Metrix, Cohesion - Adjacent Sentences, Stems	42.307%
Coh-Metrix, Cohesion - Adjacent Sentences, Arguments	50%
Coh-Metrix, Cohesion - Global, Content Words: Ratio	-0.968%

Table 1: Technical evaluation results. Change refers to CohSort, relative to ElsaSum.

5. Discussion

The technical evaluation suggests higher readability for the reordered summaries generated by CohSort, while the surveys suggest lower readability, instead favoring ElsaSum.

Readability is ultimately dependent on the reader. This therefore suggests that the cohesion measures do not fully capture readability as a phenomenon. This is not to say that cohesion measures are useless and that they do not capture readability at all. It merely suggests that there is more to readability than what the measures encompass, and that these limitations become apparent when one tries to maximize these measures as we have done in this study.

Our results may be limited by the use of self-report surveys of perceived readability, making it prone to cognitive biases. Another more suitable approach would perhaps be a reading comprehension test, where a more readable text would yield a higher understanding than a less readable one. Another approach would have been using eye-tracking technology, where longer fixations and rereading of sentences would suggest lower readability. Both these approaches would yield more objective results than subjective self-reports.

The indexes we used in CohSort operate only on sentences, and this may be the reason readability suffers when maximizing these cohesion measures. We therefore suggest combining these sentence-level measures with other measures that operate on higher levels. An example is the LSAPP1 index in Coh-Metrix (McNamara et al., 2014, p. 66) which computes the semantic similarities between adjacent paragraphs. We suggest that sentences within paragraphs could be sorted according to our CohSort approach, while these paragraphs are then sorted using LSAPP1 index. If this improves readability remains to be seen.

6. Conclusion

Our results indicate that maximizing the cohesion of a text does not necessarily make it more readable. We found that, while the CohSort summaries fairly consistently rank higher than the ElsaSum summaries in computed linguistic metrics, readers don't actually find them more readable. This indicates a discrepancy between the computationally measurable linguistics and actual reader experience.

On the other hand, while simply maximizing cohesion is not a viable approach, it does, however, not leave out the possibility of increasing readability via more sophisticated uses of the cohesion measures.

7. Bibliography

Andersson, Elsa. 2022. *Methods for increasing cohesion in automatically extracted summaries of Swedish news articles: Using and extending multilingual sentence transformers in the data-processing stage of training BERT models for extractive text summarization*. B.A. Thesis, Linköping University. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1667268

Crossley, Scott A.; Greenfield, Jerry. and McNamara, Danielle S. 2008. Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly* 42(3): 475-493. http://www.jstor.org/stable/40264479

Graesser, Arthur C.; McNamara, Danielle S.; Louwerse, Max M. and Cai, Zhiqiang. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36(2): 193-202. https://doi.org/10.3758/BF03195564

Fahlborg, Daniel. and Rennes, Evelina. 2016. Introducing SAPIS – an API Service for Text Analysis and Simplification. In *The second national Swe-Clarin workshop: Research collaborations for the digital age*. Umeå, Sweden. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-141550 (Accessed 2023-05-22).

Landauer, Thomas K. 2013. LSA as a Theory of Meaning. In Landauer, Thomas K.; McNamara, Danielle S.; Dennis, Simon. and Kintsch, Walter (eds.). *Handbook of Latent Semantic Analysis*. Routledge.

Linderholm, Tracy; Everson, Michelle G.; van den Broek, Paul; Mischinski, Maureen; Crittenden, Alex. and Samuels, Jay. 2000. Effects of Causal Text Revisions on More- and Less-Skilled Readers' Comprehension of Easy and Difficult Texts. *Cognition and Instruction* 18(4): 525-556. https://doi.org/10.1207/S1532690XCI1804_4

McNamara, Danielle S.; Graesser, Arthur C.; McCarthy, Philip M. and Cai, Zhiqiang. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Reimers, Nils. and Gurevych, Iryna. 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* https://arxiv.org/abs/1908.10084 (Accessed 2023-05-20).