JMZW: TOPIC IDENTIFICATION IN CZECH NEWSPAPER ARTICLES

Lucie SKORKOVSKÁ [1]

1 INTRODUCTION

Topic identification module is a part of the complex system for acquisition and storing large volumes of text data from the Web called *JMZW - Jazykové modelování z webu*. This module processes each acquired text item, mostly newspaper article, and automatically assigns keywords from a predefined topic hierarchy to it.The main purpose of the JMZW system is to acquire and process data for training of extensive language models used in Automatic Speech Recognition systems. Since it has been shown Psutka et al. (2003) that a smaller topic specific language model can outperform a much bigger general one, it is important to filter the gathered data according to its topics.

2 REALIZATION

Each newly downloaded news article is preprocessed by JMZW system's algorithms before automatic topic identification starts. One of the problems that we have to solve is how many topics we should assign to each article. For the current version of the algorithm we have experimentally chosen to assign 3 topics, the topics are chosen from a topic tree.

2.1 Topic tree

At present the topic tree has 32 main topic categories like `health`, `culture` or `sport`, each of this main category has its subcategories with the "smallest" topics represented as leaves of this tree. In the current system, we use the topic tree with about 450 topics and topic categories, which correspond to the keywords assigned to the articles on the news servers *ČeskéNoviny.cz* or *iDnes.cz*. The articles with these "originally" assigned topics are used as training text for identification algorithms.

2.2 Identification algorithms

Two methods for automatic topic identification was implemented so far, a TF-IDF vector space model based classification and a language model based classification (formally similar to the Naive Bayes classifier (Manning et al. (2008))).

The goal of the language modeling based approach is to find the most likely or the maximum a posteriori topic (or topics) T_{map} of an article A:

$$T_{map} = \arg\max_T \hat{P}(T|A) = \arg\max_T \prod_{t \in A} \hat{P}(t|T) \quad (1)$$

where $\hat{P(T}|A)$ is a probability of an article A belonging to a topic T and $\hat{P}(t|T)$ is a conditional probability of a term t given the topic T.

[1]Ing. Lucie Skorkovská, student of the doctoral study programme Applied Sciences and Informatics, specialization Cybernetics, e-mail: lskorkov@kky.zcu.cz

In the TF-IDF vector space model classification is the similarity of an article A and a topic T computed as:

$$sim(A,T) = \sum_{t \in A} tf_{t,T} \cdot idf_t \tag{2}$$

where $tf_{t,T} \cdot idf_t$ is the term frequency and inverse document frequency of a term t. The topics with the highest similarity are then assigned to the tested article.

2.3 Evaluation

Two types of evaluation were performed, one from the point of view of information retrieval (IR), where each article is considered as a query and precision (P), recall (R) and F_1-measure is computed for the answer topic set. The second type of evaluation is from the point of view of a topic classifier, where P, R and F_1 is computed for each topic separately. Two ways of computing the average measures are applied, *microaveraging* and *macroaveraging*. The results for the test set of 15 000 articles are shown in table 1.

Tab. 1: Average P, R and F_1 of topic identification results for 15 000 set of articles

	IR point of view			**microaveraging**			**macroaveraging**		
classification method	P	R	F_1	P	R	F_1	P	R	F_1
language modeling	0.594	0.626	0.583	0.597	0.570	0.583	0.624	0.442	0.517
vector space model	0.495	0.523	0.486	0.496	0.475	0.485	0.496	0.273	0.352

3 CONCLUSION

The language modeling approach seems to achieve better results than vector space modeling, especially for topics with the small article set, which can be seen from the *macroaverage* R and F_1 measures. It may seem that the results are not so good, but it must be taken into consideration that we have a very large set of topics that are in many cases not well distinguished. Also the articles in the test collection are taken as they were on the news server, the original reference topics was not revised in any way, so in many cases the topic we assign is also "correct", but it is not included in the reference set of topics.

For the future work, we would like to include in the topic identification module an automatic determination of the number of topics that should be assigned to each article. The number of the assigned topics should not be predefined, but it should be somehow related to the topic identification similarity score.

Acknowledgement: The work has been supported by the grant of The University of West Bohemia, project No. SGS-2010-054, and by the Ministry of Industry and Trade, project No. MPO FR-TI1/486.

REFERENCES

Manning, Christopher D., Raghavan, Prabhakar and Schütze, Hinrich, 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York.

Psutka, J., Ircing, P., Psutka, J.V., Radová, V., Byrne, W., Hajič, J., Mírovský, J. and Gustman, S., 2003. Large vocabulary ASR for spontaneous Czech in the MALACH project. *Proceedings of Eurospeech 2003*. pp. 1821–1824. Geneva, Switzerland.