LSA-Based Multi-Document Summarization

J. Steinberger[1], M. Křišťan[1]

[1] Text Mining Group,
Dept. of Computer Science and Engineering,
University of West Bohemia in Plzeň, Czech Republic

I. Introduction

Text Summarization is a research area that attracts many research groups around the world. Its aim is to take a source text and present the most important content in a condensed form in a manner sensitive to the needs of the further task. Being able to produce summaries that would be close to human abstracts would affect many tasks in information retrieval. We developed a single-document summarization method [1] that is based on Latent Semantic Analysis (LSA - [2]). The analysis captures main topics of a document. The sentences that contain the most important topics are selected for the summary. Further, we enriched the method by anaphoric information [3]. In the last years, the summarization field has turned to a more compex task - multi-document summarization. It assumes that we have a set of documents related to a specific event/topic. The task is to create a short document that would summarize the event/topic. Moreover, the summary can be influenced by a user query.

When turning to multi-document summarization some new problems arise. For example, we do not want in the summary very similar sentences from different documents. Our solutions of these problems are discussed in the paper.

The structure of the paper is as follows. In Section II, some background information is presented. Then, the single-document LSA-based method is described and its extension to multi-dcoument summarization is proposed. In section IV, we firstly describe data and evaluation process and then the own results are discussed. In the end we conclude all.

II. Multi-Document Summarization

Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary allows users to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are complementing the news aggregators performing the next step down the road of coping with information overload [4].

Multi-document summarization creates information reports that are both concise and comprehensive. With different opinions being put together and outlined, every topic is described from multiple perspectives within a single document. While the goal of a brief summary is to simplify information search and cut the time by pointing to the most relevant source documents, comprehensive multi-document summary should itself contain the required information, hence limiting the need for accessing original files to cases when refinement is required. Automatic summaries present information extracted from multiple sources algorithmically, without any editorial touch or subjective human intervention, thus making it completely unbiased.

The multi-document summarization task has turned out to be much more complex than summarizing a single document, even a very large one. This is evidently due to inevitable thematic diversity within a large documents set. A good summarization technology aims at combination of the main theme compliance and completeness, good readability and conciseness. Document Understanding Confer-

ences (DUC), conducted annually by NIST (National Institute of Standards and Technology), have developed sophisticated evaluation criteria for techniques accepting the multi-document summarization challenge.

An ideal multi-document summarization system not just shortens the source texts but presents information organized around the key aspects so as the wider diversity of views on the topic. When such quality is achieved, an automatic multi-document summary is perceived more like an overview of a given news topic. The latter implies that such text compilations should also meet other basic requirements for an overview text compiled by a human.

The multi-document summary quality criteria are firstly *clear structure*, including an outline of the main contents items, from which it is easy to navigate to the full text sections; secondly text within sections is divided into *meaningful paragraphs*; then *gradual transition* from more general to more particular thematic aspects; and finally *good readability*, where the automatic overview should show no unrelated information noise,no dangling references to what is not mentioned or explained in the overview, no text breaks across a sentence, and no semantic redundancy.

III. Our LSA-Based Method

We based our work on our LSA-based single-document summarization method [1] and we extended it to work with a set of documents. The extension is proposed in this paper.

1. Single-Document LSA-Based Method

Our approach to summarization follows what has been called a term-based approach [5]. In term-based summarization, the most important information in a document is found by identifying its main 'terms' (or 'topics' - combinations of terms), and then extracting from the document the most important information about these terms/topics.

LSA is a fully automatic mathematical/statistical technique for extracting and representing the contextual usage of words' meanings in passages of discourse. The basic idea is that the aggregate of all the word contexts in which a given word does and does not appear provides mutual constraints that determine the similarity of meanings of words and sets of words to each other. LSA has been used in a variety of applications (e.g., information retrieval, document categorization, information filtering, and text summarization).

The heart of the analysis in summarization background is a document representation developed in two steps. The first step is the creation of a term by sentence matrix, where each column represents the weighted term-frequency vector of a sentence in the set of documents under consideration. The terms from a user query get higher weight. The next step is to apply Singular Value Decomposition (SVD) to matrix A:

$$A = U\Sigma V^T \tag{1}$$

From an NLP perspective, what SVD does is to derive the *latent semantic structure* of the document represented by matrix A: i.e. a breakdown of the original document into r linearly-independent base vectors which express the main 'topics' of the document. SVD can capture interrelationships among terms, so that terms and sentences can be clustered on a 'semantic' basis rather than on the basis of words only. Furthermore, as demonstrated in [6], if a word combination pattern is salient and recurring in document, this pattern will be captured and represented by one of the singular vectors. The magnitude of the corresponding singular value indicates the importance degree of this pattern within the document. Any sentences containing this word combination pattern will be projected along this singular vector, and the sentence that best represents this pattern will have the largest index value with this vector. Assuming that each particular word combination pattern describes a certain topic in

the document, each singular vector can be viewed as representing such a topic [7], the magnitude of its singular value representing the degree of importance of this topic.

The method selects for the summary the sentences whose vectorial representation in the matrix $\Sigma^2 \cdot V^T$ has the greatest 'length'. Intuitively, the idea is to choose the sentences with greatest combined weight across all important topics.

2. *Multi-Document Extension*

In this paper we propose the extension of the method to process a cluster of documents written about the same topic. Multi-document summarization is one step more complex task than single-document summarization. It brings into new problems we have to deal with.

The first step is again to create a term by sentence matrix. In this case we include in the matrix all sentences from the cluster of documents. (On the contrary, in the case of single-document summarization we included the sentences from that document.) Then we run sentence ranking. Each sentence gets the score, which is computed in the same way as when we summarize a single document - vector length in the matrix $\Sigma^2 \cdot V^T$. Now, we are ready to select the best sentences (the ones with the greatest score) for the summary.

However, two documents written about the same topic/event can contain similar sentences and thus we need to solve redundancy. We propose the following process: before adding a sentence into the summary, look if there is a similar sentence already in the summary. The similarity is measured by the cosine similarity in the original term space. We determine a threshold here[1].

Extracted sentence should be close to the user query. To satisfy this, query terms get a higher weight in the input matrix[2].

Another problem of this approach is that it favours long sentences. Is the natural because a longer sentence problably contains more significant terms than a shorter one. We solve this by dividing the sentence score by $number - of - terms^{lk}$, where lk is the length coeficient[3].

Experiments showed good results with a low dimensionality. It is enough to use up to 10 dimensions (topics). However, the topics are not equally important. The magnitude of each singular value holds the topic importance. To make it more general we experimented with different power functions in the computation of the final matrix used for determination of sentence score: $\Sigma^{power} \cdot V^T$

IV. Experiments

1. *Data*

Our goal is to create a multilingual summarizer. We performed the evaluation on an English collection DUC and our czech collection CMDSC.

DUC 2005 included a multi-document summarization task, in which 32 systems participated. The DUC 2005 corpus contains more than 1300 documents organized in 50 clusters. For each cluster there is one task/query; and four 250-word abstracts written by assessors.

Unfortunately, there is no czech corpus annotated for summarization. Thus, we started to develop one, similarly to DUC. We call the corpus Czech Multi-Document Summarization Corpus (CMDSC). By now, it contains 80 documents organized in 7 clusters. For each cluster there are three different queries. The first is a gereral query. A user that has almost no information about the cluster topic would ask this way. The second is a specific query. In this case a user knows the basics of the topic but he would like to know something specific from the topic. And the third query is incremental. In this case a reader has read some of the documents, the ones with a date of publication less then a specified

[1]The experiments showed that the appropriate threshold value of the cosine of the angle between the two document vectors is 0.6.

[2]The best weight shown by experiments is 2. Thus, the query terms get twice higher weight than non-query terms.

[3]We obtained the best results with lk = 0.4.

date, and he would like to know what has happened in the topic since that date. Three annotators provided 250-word summaries for each cluster-query pair.

2. Evaluation Method

We used the same evaluation process as in DUC 2005. The only fully automatic method was ROUGE [8], N-gram comparison of human written abstracts with system summaries. Two scores were used: ROUGE-2 - bigram match, ROUGE-SU4 - skip-bigram measure. (For details see [8].) DUC 2005 data contains summaries of 32 systems which participated in the evaluation. We could compare our results with them.

3. Results

The results are summarized in table 1.

	ROUGE-2	**ROUGE-SU4**
Better then	27	27
Worse than	5	5
Sig. better than	9	11
Sig. worse than	0	0

Table 1: Comparison with DUC 2005 participating systems.

We can observe that our system is better than 27 systems in both ROUGE-2 and ROUGE-SU4, statistically significantly better (95% confidence) than 9, resp. 11 systems. It is worse than only 5 systems, but none of the differences is statistically significant.

Unfortunaly, in the case of czech language we do not have such systems for comparison.

To demostrate the result of our summarizer, we show here one of the summaries (2). The task was *What is the World Court? What types of cases does the World Court hear?*.

(2) THE WORLD COURT

THE International Court of Justice in the Hague will consider today Bosnian accusations that Serbs have been carrying out a campaign of territorial expansion through 'ethnic cleansing' and genocide. The International Court of Justice in the Hague yesterday ordered Serbia and Bosnia to stop acts of genocide in Bosnia, reaffirming an earlier ruling. The court would comprise 11 judges and sit at The Hague in the Netherlands, where the International Court of Justice is located. Earlier yesterday the International Court in The Hague rejected Libya's plea to bar the US and Britain from taking punitive measures. The court said Security Council Resolution 748 imposing sanctions should override all other international agreements. Col Gadaffi has separately opened an action against the US and Britain at the World Court in The Hague, accusing them of breaking international law by failing to hand over to Libya the evidence allegedly pointing to Libyan complicity in the bombings. The Court is due to hold a preliminary hearing on the action on March 26. The Panamanian government charged a violation of international law, and a famous Harvard Law School professor volunteered to argue the case before the International Court of Justice in The Hague. After six months of uproar, the U.S. district court judge in Miami ordered that the case proceed to trial. Nevertheless, Bush might spurn the World Court. Ever since the Reagan Administration walked out of the Hague to protest Nicaragua's claim of illegality in U.S. aid to the Contras, the State Department has opposed submitting to the World Court any case that involves the use of military force.

V. Conclusion

We presented our first multilingual multi-document summarizer. The proposed LSA-based method satisfies the multilinguality constraint because it works only with the context of terms. We experimented with Czech and English Corpora. The experiments showed that the summarizer is comparable with the best DUC participated systems. However, there is a lot of work to be done. We plan to further reduce sentences in the summary. We do experiments with removing unimportant clauses from them. Further, we need to improve the summaries in other ways. We need to correct the anaphoric expressions that cannot be resolved in the context of the summary. On the orher hand, there are expressions (like *the International Court of Justice in the Hague*) that appear in the summary more than once with their full noun phrase. The first occurence should be left in the full form, but the others should be shortened. Another problem that has to be solved is to sort the sentences that come into the summary from different documents.

Acknowledgement

This work was supported by grant no. 2C06009 Cot-Sewing.

References

[1] J. Steinberger and K. Ježek, "Text summarization and singular value decomposition," in *Lecture Notes in Computer Science 2457, Springer-Verlag*, pp. 245–254, 2004.

[2] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge," *Psychological Review*, 104:211–240, 1997.

[3] J. Steinberger, M. A. Kabadjov, M. Poesio, and O. Sanchez-Graillet, "Improving lsa-based summarization with anaphora resolution," in *Proceedings of Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, pp. 1–8, 2005.

[4] Wikimedia Foundation, Inc., *Multi-document summarization*, URL: http://en.wikipedia.org/wiki/Multi-document_summarization.

[5] E. Hovy and C. Lin, "Automated text summarization in summarist," in *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 1997.

[6] S. T. D. M. W. Berry and G. W. O'Brien, "Using linear algebra for intelligent ir," *SIAM Review*, 37(4), 1995.

[7] C. H. Q. Ding, "A probabilistic model for latent semantic indexing," *Journal of the American Society for Information Science and Technology*, 56(6):597–608, 2005.

[8] C. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of HLT-NAACL*, 2003.