AUTOMATIC KEYPHRASE EXTRACTION BASED ON NLP AND STATISTICAL METHODS

Martin DOSTAL[1], and Karel JEŽEK[2]

1 INTRODUCTION

We would like to present our experimental approach to automatic keyphrase extraction based on statistical methods and Wordnet-based pattern evaluation. Automatic keyphrases are important for automatic tagging and clustering because manually assigned keyphrases are not sufficient in most cases. Keyphrase candidates are extracted in a new way derived from a combination of graph methods (TextRank) and statistical methods (TF*IDF). Keyword candidates are merged with named entities and stop words according to NL POS (Part Of a Speech) patterns. Automatic keyphrases are generated as TF*IDF weighted unigrams. Keyphrases describe the main ideas of documents in a human-readable way. Evaluation of this approach is presented in articles extracted from News web sites. Each article contains manually assigned topics/categories which are used for keyword evaluation.

2 OUR APPROACH

As first we would like to mention related method to automatic keywords extraction. TextRank (Mihalcea et al., 2004) can be used in individual documents without any other knowledge. Graph nodes ranking is made upon co-occurrences of tokens and the score of the other nodes with edges to the current one. In our approach, the TF*IDF score is used for better text token evaluation instead of the measure based only on n-grams (as Text-Rank). Our approach can be divided into two main steps:

A. Keyword extraction
B. Keyphrase extraction

Ad A) Rule for token's TF*IDF score is applied with restrictive boundary of acceptation between 80% to 20% of maximal score except named entities. We want to remove too general and too specific tokens. This boundary can be changed by the number of requested automatic keywords.

Ad B) The keyphrase extraction part can be described by these steps:

1) NLP method – interesting n-grams are chosen. The choice is based on their POS tag patterns and the corpus frequency is counted only for these n-grams. These n-grams can be marked as keyphrase candidates.

2) A score of importance is counted for each keyphrase candidate. This score contains the n-gram corpus frequency and TF*IDF score for each word. The score is used for document keyphrase selection.

[1] Martin Dostal, student of the doctoral study programme Computer Engineering, specialization Software Engineering, University of West Bohemia, e-mail: madostal@kiv.zcu.cz.

[2] Karel Ježek, professor at Department of Computer Science and Engineering, University of West Bohemia, email: jezek_ka@kiv.zcu.cz.

3) Derivation – keyphrase candidates are merged with named entities or individual keywords if their co-occurrence is significant for this document.

3 EXPERIMENTAL RESULTS

We have used two data sets:

1) Corpus of 50 at random selected articles with a small number (approximately 3) of manually assigned topics. Precision and recall for this corpus are shown in table 1.

Boundary	1%	10%	30%	50%	70%	90%
Precision	30%	38.6%	48%	49.4%	50.7%	55.2%
Recall	49%	33%	23.7%	18.5%	13.6%	12.9%

Tab. 1: Precision and recall for the corpus of 50 articles.

2) Corpus of 50 at random selected articles with manually assigned topics (by author) and expanded, by 2-3 another human annotators, to other important topics covered by the article. Each article contains approximately 5 manual topics. Precision and recall for this corpus are shown in table 2.

Boundary	1%	10%	30%	50%	70%	90%
Precision	37.4%	47.4%	55.8%	59.4%	60.6%	64%
Recall	54.6%	35.9%	23.7%	18.5%	14.1%	13.5%

Tab. 2: The corpus of 50 articles with additional human annotations.

4 CONCLUSION

The proposed approach seems to be efficient enough to be comparable with other automatic keyword extraction systems. For example, the RAKE system (Rose et al., 2010) achieved 33.7% precision with 41.5% recall and the undirected TextRank (Mihalcea et al., 2004) achieved 31.2% precision with 43.1% recall. Our approach achieved 37.4% precision and 54.6% recall for a small corpus with expanded number of annotations, including the problem of keyword generation and automatic clustering. We can assume that precision and recall will be a little bit lower for a bigger corpus. The most significant feature of the corpus is the number of exact manual annotations which are used for performance tests.

In the future, we would like to compare our approach with other methods on their data corpuses. These corpuses were not available at this moment so we had to use our data collection for the first evaluation tests. Automatic keywords will be used for mapping the Linked Data topics to the articles and graph-based knowledge extraction.

REFERENCES

Mihalcea, R., nad Tarau, P., 2004. Textrank: Bringing order into texts. *Proceedings of EMNLP 2004* (ed. Lin D and Wu D), pp. 404–411, Barcelona, Spain.

Rose, S., Engel, D., Cramer, N., and Cowley, D. 2010. Automatic keyword extraction from individual documents in *Text mining applications and theory*. pp 3-19.