Automatic Topic Identification for Large Scale Language Modeling Data Filtering

Lucie Skorkovská, Pavel Ircing, Aleš Pražák, and Jan Lehečka

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{lskorkov,ircing,aprazak,jlehecka}@kky.zcu.cz

Abstract. The paper presents a module for topic identification that is embedded into a complex system for acquisition and storing large volumes of text data from the Web. The module processes each of the acquired data items and assigns keywords to them from a defined topic hierarchy that was developed for this purposes and is also described in the paper. The quality of the topic identification is evaluated in two ways - using classic precision-recall measures and also indirectly, by measuring the ASR performance of the topic-specific language models that are built using the automatically filtered data.

Keywords: topic identification, language modeling, automatic speech recognition.

1 Introduction

Statistical language models (LM) that constitute the state-of-the-art language modeling technique in many areas of natural language processing (automatic speech recognition, machine translation, etc.) require an extensive amount of training data in order to ensure the robust estimation of their parameters. It might seem that the problem of data availability has already disappeared as the quantity of electronic texts available on-line nowadays exceeds every conceivable limit. However, when we want to use those data for language modeling, there are still several important problems that have to be solved. We must tackle the technical issues related to the actual download of the on-line content, the algorithms for stripping of the HTML (or other) markup, methods for text tokenization and normalization and, last but not least, also the detection of possible duplicate documents. The system that deals with the mentioned tasks is introduced in [8].

Once we have the "cleaned" data available, it is still not practical to use them for language modeling right away. First, the data are typically huge to the extent that it complicates the actual language model construction. Even more importantly, there is the evidence that the data quantity by itself might not be sufficient for good language model performance and what is more important is the right scope of the LM training texts. When the topic of the LM target domain is really specific, it happens that the "in-domain" language model estimated on a moderate-sized corpus vastly outperforms the model built using the data that are one or two orders of magnitude bigger but constitute just a general corpus [5].

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 64–71, 2011.

Thus, when we download and store texts that are meant for future LM training, the information about the document topic is extremely valuable but, at the same time, often not available from the data source. This paper therefore introduces a method for automatic identification of the document topic and presents two different evaluation scenarios for determining the method efficiency.

2 Topic Identification

As mentioned before, the main purpose of our topic identification module is to filter the huge amount of data according to their topics for the future use as the LM training data. We decided that more than one topic should be assigned to each article in our database and that the topics should form some sort of hierarchical system - a topic tree.

The topic identification module is a part of the system described in [8], each newly downloaded article is preprocessed by that 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 to each article.

2.1 Topic (Keyword) Tree

When we started to design our topic identification module, we searched for some kind of an existing topic hierarchy, but we found out that there is no such suitable hierarchical system. Consequently, we have build our own topic hierarchy in the form of topic tree, based on our expert findings in topic and keyword distribution in the articles on the favourite news servers like *ČeskéNoviny.cz* or *iDnes.cz*.

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. An example of a branch representing the topic category `justice & courts` from the topic tree can be seen on figure 1.

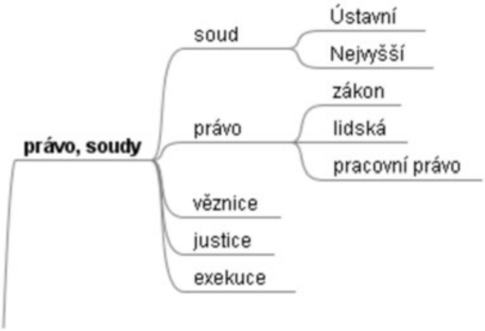

Fig. 1. Branch of the topic tree representing the topic `justice & courts`

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 mentioned news servers. 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 classification based on TF-IDF[1] vector space model and a language modeling based classification. These methods were selected due to the good results in our information retrieval experiments [2], since we had no experience with the topic identification task so far.

Language Modeling Based Classification. The language modeling based approach chosen for the first experiments is similar to the Naive Bayes classifier [3], 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 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 (or topics) T_{map} of an article A:

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

The prior probability of the topic $\hat{P}(T)$ was implemented as the relative frequency of the articles belonging to the topic in the training set, but we found out that it has no effect on the identification results.

Vector Space Model Classification. The second tested algorithm is the TF-IDF vector space model based classification. For each term t in the topic T the term frequency $tf_{t,T}$ and inverse document frequency is computed:

$$idf_t = \log \frac{N}{N_t} \tag{4}$$

[1] Term Frequency - Inverse Document Frequency.

where N is the total number of topics and N_t is the number of topics containing the term t. The similarity of an article A and a topic T is then computed as:

$$sim(A,T) = \sum_{t \in A} tf_{t,T} \cdot idf_t \ . \tag{5}$$

The topics with the highest similarity are then assigned to the tested article.

2.3 Evaluation

For the evaluation of the chosen topic identification methods a smaller collection of articles from the news server *ČeskéNoviny.cz* was separated. This collection contains 158 000 articles, 140 000 of these articles were used as topic training data, remaining 18 000 is available for evaluation testing. The articles from *ČeskéNoviny.cz* have included the originally assigned keywords from their authors (in average 3.5 keywords for one article), which were used as the training and reference topics.

Two types of evaluation were performed on the test collection. The first one is more from the point of view of information retrieval (IR), 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:

$$P = \frac{T_C}{T_A}, \qquad R = \frac{T_C}{T_R}, \qquad F_1 = 2\frac{P \cdot R}{P + R} \tag{6}$$

where T_A is the number of topics assigned to the article, T_C is the number of correctly assigned topics and T_R is the number of relevant reference topics. An average of these measures is then computed across a set of testing articles.

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 can be applied in this case, *microaveraging* (topics count proportionally to the size of the topic article set):

$$P_{micro} = \frac{\sum_T T_C}{\sum_T T_A}, \qquad R_{micro} = \frac{\sum_T T_C}{\sum_T T_R} \tag{7}$$

and *macroaveraging* (all topics count the same):

$$P_{macro} = \frac{\sum_T P_T}{|T|}, \qquad R_{macro} = \frac{\sum_T R_T}{|T|} \tag{8}$$

In this case T_A refers to the number of articles assigned to a topic, T_C is the number of articles correctly assigned to the topic i.e. the “true positives”, T_R is the true number of articles with the topic and $|T|$ is the total number of topics. The *macroaverage* measures are more important in our case, because we want our classifier to perform well on infrequent topics, too.

First, we wanted to find out the best number of topics to assign to each article. The relation between the number of topics and P, R and F_1 measures from the IR point of view is shown on figure 2, it can be seen that best results are obtained for 3 assigned

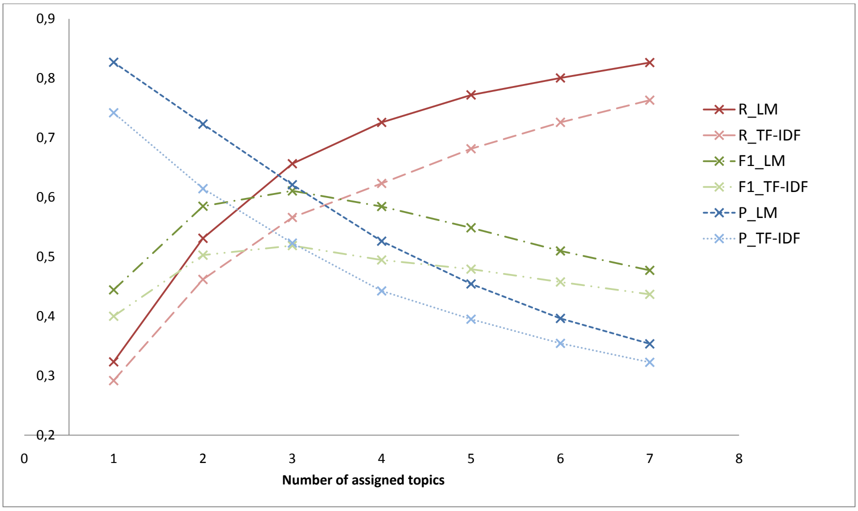

Fig. 2. Dependency of P, R and F1 on the number of assigned topics

Table 1. Average P, R and F_1 measures 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

topics. P, R and F_1 measures obtained for the test set of 15,000 articles and 3 assigned topics are shown in table 1.

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 to the article is also "correct", but it is not included in the reference set of topics. For example, the article about the achievements of the hockey representation has only `hockey` in reference topics, but our topic identification module assigned the topics `hockey, representation`, which is correct as well.

3 Language Modeling and ASR Experiments

The main motivation for the development of an automatic topic identification method introduced in the previous chapter was that we wanted to be able to effectively retrieve large amounts of domain-specific data for language model training. In this chapter, we will therefore present several experiments with language models estimated on the text corpora that were filtered from the large database of newspaper articles using various selection criteria. Since the ultimate measure of the language model quality is the performance of the system where the LM is employed (in this case the ASR decoder), we will also describe the speech recognition system that we have used and report relevant Word-Error-Rates (WER).

All language models perplexities (PPL) and WER were evaluated on test set consisting of speech obtained during the testing phase of the automatic closed-captioning system that employs the so-called “shadow-speaker” approach [4]. It means that the potentially noisy and/or overlapping broadcast speech is respoken with a trained speaker in controlled acoustic conditions in order to ensure higher recognition accuracy. The evaluation set contains recordings from just a single female speaker. The total length of the test set audio is 98 minutes. Since the speaker, whose utterances are in the test set, recorded in fact over 25 hours of data in total, we were able to tailor the acoustic models of the ASR system to this particular speaker (see [7] and [9] for details). This gives us a very high quality acoustic model and consequently, we can safely assume that any ASR performance gain from the improved language model would be even more prominent in the case when the acoustic model is less effective. All the language models described in the following paragraphs are trigram LMs estimated using the SRI Language Modeling Toolkit (SRILM) [6] employing the default Good-Turing discounting method. The resulting models always contain all the lexicon word bigrams that are found in the training data; the trigrams must occur at least twice to be included in the model.

The test set consists of samples of the dialogues that took place during the political talk show (“Otázky Václava Moravce”) broadcast by the Czech Television on July 18th, 2010. This particular show discussed mainly the newly appointed Czech government, state budget and also the health care issues. The appropriate keywords from the first-tier of the tree would then be `politics & diplomacy`, `economy` and `health`.[2] The first three lines of the Table 2 thus describe the language models that were trained using the articles published between January 1st, 2009 and July 17th, 2010 and are labeled with any keyword that comes from the subtree with the headword `politics & diplomacy`, `politics & diplomacy` and `economy`, and `politics & diplomacy`, `economy` and `health`. The results for these topic-specific LMs are compared with the models that are trained from all the articles that were published in the defined period just prior the broadcast day (lines 4 to 6). Such criterion is also a powerful filter of the retrieved data topics as the “hot” issues tend to be discussed across different mass media at the same time. The length of the periods was chosen in order to obtain roughly the same amount of selected data for both topic-defined and time-defined selections.

It can be seen from the results that topic-defined language models moderately outperform the time-defined ones in term of WER (8 to 12% relative improvement). It should be noted, however, that `politics & diplomacy` is the second most frequent of all first-tier topics and constitutes over 13% of the articles (there are 32 topic in the

[2] Note that assuming to know the topics before the actual broadcasting is not unrealistic - the main themes of each debate are published on the broadcaster website beforehand.

Table 2. Properties of language models trained on different data selections

	Selection ID	# tokens	Lex. size	LM size [MB]	OOV [%]	PPL	WER [%]
1	politics & dipl.	42M	281k	348	1.19	777	4.56
2	pol.+econ.	51M	302k	426	1.19	718	4.44
3	pol.+econ.+heal.	62M	358k	512	1.09	716	4.42
4	4-months	37M	307k	313	1.31	1,167	5.23
5	5-months	47M	332k	382	1.20	1,087	5.05
6	7-months	67M	378k	551	1.17	983	4.80
7	7-months P.E.H. only	28M	279k	268	1.37	850	4.97
8	sport	48M	190k	262	4.53	4,493	8.80

first level of the tree in total). Only the somehow fuzzy topic `relax` is slightly more frequent and thus the rather topicaly coherent `politics` articles probably dominate even the general language model.

In order to make the direct comparison between general and topic-specific corpus, we have further applied the keyword filter from line 3 to the selection from line 6. As can be seen from line 7, a 3.5% relative increase of WER was observed, however, the topic-filtered model size is only about 50% of the general one. Such finding is important for the potential future deployment of limited-resource ASR systems. Finally, to show that we really cannot just use any large-enough text data to train a good language model, we have performed a somehow extreme experiment by taking the articles labeled with the all the "sport" keywords. As you can see from line 8, the WER increased by almost a 100%.

4 Conclusions and Future Work

Both the evaluation of the topic identification accuracy itself and the indirect evaluation of the WER of the resulting topic-specific language models suggest that the topic identification algorithms presented in this paper work reasonably well. However, there is still some room for improvement. First, some topics are clearly ill-defined, especially the ones concerning geography - there are often either too broad (e.g. `USA`) or too narrow (e.g. `Vítkov` - small town connected with one xenophobic cause). Some other topics make a good sense intuitively, but are extremely hard for the automatic system to distinguish between as they use virtually identical phraseology (e.g. men's and women's tennis competitions `Davis Cup` and `Fed Cup`, respectively).

Second, we would like to implement some improvements for the presented topic identification methods like the k-NN classifier for the vector space model or the use of topic bigram LMs for the language modeling based classification and test the effects of these improvements on the topic identification results. More sophisticated methods like Support Vector Machines for text classification [1] could also be explored.

Finally, one of the most challenging improvements that we would like to include in the future version of the topic identification module is the automatic determination of the number of topics that will 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.

Acknowledgments. This research was supported by the Technology Agency of the Czech Republic, project No. TA01011264 and by the grant of the University of West Bohemia, project No. SGS-2010-054.

References

1. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
2. Kanis, J., Skorkovská, L.: Comparison of different lemmatization approaches through the means of information retrieval performance. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 93–100. Springer, Heidelberg (2010)
3. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
4. Pražák, A., Loose, Z., Psutka, J., Radová, V., Müller, L.: Four-phase re-speaker training system. In: Proceedings of SIGMAP 2011, Seville (2011)
5. Psutka, J., Ircing, P., Psutka, J.V., Radová, V., Byrne, W., Hajič, J., Mírovský, J., Gustman, S.: Large vocabulary ASR for spontaneous Czech in the MALACH project. In: Proceedings of Eurospeech 2003, Geneva, pp. 1821–1824 (2003)
6. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of ICSLP 2002, Denver, pp. 901–904 (2002)
7. Vaněk, J., Psutka, J.: Gender-dependent acoustic models fusion developed for automatic subtitling of parliament meetings broadcasted by the Czech TV. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 431–438. Springer, Heidelberg (2010)
8. Švec, J., Hoidekr, J., Soutner, D., Vavruška, J.: Web text data mining for building large scale language modelling corpus. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS(LNAI), vol. 6836, pp. 356–363. Springer, Heidelberg (2011)
9. Zajíc, Z., Machlica, L., Müller, L.: Robust statistic estimates for adaptation in the task of speech recognition. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 464–471. Springer, Heidelberg (2010)