Annotated suffix tree similarity measure for text summarization

Maxim Yakovlev and Ekaterina Chernyak

National Research University – Higher School of Economics
Moscow, Russia
myakovlev,echernyak@hse.ru

Abstract. The paper describes an attempt to improve the TextRank algorithm. TextRank is an algorithm for unsupervised text summarisation. It has two main stages: first stage is representing a text as a weighted directed graph, where nodes stand for single sentences, and edges are weighted with sentence similarity and connect sequential sentences. The second stage is applying the PageRank algorithm [1] as is to the graph. The nodes that get the highest ranks form the summary of the text.
We focus on the first stage, especially on measuring the sentence similarity. Mihalcea and Tarau [4] suggest to employ the common scheme: use the Vector space model (VSM), so that every text is a vector in the space of words or stems, and compute cosine similarity between these vectors. Our idea is to replace this scheme by using the annotated suffix trees (AST) [5] model for sentence representation. The AST overcomes several limitations of the VSM model, such as being dependent on the size of vocabulary, the length of sentences and demanding stemming or lemmatisation. This is achieved by taking all fuzzy matches between sentences into account and computing probabilities of matched coocurrencies.
More specifically we develop an algorithm for common subtree construction and annotation. The common subtrees are used to score the similarity between two sentences. Using this algorithm allows us to achieve slight improvements according to cosine baseline on our own collection of Russian newspaper texts. The AST measure gained around 0.05 points of precision more than the cosine measure. This is a great figure for natural language processing task, taking into account how low the baseline precision of the cosine measure is. The fact that the precision is so low can be explained by some lack of consistency in the constructed collection: the authors of the articles use different strategies to highlight the important sentences. The text collection is heterogeneous: in some articles there are 10 or more sentences highlighted, in some only the first one. Unfortunately, there is no other test collection for text summarisation in Russian. For further experiments we might need to exclude some articles, so that the size of summary would be more stable. Another issue of our test collection is the selection of sentences that form summaries. When the test collections are constructed manually, summaries are chosen to common principles. But we can not be sure that the sentences are not highlighted randomly.
Although the AST technique is rather slow, it is not a big issue for the text summarisation problem. The summarisation problem is not that

In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Porto, Portugal, 2015.

kind of problems where on-line algorithms are required. Hence the precision plays more significant part than time characteristics.
There are several directions of future work. First of all, we have to conduct experiments on the standard DUC (Document Understanding Conference [2]) collections in English. Second, we are going to develop different methods for construction and scoring of common subtrees and compare it to each other. Finally, we may use some external and more efficient implementation of the AST method, such as EAST Python library by Mikhail Dubov [3], which uses annotated suffix arrays. More details on this work can be found in [6].
Keywords: TextRank, annotated suffix tree

References

1. Brin S., Page L. : The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the seventh international conference on World Wide Web 7, 107-117 (1998)
2. Document Understanding Conference, `http://www-nlpir.nist.gov/`
3. Enhanced Annotated Suffix Tree, `https://pypi.python.org/pypi/EAST/0.2.2/`
4. Mihalcea R., Tarau P. : TextRank: bringing order into text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 404-411 (2004)
5. Pampapathi R., Mirkin B., Levene M. (2008): A suffix tree approach to anti-spam email filtering. In: Machine Learning, 65(1), 309-338 (2008)
6. Yakovlev M., Chernyak E. (2015): Using annotated suffix tree similarity measure for text summarization. In: Proccedings of European Conferenece on Data Analysis, 2015.