Studentská Vědecká Konference 2010

FIRST EXPERIMENTS WITH AUTOMATIC TOPIC IDENTIFICATION OF CZECH NEWSPAPER ARTICLES

Lucie SKORKOVSKÁ [1]

1 INTRODUCTION

Topic identification module is an important part of the system for generation of language models with the use of internet data (in development at the Department of Cybernetics). A language model is an essential component of every automatic speech recognition system and it requires large amounts of data to be well trained. Our effort is to use the large potential of the internet for gathering the topic oriented training data.

We intend to use articles from internet newspapers to train the topic oriented language models. As a typical article from most newspapers has no assigned keywords and no precisely defined topic, we have to automatically identify the topic of each article.

Our goal is to automatically identify the topic of an article from among the set of topics. For our experiments we used data from news server České Noviny . The articles here have manually specified keywords. We have used 10412 articles in our test, 9000 as training data, the rest as test data. The training articles contained 3434 keywords - topics for our purpose.

2 REALIZATION

For the first experiments the Naive Bayes classifier was chosen (Manning et al. (2008)), where the probability $P(T|A)$ of an article A belonging to a class (topic in our case) T is computed as

$$P(T|A) \propto P(T) \prod_{t \in A} P(t|T) \tag{1}$$

where $P(T)$ is the prior probability of a topic T and $P(t|T)$ is a conditional probability of a term t given the topic T. This probability can be estimated by the maximum likelihood estimate (MLE) simply as the relative frequency of the term t in the training articles belonging to the topic T:

$$\hat{P}(t|T) = \frac{tf_{t,T}}{N_T} \tag{2}$$

where $tf_{t,T}$ is the frequency of the term t in T and N_T is the total number of tokens in articles of the topic T.

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

$$T_{map} = arg\ max_T \hat{P}(T|A) = arg\ max_T \hat{P}(T) \prod_{t \in A} \hat{P}(t|T). \tag{3}$$

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

The linear interpolation smoothing method (Jelinek and Mercer (1980)) was also implemented, where the computation of the probability $P(T|A)$ is:

$$P(T|A) \propto P(T) \prod_{t \in A} (\lambda P(t|T) + (1 - \lambda) P(t|M_C)) \quad (4)$$

where $P(t|M_C)$ is a conditional probability of a term t in the whole article collection and λ is an interpolation parameter.

3 CONCLUSION

These first experiments with topic identification have disclosed several findings usable for the future research. First, the experiments have shown that automatic use of all keywords as topics is not suitable - many of them are too detailed to have enough articles assigned to proper model the topic. So for the future work the development of some keywords clusters is needed.

Second, the results have shown that the use of the prior probability of a topic has not affected the identification process. Identified topics and also their order is the same as when $P(T)$ is considered equal for all topics, so it can be ignored in future experiments.

Finally, as smoothing is considered to be an essential part of the language modeling approach to the information retrieval, it was expected to improve topic identification as well. On the contrary, the first analysis of the results of our experiments has shown that smoothing may not be necessary in this task. Deeper examination of this hypothesis is a suitable matter for further research.

Acknowledgement: The work has been supported by the grant of The University of West Bohemia, project No. SGS-2010-054 - "Inteligentní metody strojového vnímání a porozumění"

REFERENCES

České Noviny. *www.ceskenoviny.cz*.

Manning, Christopher D., Raghavan, Prabhakar and Schütze, Hinrich, 2008. Chapter 13 - Text classification & Naive Bayes. *Introduction to Information Retrieval*. Cambridge University Press, New York.

Jelinek, Frederick and Mercer, Robert L., 1980. Interpolated estimation of Markov source parameters from sparse data. *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland.