Combined models for topic spotting and topic-dependent language modeling

Brigitte Bigi, Renato de Mori, Marc El-Bèze, Thierry Spriet

COMBINED MODELS FOR TOPIC SPOTTING AND TOPIC-DEPENDENT LANGUAGE MODELING

Brigitte Bigi, Renato De Mori, Marc El-Bèze, Thierry Spriet
{bigi,demori,elbeze,spriet}@univ-avignon.fr
LIA, University of Avignon
CERI-IUP BP 1228
84911 Avignon Cedex 9 - France

Abstract - A new statistical method for Language Modeling and spoken document classification is proposed. It is based on a mixture of topic dependent probabilities. Each topic dependent probability is in turn a mixture of n-gram probabilities and the probability of Kullback-Lieber (KL) distances between key-word unigrams and distribution obtained from the content of a cache memory. Experimental result on topic classification using a corpus of 60 Mword from the French newspaper *Le Monde* show the excellent performance of the cache memory and its complementary role in providing different statistics for the decision process.

1. Introduction

Topic classification has been an important subject in Information Retrieval (IR). A number of methods have been developed for the classification of text documents. Recently, statistical methods based on unigram word probabilities have been proposed [5].

The use of topic-dependent language models (LM) is also now the object of research efforts [2,6,8]. For the automatic transcription of spoken documents [2,6,8], mixtures of topic-dependent LMs can be considered [3,5].

In general, a mixture combines different LM probabilities to compute the probability of a word w_i given its history h_i in the following way :

$$P(w_i|h_i) = \sum_{j=1}^{J} \lambda_j P_j(w_i|h_i) \qquad (1)$$

where $P_j(w_i|h_i)$ is the probability provided by the *j*-th LM and λ_i is a constant satisfying the constraint :

$$\sum_{j=1}^{J} \lambda_j = 1 \qquad (2)$$

It is possible to make the λ_j topic-dependent and equal to scores that are useful for topic spotting too. In fact, $P(w_i|h_i)$ can also be expressed as :

$$P(w_i|h_i) = \sum_{j=1}^{J} P(w_i T_j|h_i) = \sum_{j=1}^{J} P(w_i|T_j h_i) P(T_j|h_i) \qquad (3)$$

where T_j is a topic. The term $P(w_i|T_j h_i)$ may be computed with an n-gram LM that is topic-dependent. This corresponds to make the approximation that h_i is represented only by the *n-1* words preceding w_i. The other term $P(T_j|h_i)$ can be computed using a different approximation for h_i. This probability can be used for *dynamically* modulating the importance of the elements of the mixture and for topic spotting as well.

It can, in turn, be expressed as a mixture of probabilities obtained with different models, and, more important, it can vary in time as new words are made available, thus making the topic-dependent LM ***adaptable***.

This paper proposes to use two models, the first one based on unigram or n-gram probabilities involving all the words of a vocabulary and the second one based on a comparison between the content of one or more *cache memories* [4] and the *static* unigram distributions of topic keywords with a constant floor value assigned to all the other words as well as the words not in the cache.

A linear combination of the two models may use probabilities with different history approximations :

$$P(T_j|h_i) = \alpha_1 P_1(T_j|W_1^{i-1}) + \alpha_2 P_2\{T_j|d[R_j, cc(i-1)]\} \qquad (4)$$

where W_1^{i-1} is the sequence $\{w_1,...w_{i-1}\}$ of the first *(i-1)* words of a document. If the first model is based on unigrams, then:

$$P_1(T_j|W_1^{i-1}) = \prod_{t=1}^{i-1} \frac{P_1(w_t|T_j)P(T_j)}{\sum_{k=1}^{J} P_1(w_t|T_k)P(T_k)} \qquad (5)$$

The problem when using this model alone is limited accuracy and the fact that non-keywords have the same importance as topic keywords.

Topic characterization accuracy can be improved by introducing the second term of the (4). Here, a set Γ_j of keywords is selected for each topic and its statistical distribution R_j obtained from a training corpus is compared with the distribution of content $cc(i-1)$ of a cache memory when word w_{i-1} is read in.

The result of such a comparison is $d[R_j, cc(i-1)]$, a symmetric Kullback-Lieber distance that varies in time as new words are considered. Probability P_2 may be conditioned only on the distance between the distribution of the cache content and R_j or can be conditioned by a vector of whose elements are distances between the distribution of the cache content and the distribution of the keywords of each topic. A single cache is used for the sake of simplicity (a different cache could be considered for each part-of-speech, or word class).

The first term of the (4) cumulates information of a long history using all the words in the vocabulary.

In the experiments described in this paper, involving a corpus of articles from the French newspaper "*Le Monde*", there are more than 500,000 different words.

Let *L* be the size of such a vocabulary. For each topic, it is possible to order the *L* words according to their frequency in a training corpus and select the first *C*<*L* words of this ordered sequence.

These first *C* words are called *topic keywords*. The comparison of the cache content is limited only to topic key-words. For this purpose, it is desirable to normalize the *keyword* probabilities of a topic in such a way that the *L-C* less probable words have all the same floor unigram probability.

It is also possible to make this floor probability independent from the topic. In this way, the contribution of the second term in the mixture (4) differs from the first one. In fact, rather than being a product of n-gram probabilities involving all vocabulary words, it is a probability of distance of distributions limited to topic-dependent keywords.

An accurate selection of topic-dependent key-words is very important for the use which is made in the following. For this purpose, words that are frequent in more than one topic are not good key-word candidates. These words have been identified and placed into a *stop-list*. Stop words of the stop-list cannot become topic key-words.

2. Topic dependent probabilities

Let $P_{jC} = \sum_{i=1}^{C} P_1(w_i \mid T_j)$ be the sum of the *static* unigram probabilities in topic T_j. For the purpose of distance computation, the probabilities of these *C* words are multiplied by a topic-dependent constant γ_j. The floor probability of the other *L-C* words is :

$$P_{jr}(w) = \frac{1-\gamma_j P_{jC}}{L-C} \qquad (6)$$

In practice *C* has been set equal to 4,000.

Assuming there are *m* words in the cache, *G* of which are different, then the cache probabilities of these words are given by :

$$P_{cache}(w) = \frac{n(w)+\beta}{\beta L + m} \qquad (7)$$

Where β is a floor constant and $\sum n(w) = m$. The floor probability for words not in the cache is given by:

$$P_{flC}(w) = \frac{\beta}{\beta L + m} \qquad (8)$$

This corresponds to an estimate of the cache probability for the unseen words.

In order to have floor probabilities equal in the cache and in the normalized topic model, it is sufficient to impose:

$$P_{jr}(w) = P_{flC}(w); \frac{1-\gamma_j P_{jC}}{L-C} = \frac{\beta}{\beta L + m}$$
$$\gamma_j = \left\{1-\beta\frac{L-C}{\beta L+m}\right\}\frac{1}{P_{jC}} = \frac{1}{P_{jC}}(1-A) \quad (9)$$

The normalized probability for a word in topic T_j becomes:

$$P'(w_k \mid T_j) = \frac{n_j(w_k)}{\sum_{i=1}^{C} n_j(w_i)}(1-A)$$
$$= (1-A)f'(w_k \mid T_j); f'(w_k \mid T_j) = \frac{n_j(w_k)}{\sum_{i=1}^{C} n_j(w_i)} \quad (10)$$

where $n_j(w_i)$ is the count of the words in topic T_j. Better choices of the set of *C* keywords for a topic are possible. In fact, given a choice of keywords, distance between topic distributions can be evaluated and the choices can be modified by a search process that attempts to maximize the overall inter-set distance. Choices can also be based on tagging and be limited to words of certain syntactic classes.

3. Distance computation

Distance computation can start when G is above a threshold. The KL symmetric distance can be computed as follows :

$$d_j(n) = \sum_{i=1}^{C}\left\{P_{cache}(n, w_i) - Bf'(w_i|T_j)\right\}\log\left(\frac{P_{cache}(n, w_i)}{Bf'(w_i|T_j)}\right) \quad (11)$$

where $d_j(n)$ is the distance related to topic *j* after reading the *n*-th word. $P_{cache}(n, w_i)$ is $P_{cache}(w_i)$ at the insertion into the cache of the n-th word of the incoming document. All the words that are not keywords in any topic have equal floor probability in the topic-dependent distribution and in the cache, then their contribution to the KL distance is zero. All these words can be ignored in the computation of the second term of the (4).

A word, that is not a keyword for any topic and appears in the cache, would provide an equal contribution to the KL distance in all topics. For such a reason, words that are not keyword of any topic do not need to be stored in the cache. Figure 1 shows an example of unigram distributions in a topic and in the cache after normalization. As the cache is filled up, better estimates become available and can be used to produce more reliable distances. Coefficients α_1 and α_2 in the (4) must sum to 1. They can be determined by deleted interpolation as suggested in [4]. Coefficient α_2 can be made lower if the cache is not full. Its value may vary and reach the computed value only when the cache is full.

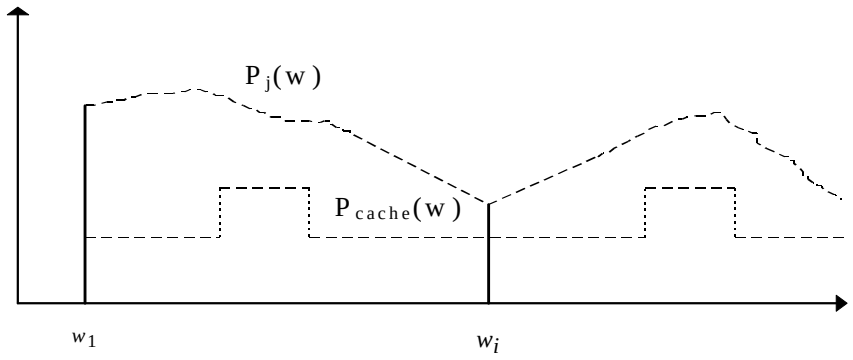

Figure 1 - Example of word probability distribution in a topic and in the cache.

It is worth noticing that, if the cache does not contain many words, it is possible that certain key-words of the right topic did not appear yet in the cache. If a zero probability would be assigned the these words in the cache, then an infinite distance would be obtained. This is avoided by the introduction of floor probabilities.

It is possible to find topic changes in the same article with the following approach. When there is a deep even short valley in the unigram score, a new cache is created and distances with it are also computed. If after some word, the second cache and the first cache agree on the topic, then the valley is considered as a false alarm, otherwise there is a topic switch. The contribution of the KL distances to the computation of a probability for a word can be of different types.

The parameter $P^{*}_{2j}(n) = 1 - \frac{d_j(n)}{D_{MAX}}$ where D_{MAX} is a normalizing constant, has been considered for the experiments described in the next section. It is interesting to see how rapidly this parameter reaches a plateau maximum value for the correct topic.

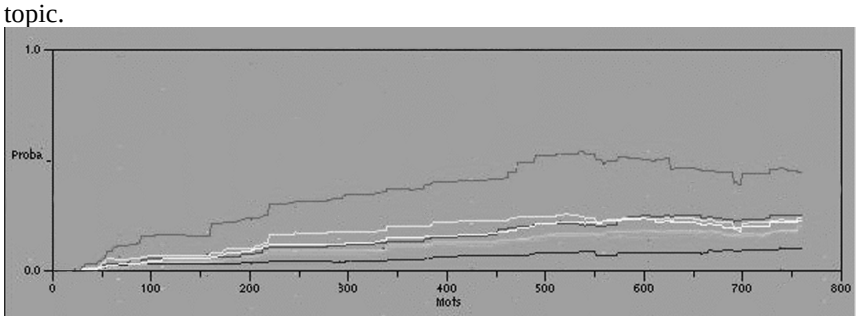

Figure 2 : Time evolution of $P^{*}_{2j}(n)$ for the seven topics. The correct one is the one with highest probability. The unit of the x axis is the number of words.

Figure 2 shows the evolution of $P_{2j}^{*}(n)$ as words of a test sample are considered. The highest curve corresponds to the correct topic. The other curves correspond to other topics. It can be seen that after 50 words, in this case, the correct topic emerges as the winner.

4. Results

Three years of articles from the French newspaper "*Le Monde"* were considered for a total of 60 Mwords, about 500,000 words of which are different. The topic of the articles is not known, but, as a first approximation, the section of the journal in which the article appears has been taken as topic. Seven topics were considered. The topic codes are shown in Table 1.

1	Foreign news	5	Business
2	History	6	Culture
3	Sciences	7	Politics
4	Sports		

Table 1: Topics and their codes

A test set of 1021 segments taken from articles was extracted from the corpus and not used for training. By just considering the topic with minimum distance, preliminary topic classification results are shown in Table 2.

Symbol = represents agreement, <> represents disagreement, *U* represents the topic selected by just unigrams, *C* represents the topic selected by just cache distances, *S* represents the following journal page header topics labeled, *C out* represents the case in which *G* was too small for using the cache.

	U<>S C<>S U<>C	U<>S C<>S U==C	U==S C<>S U<>C	U<>S C==S U<>C	U==S C==S U==C	U<>S C out	U==S C out	Total
1	4	10	29	13	75	2	6	*139*
2	0	0	1	0	6	0	0	*7*
3	19	4	2	17	48	1	0	*91*
4	15	7	12	57	21	0	2	*114*
5	2	3	16	6	129	0	3	*159*
6	15	21	55	24	244	6	16	*381*
7	15	21	18	18	51	3	4	*130*
Total	*70*	*66*	*133*	*135*	*574*	*12*	*31*	***1021***

Table 2: preliminary topic classification results in number of segments

Table 3 summarizes over all topic identification results on the test set for

different recognition strategies.

The first row refers to the case in which cache and unigrams agree with the label. In the second row, an additional rule is added for which unigrams agree with the label and the cache does not have enough data to be reliable. The third row corresponds to decision made with the unigrams only. The fourth row corresponds to the use of cache only except when the cache does not have enough data. The final row corresponds to decision made with the cache or the unigrams.

This mixed strategy consists in using the cache for decision only if there are enough words in the cache and if the difference between the cache scores of the first and the second candidate is greater than a threshold. Such a threshold depends of the first candidate and has been determined using the training set.

These preliminary results with a crude decision rule show a significant overall increase in performance when using the cache with respect to the use of unigram only. With the contribution of the cache, a precision greater than 78% is obtained.

Strategy	N	%
U=C=S	574	56.22
+ U=S, C out	605	59.26
+ U=S	738	72.28
First two + Cache	740	72.48
Strategy combination	*800*	*78.35*

Table 3 : Strategy combination results

5. Conclusion

The experiments described in this paper show the benefits using complementary two models and the fact that they may be used, when they agree, to refine the crude labeling. A wise combination of the two models may lead to a substantial improvement of topic classification.

It is worth mentioning the difficulty of the task, since topic classification is based on human decision which is often questioned by other humans. It is also important to notice that newspaper articles describe a very large variety of facts with a very large vocabulary (of the order of 500,000 words).

Research will continue along many directions. Lemmas instead of full words will be considered [1]. This will reduce the vocabulary size and the amount of gaps in the cache distribution. Various types of distance probabilities will also be considered in view of their use in adaptable LMs for Automatic Speech Recognition (ASR). Multiple cache memories will be introduced. Words will be dispatched to them based on the labels provided by a statistical tagger [7]. Other uses of tagged words will also be investigated.

6. REFERENCES

[1] M. El-Bèze, B. Mérialdo, (1997), "HMM Based Taggers" in *Syntactic Wordclass Tagging*, Edited by N. Ide and J. Veronis, Kluwer Academic Publishers.

[2] B. A. Carlson, (1996), "Unsupervised topic clustering of switchboard speech messages." *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 315-319. Atlanta GA.

[3] T. Imai, R. Schwartz, F. Kubala and L. Nguyen (1997) , "Improved Topic Discrimination of Broadcast News Using a Model of Multiple Simultaneous Topics", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 727-730. Munich, Germany.

[4] R. Kuhn, R. de Mori (1990), "A cache-based natural language model for speech recognition.", IEEE Trans. Pattern anal. Machine Intell., PAMI-12(6):570-582.

[5] H. Li and K. Yamamishi, (1997) "Document classification using a finite mixture model", *Proc. of the Conference of the Association for Computational Linguistics*, pp. 39-47, Madrid, Spain.

[6] B. Peskin, S. Conolly, L. Gillick, S. Lowe, D. McAllaster, V.Nagesha, P. van Mulbregt, S. Wegmann (1996), "Improvements in SWITCHBOARD recognition and topic identification.", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 303-306. Atlanta GA.

[7] T. Spriet, M. El-Bèze, (1997) "Introduction of Rules into a Stochastic Approach for Language Modeling", in *Computational Models of Speech Pattern Processing*, NATO ASI series F, edited by K. M. Ponting, Sprnger Verlag, Berlin New York.

[8] J.H. Wright, M.J. Carry and E.S.. Parris (1996) "statistical models for topic identification using phoneme substrings." *Proc. of the IEEE International Conference On Acoustics, Speech And Signal Processing*, pp. 307-310. Atlanta GA.