The UWB Summariser at Multiling-2013

Josef Steinberger
University of West Bohemia
Faculty of Applied Sciences
Department of Computer Science and Engineering, NTIS Centre
Univerzitni 8, 30614 Plzeň, Czech Republic
jstein@kiv.zcu.cz

Abstract

The paper describes our participation in the Multi-document summarization task of Multiling-2013. The community initiative was born as a pilot task for the Text Analysis Conference in 2011. This year the corpus was extended by new three languages and another five topics, covering in total 15 topics in 10 languages. Our summariser is based on latent semantic analysis and it is in principle language independent. Its results on the Multiling-2011 corpus were promising. The generated summaries were ranked first in several languages based on various metrics. The summariser with minor changes was run on the updated 2013 corpus. Although we do not have the manual evaluation results yet the ROUGE-2 score indicates good results again. The summariser produced best summaries in 6 from 10 considered languages according to the ROUGE-2 metric.

1 Introduction

Multi-document summarization has received increasing attention during the last decade. This was mainly due to the requirement of news monitoring to reduce the big bulk of highly redundant news data. More and more interest arises for approaches that will be able to be applied on a variety of languages. The summariser should be of high quality. However, when applied in a highly multilingual environment, it has to be enough language-independent to guarantee similar performance across languages.

Given the lack of multilingual summarisation evaluation resources, the summarisation community started to discuss the topic at Text Analysis Conference (TAC[1]) 2010. It resulted in the first multilingual shared task organised as part of TAC 2011 – Multiling-2011 (Giannakopoulos et al., 2012). Each group took an active role in the creation of their language subcorpus. Because no freely available parallel corpus suitable for multi-document summarisation was found, news clusters from WikiNews (in English) needed to be first translated to six other languages. Three model summaries for each cluster were then written and both model and peer summaries were manually evaluated. For Multiling-2013, three new languages were added (Chinese, Romanian and Spanish) and 5 new topics (news clusters) were added to the corpus.

This article contains the description of our system based on latent semantic analysis (LSA) which participated in Multiling-2013. We first briefly discuss the multi-document task in section 2. Then we show our summarisation approach based on LSA (Section 3). The next section (4) compares the participating systems based on the ROUGE-2 score. Manually assigned scores were not available at the time of creation of this report. We conclude by a discussion of possible improvements of the method which require language-specific resources.

2 Multi-document summarisation task at Multiling'13

MultiLing-2013 is a community effort, a set of research tasks and a corresponding workshop which covers three summarisation tasks, focused on the multilingual aspect. It aims to evaluate the application of (partially or fully) language-independent summarization algorithms on a variety of languages.

The annotation part consisted of four phases. The first phase was to select English WikiNews articles about the same event and to create the topics. The articles were then manually translated to the other languages. Model summaries were created

[1] http://www.nist.gov/tac/

Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 50–54,
Sofia, Bulgaria, August 9 2013.

separately for each language by native speakers. In a certain time frame, participating groups ran their summarisers and the automatic summaries were then evaluated, both manually (on a 5-to-1 scale) and automatically by ROUGE (Lin, 2004) and the AutoSummENG metric (Giannakopoulos and Karkaletsis, 2010).

We participated with our summariser in the main multi-document task, which requires to generate a single, fluent, representative summary from a set of 10 documents describing an event sequence. The language of the document set (topic) was within a given range of 10 languages (Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian and Spanish) and all documents in a set share the same language. The output summary should be of the same language as its source documents. The output summary should be 250 words at most. The corpus was extended to 15 topics (Chinese, French and Hindi subcorpora contained only 10 topics).

3 LSA-based summarisation approach

Originally proposed by Gong and Liu (2002) and later improved by Steinberger and Ježek (2004), this approach first builds a term-by-sentence matrix from the source, then applies Singular Value Decomposition (SVD) and finally uses the resulting matrices to identify and extract the most salient sentences. SVD finds the latent (orthogonal) dimensions, which in simple terms correspond to the different topics discussed in the source.

More formally, we first build matrix $\mathbf{A}$ where each column represents the weighted term-frequency vector of a sentence in a given set of documents. The weighting scheme we found to work best is using a binary local weight and an entropy-based global weight (for details see Steinberger and Ježek (2009)).

After that step Singular Value Decomposition (SVD) is applied to the above matrix as $\mathbf{A} = \mathbf{USV}^T$, and subsequently matrix $\mathbf{F} = \mathbf{S} \cdot \mathbf{V}^T$ reduced to r dimensions[2] is derived.

Sentence selection starts with measuring the length of sentence vectors in matrix $\mathbf{F}$ computed as the Euclidean norm. The length of the vector (the sentence score) can be viewed as a measure for importance of that sentence within the top cluster topics.

The sentence with the largest score is selected as the first to go to the summary (its corresponding vector in $\mathbf{F}$ is denoted as $\mathbf{f}_{best}$). After placing it in the summary, the topic/sentence distribution in matrix $\mathbf{F}$ is changed by subtracting the information contained in that sentence:

$$\mathbf{F}^{(it+1)} = \mathbf{F}^{(it)} - \frac{\mathbf{f}_{best} \cdot \mathbf{f}_{best}^T}{|\mathbf{f}_{best}|^2} \cdot \mathbf{F}^{(it)}. \quad (1)$$

The vector lengths of similar sentences are decreased, thus preventing within summary redundancy. After the subtraction of information in the selected sentence, the process continues with the sentence which has the largest score computed on the updated matrix $\mathbf{F}$. The process is iteratively repeated until the required summary length is reached.

4 Experiments and results

Although the approach works only with term co-occurrence, and thus it is completely language-independent, pre-processing plays an important role and greatly affects the performance. When generating the summaries for Multiling-2013 each article was split into sentences. We used the old DUC sentence splitter[3], although a different sentence-splitting character was used for Chinese. It was a simplification because the sentence splitter should be adapted for each language (e.g. a different list of abbreviations should be used or language specific features should be added). If LSA is applied on a large matrix stopwords can be found in the first linear combination which could be then filtered out. However, in our case we apply it on rather small matrices and stopwords could affect negatively the topic distribution. Thus the safer option is to filter them out. This brings a dependency on a language but, on the other hand, acquiring lists of stop-words for various languages is not difficult. Filtering these insignificant terms does not also slow down the system. The stopwords were filtered out for all the languages of Multiling. The approach discussed in section 3 was then used to select sentences until the required summary length (250 words) has not been reached. Sentence order is important for event-based stories. In the case of the Multiling corpus,

[2]The degree of importance of each 'latent' topic is given by the singular values and the optimal number of latent topics (i.e., dimensions) r can be fine-tuned on training data. Our previous experiments led us to set r to 8% from the number of sentences for 250-word summaries.

[3]http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz

Language	Topics	Avg. Model	ID1	ID11	ID2	ID21	ID3	**ID4 (rank/total)**	ID5	ID51	Baseline
Arabic	15	.137	.132	.132	.118	.105	.052	**.167 (1/9)**	.105	.088	.086
Chinese	10	.462	.430	.457	.212	.354		**.354 (5/6)**			.867
Czech	15	.195	.155	.166	.123	.151		*.179 (1/6)*			.085
English	15	.185	.161	.161	.147	.142	.083	**.171 (1/9)**	.117	.101	.118
French	10	.198	.201	.201	.166	.177		**.214 (1/6)**			.130
Greek	15	.111	.120	.124	.100	.112		**.110 (4/6)**			.088
Hebrew	15	.076	.088	.100	.076	.084		**.092 (2/8)**	.087	.084	.072
Hindi	10	.342	.125	.132	.123	.123		**.129 (2/6)**			.114
Romanian	15	.543	.147	.139	.120	.138		**.166 (1/6)**			.098
Spanish	15	.239	.198	.218	.180	.175		**.228 (1/6)**			.164
Avg. rank			2.7	1.9	5.0	4.3	9	**1.9**	5.7	7.0	5.9

Table 1: ROUGE-2 scores of the average model and paricipating systems. Our LSA-based system is ID4 and we report its rank from the total number of systems which submitted summaries for the particular language. We included the baseline (the start of a centroid article) and excluded the topline which uses model sentences.

much attention has to be given to sentence ordering because some topics contained articles spread over a long period, even 5 years. We did not perform any temporal analysis at sentence level. The sentences in the summary were ordered based on the date of the article they came from. Sentences from the same article followed their order in the full text. Even if they were sometimes out of context, when extracted, the adjacent sentences at least dealt with the same (or temporary close) event.

We analysed ROUGE scores which we received from the organisers. We discuss here ROUGE-2 (bigram) score, a traditionally used metric in summarisation evaluation (Table 1). ROUGE-2 ranked our summariser on the top of the list for 6 from 10 languages (Arabic, Czech, English, French, Romanian, Spanish). System ID11 performed better twice (Hebrew and Hindi), there were three better systems in Greek and the baseline won in Chinese. In the following, we will discuss the results for each language separately.

For **Arabic**, our system received the best ROUGE-2 score. It was significantly better (at confidence 95%) then 5 other systems, including baseline. It performed on the same level as models.

It was our first attempt to run the summariser on **Chinese**. We did not use any specific word-splitting tool and we considered each character to be a context feature for LSA. The ROUGE results say that the summariser was not that successful compared to the others. It was significantly better than one system and worse than two and the baseline which received suspiciously high score.

We annotated the **Czech** part of the corpus, and therefore the result of our system can be considered only as another baseline for this language. It received the largest ROUGE-2 score, however, there was no significant difference among the top four systems.

For **English**, our system together with the following systems ID1 and ID11 were significantly better than the rest. A similar conclusion can be driven by observing the **French** results. In the case of **Greek** only baseline performed poorly. Our approach was ranked fourth although there were marginal differences between the systems. For **Hebrew** and **Hindi** system ID11 performed the best, followed by our system. For **Romanian**, a newly introduced language this year, our system received a high score, however, a larger confidence interval did not show much significance. For another newly-introduced language, **Spanish**, only system ID11 was not significantly worse than our system.

As a try to compare the systems across languages, an average rank was computed. (Computing an average of absolute ROUGE-2 scores did not seem to have sense.) Our system and system ID11 received the best average rank: 1.9.

For several languages (Arabic, French, Hebrew), our summaries were better (not significantly) then the average model according to ROUGE-2.

The AutoSummENG method (Giannakopoulos and Karkaletsis, 2010) gave results similar to those of ROUGE. The only difference was in Chi-

nese: ROUGE-2 ranked our system 5th, AutoSummENG 1st.

One question remains: are the ROUGE scores correlated with human grades? Unfortunately, the human grades were not available at the time of the system reports submission. However, because we were managing annotation of the Czech subcorpus we had access to human grades for that language. The system ranking provided by ROUGE mostly agree with the human grades, reaching Pearson correlation of .97 for the systems-only scenario. The human grades ranked our system as significantly better than any other submission in the case of Czech.

5 Conclusion

The evaluation indicates good results of our summariser, mainly for European Latin-script languages (Czech, English, French, Romanian and Spanish). It could be connected to good-enough pre-processing (sentence and word splitting). The last two languages were added this year and the good results show that the LSA-based summariser can produce good summaries when run on an 'unseen' language.

We experiment with several improvements of the method which require language-specific resources. Entity detection can improve the LSA model by adding entity features as new rows in the SVD input matrix (Steinberger et al., 2007). From the Multiling-2013 languages we have developed the NER tool only for 6 languages (Arabic, Czech, English, French, Romanian and Spanish) so far (Pouliquen and Steinberger, 2009). A coreference- (anaphora-) resolution can help in checking and rewriting the entity references in a summary (Steinberger et al., 2007) although there is usually a high dependency on the language (e.g. in the case of pronouns).

Event extraction can detect important aspects related to the category of the topic (e.g. detecting victims in a topic about an accident) (Steinberger et al., 2011). The aspect information can be used in the model weighting or during sentence selection. We have developed the tool for 5 languages considered in Multiling-2013 (Arabic, Czech, English, French and Spanish). Temporal analysis could improve sentence ordering if a correct temporal mark, which contains information about time of a discussed event, is attached to each summary sentence (Steinberger et al., 2012). So far, we experimented with English, French and Spanish from the list of the Multiling languages. By compressing and/or rephrasing the saved space in the summary could be filled in by the next most salient sentences, and thus the summary can cover more content from the source texts. We have already tried to investigate language-independent possibilities in that direction (Turchi et al., 2010).

Acknowledgments

This work was supported by project "NTIS - New Technologies for Information Society", European Center of Excellence, CZ.1.05/1.1.00/02.0090.

References

G. Giannakopoulos and V. Karkaletsis. 2010. Summarization system evaluation variations based on n-gram graphs. In *Proceedings of the Text Analysis Conference (TAC)*.

G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma. 2012. Tac 2011 multiling pilot overview. In *Proceedings of the Text Analysis Conference (TAC)*. NIST.

Y. Gong and X. Liu. 2002. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of ACM SIGIR*, New Orleans, US.

C.-Y. Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain.

B. Pouliquen and R. Steinberger. 2009. Automatic construction of multilingual name dictionaries. In Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster, editors, *Learning Machine Translation*. MIT Press, NIPS series.

J. Steinberger and K. Ježek. 2004. Text summarization and singular value decomposition. In *Proceedings of the 3rd ADVIS conference*, Izmir, Turkey.

J. Steinberger and K. Ježek. 2009. Update summarization based on novel topic distribution. In *Proceedings of the 9th ACM Symposium on Document Engineering, Munich, Germany*.

J. Steinberger, M. Poesio, M. Kabadjov, and K. Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680. Special Issue on Text Summarisation (Donna Harman, ed.).

J. Steinberger, H. Tanev, M. Kabadjov, and R. Steinberger. 2011. Aspect-driven news summarization. *International Journal of Computational Linguistics and Applications*, 2(1-2).

J. Steinberger, M. Kabadjov, R. Steinberger, H. Tanev, M. Turchi, and V. Zavarella. 2012. Towards language-independent news summarization. In *Proceedings of the Text Analysis Conference (TAC)*. NIST.

M. Turchi, J. Steinberger, M. Kabadjov, R. Steinberger, and N. Cristianini. 2010. Wrapping up a summary: from representation to generation. In *Proceedings of CLEF*.