Text Summarization using Random Indexing and PageRank

Pär Gustavsson, Arne Jönsson

Department of Computer and Information Science, Santa Anna IT Research Institute AB
Linköping University, SE-581 83, LINKÖPING, SWEDEN
pargu814@student.liu.se, arnjo@ida.liu.se

Abstract

We present results from evaluations of an automatic text summarization technique that uses a combination of Random Indexing and PageRank. In our experiments we use two types of texts: news paper texts and government texts. Our results show that text type as well as other aspects of texts of the same type influence the performance. Combining PageRank and Random Indexing provides the best results on government texts. Adapting a text summarizer for a particular genre can improve text summarization.

1. Introduction

CogSum (Jönsson et al., 2008) is a tool for creating extraction based text summaries based on the vector space technique Random Indexing. To further improve sentence ranking CogSum also uses PageRank (Brin and Page, 1998). To use PageRank we create a graph where a vertex depicts a sentence in the current text and an edge between two different vertices is assigned a weight that depicts how similar these are, by a cosine angle comparison. Sentences with similar content will then contribute with positive support to each other. This effect doesn't exclusively depend on the number of sentences supporting a sentence, but also on the rank of the linking sentences. This means that a few high-ranked sentences provide bigger support than a large number of low-ranked sentences. This leads to a ranking of the sentences by their importance to the document at hand and thus to a summary including only the most important sentences.

2. Experiment

To evaluate CogSum for text summarization on various text types, two studies were performed. The first compared summaries created by CogSum with or without PageRank activated. This study was conducted on news texts and we used another summarizer, SweSum (Dalianis, 2000), as a baseline. SweSum is basically a word based summarizer but with additional features such as letting users add keywords, extracting abbreviations and having a morphological analysis. SweSum has been been tailored to news texts in various ways, e.g. by increasing the probability to include the first sentences in an article in the summary.

The created summaries were compared to existing gold standards in the KTH eXtract Corpus (KTHxc) by an overlap measure on sentence level (Hassel and Dalianis, 2005). We used 10 Swedish news texts with an average length of 338 words.

The second study was conducted to compare summaries created by the same systems but with other texts, namely 5 fact sheets from the the Swedish Social Insurance Administration (Sw. Försäkringskassan). The length of the fact sheets ranged from 1000 to 1300 words. The gold standards for these texts were created by Carlsson (2009). The evaluation for this experiment was conducted in AutoSummENG, by means of the metric Graph Value Similarity (Giannakopoulos et al., 2008), as this allows taking content similarity between different sentences into consideration during the evaluation.

The Random Indexing dimensionality was kept constant to 100 through the first experiment, as done previously by Chatterjee and Mohan (2007) on texts of equal length. Different dimensionalities ranging from 100 to 1000 were initially used in the second study as these texts were longer on average. The summaries created in the second study were more or less identical, especially the ones with a dimensionality of 500 and upwards. Results from previous studies imply that as low dimensionality as possible is desirable to deal with time and memory usage while it's unimportant to optimize the variable because of the small difference between the created summaries (Sjöbergh, 2006). With this in mind a dimensionality of 500 was used for the second study.

3. Results

Text	CogSum	CogSumPR	SweSum
Text001	85.71	85.71	85.71
Text002	30.00	9.09	38.10
Text003	20.00	0.00	80.00
Text004	57.14	54.54	52.63
Text005	70.59	35.29	66.67
Text006	66.67	66.67	50.00
Text007	50.00	50.00	85.71
Text008	42.86	66.67	50.00
Text009	40.00	37.50	70.59
Text010	28.57	33.33	66.67
Average	49.15	43.88	64.61

Table 1: Sentence overlap on news texts (%)

Table 1 shows results from the first study for the summaries created by CogSum with or without PageRank and SweSum for 10 news texts from the KTHxc corpus. The table shows the overlap on sentence level compared to the gold standards expressed in percentage. We can see that SweSum gained the highest average sentence overlap of 64.61% followed by CogSum (49.15%) and CogSumPR (43.88%).

The results from the second study, where we use government texts are presented in Table 2. The table shows the N-gram Value Similarity between the created summaries and

the gold standards. The value of this metric ranges from 0 to 1.

Text	CogSum	CogSumPR	SweSum
Text001	0.532	0.491	0.227
Text002	0.284	0.356	0.353
Text003	0.416	0.443	0.293
Text004	0.292	0.383	0.168
Text005	0.370	0.342	0.246
Average	0.379	0.403	0.258

Table 2: Graph Value Similarity on government texts

As shown in Table 2 the summaries created by CogSumPR gained the highest average value of 0.403 followed by CogSum (0.379) and SweSum (0.258).

To further investigate the various evaluation metrics used in our study, we evaluated the news paper texts, i.e. the first experiment, using AutoSummENG.

Graph Value	CogSum	CogSumPR	SweSum
Average	0.526	0.516	0.584

Table 3: Graph Value Similarity on news texts

Table 3 presents the results, and as can be seen they are consistent with the first study as the systems get ranked in the same order as they did when ranked according to sentence overlap, c.f. Table 1.

4. Discussion

The results of the first study showed that SweSum achieved the best results. This is not surprising as this system is tailored to summarize news texts. The results for CogSum and CogSumPR were equal for most of the texts in the corpus with a slight advantage for CogSum. One particularly interesting result is the one for Text003 where SweSum got an 80% overlap while CogSum gained 20% and CogSumPR 0%, which call for further analysis in the future to be properly explained. It was hard to draw any definite conclusions from this data and the possibility that CogSum performed better than CogSumPR by chance exists. Still, it's possible that Random Indexing works well as it is and that the incorporation of a PageRank algorithm doesn't improve the created summaries.

The second study revealed that the summaries created by CogSum with PageRank activated were closest to the gold standards which means that they were created by a better system. This is only the case for the 5 texts used in this study and a larger evaluation would strengthen the reliability of the study. The results showed that CogSum with and without PageRank performed relatively equal results for all of the texts which indicates that the two systems gained an accurate "understanding" of all of them. The fact that the activation of PageRank led to a better average result for these five fact sheets thus suggest that this version of the summarizer may be preferable for this kind of texts in general. No statistical significance testing was conducted in either study due to the fairly small number of texts used, but further studies involving a larger amount of texts are close at hand.

One possible explanation to the results could be properties of the texts. The fact sheets were longer than the news texts. It is possible that PageRank works better for texts with more sentences as a larger number of sentences can be used to strengthen the mutual effect. Another possible explanation is the structure of the texts used in the two studies. The fact sheets aim to contribute with as much information as possible regarding a certain topic and thus have a fair number of headings. The news texts on the other hand only include a main header and read up on a news item with the most important information presented in the beginning of the text.

The evaluations were done automatically with no qualitative input from people in potential target groups. Although humans were involved in the creation of the gold standards and thus affected the results indirectly, no information regarding readability or usefulness of the summaries were collected. The results only show how different extraction techniques mimic human choice of extraction units.

Acknowledgment

This research is financed by Santa Anna IT Research Institute AB.

5. References

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Bertil Carlsson. 2009. Guldstandarder – dess skapande och utvärdering. Master's thesis, Linköping University.

Nilhadri Chatterjee and Shiwali Mohan. 2007. Extraction-based single-document summarization using random indexing. In *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence – (ICTAI 2007)*, pages 448–455.

Hercules Dalianis. 2000. Swesum – a text summarizer for swedish. Technical Report TRITA-NA-P0015, IPLab-174, NADA, KTH, Sweden.

George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech Language Processing*, 5(3):1–39.

Martin Hassel and Hercules Dalianis. 2005. Generation of Reference Summaries. In *Proceedings of 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland, April 21-23.

Arne Jönsson, Mimi Axelsson, Erica Bergenholm, Bertil Carlsson, Gro Dahlbom, Pär Gustavsson, Jonas Rybing, and Christian Smith. 2008. Skim reading of audio information. In *Proceedings of the The second Swedish Language Technology Conference (SLTC-08), Stockholm, Sweden*.

Jonas Sjöbergh. 2006. *Language Technology for the Lazy - Avoiding Work by Using Statistics and Machine Learning*. Ph.D. thesis, KTH, Stockholm, Sweden.