JMZW: Application of Summarization Method in the Topic Identification of Czech Newspaper Articles

Lucie Skorkovská[1]

1 Introduction

The topic identification module, which is a part of a complex system for acquisition and storing large volumes of text data (Švec et al. (2011)), processes each acquired data item and assigns to it topics from a defined topic hierarchy. The topic hierarchy is quite extensive - it contains about 450 topics and topic categories. Since the system is used for processing large amounts of data, a summarization method was implemented and the effect of using only the summary of an article on the topic identification accuracy is studied.

The main purpose of the topic identification module is to filter the huge amount of data according to their topics for the future use as the language modeling training data. The module uses a language modeling based approach similar to the Naive Bayes classifier for the implementation of the topic identification and assigns 3 topics to each article. Topics are chosen from a hierarchical system - a "topic tree". Further information about the topic identification module can be found in Skorkovská et al. (2011).

2 Summarization module

For the automatic summarization module an extractive generic summarization was chosen, as we want our summaries to preserve all the information contained in the original text, so the topic identification module can assign the correct topics. The implemented summarization algorithm selects the most important sentences in a text, where an importance of a sentence is measured by the importance of its words. One of the most commonly used measure for assessing the word importance in information retrieval area is the TF-IDF measure (Term Frequency - Inverse Document Frequency), so we have decided to use it as well. The summary is created in a following way:

1. Split text to sentences and sentences to words.
2. For each term t in the document compute an idf weight:
$$idf_t = \log \frac{N}{N_t} \tag{1}$$
where N is the total number of sentences in the document and N_t is the number of sentences containing the term t.
3. For each sentence s compute a term frequency $tf_{t,s}$ for each term. We have used the normalization of the term frequency by the maximum term frequency in the sentence.
4. The importance score S of each sentence in the document is computed as:
$$S_s = \sum_{t \in s} tf_{t,s} \cdot idf_t \tag{2}$$

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

5. The five sentences with the highest score S are included in the summary.

3 Evaluation

For the experiments three smaller collections containing 5000, 10000 and 31419 articles were separated from the whole corpus. Evaluation from the point of view of information retrieval (IR) was performed on the collections, where each newly downloaded article is considered as a query in IR and precision (P), recall (R) and F_1-measure is computed for the answer topic set. The results can be seen in table 1. For the comparison the collections trained also using the preprocessing with the lemmatization module are shown.

Table 1: Results of topic identification on different collections

coll./art.		word	lemma	summary	lemma/summary	word/summary
5k	P	0.5366	**0.5547**	0.5028	*0.5457*	0.5374
	R	0.5544	**0.5754**	0.5155	*0.5686*	0.5546
	F_1	0.5454	**0.5649**	0.5091	*0.5569*	0.5459
10k	P	0.5481	**0.5536**	0.5024	*0.5378*	0.5293
	R	0.5472	**0.5555**	0.4979	*0.5421*	0.53
	F_1	0.5476	**0.5546**	0.5002	*0.54*	0.5296
30k	P	0.5864	**0.5859**	0.5387	*0.5588*	0.5598
	R	0.6125	**0.6155**	0.5616	*0.5921*	0.5884
	F_1	0.5992	**0.6003**	0.5499	*0.575*	0.5737

4 Conclusion

From the table we can draw following conclusions: first, the summarized text is not suitable for training topic identification statistics - results in column summary are the worst for all sizes of collections. This is not surprising, as much less text is used for counting the statistics so the topic important words may be missing.

On the other hand, interesting finding can be seen in column lemma/summary. When needed, a faster computation of topic identification using summarized and lemmatized texts can be used with a minimum loss on the topic identification accuracy. The time needed for the topic identification of an article is reduced as the computation of the probability $P(T|A)$ of an article belonging to a topic is done over a reduced set of words.

Acknowledgement

The work has been supported by the grant of The University of West Bohemia, project No. SGS-2010-054.

References

Skorkovská, L., Ircing, P., Pražák, A., Lehečka, J., 2011. Automatic topic identification for large scale language modeling data filtering. *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 6836, pp. 64–71. Springer Berlin / Heidelberg

Švec, J., Hoidekr, J., Soutner, D., Vavruška, J., 2011. Web text data mining for building large scale language modelling corpus. *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 6836, pp. 356–363. Springer Berlin / Heidelberg