DETECTING TOPIC SHIFTS USING A CACHE MEMORY

Brigitte Bigi, Renato De Mori, Marc El-Bèze, Thierry Spriet
{bigi,demori,elbeze,spriet}@lia.univ-avignon.fr

LIA CERI-IUP
University of Avignon , BP 1228
84911 Avignon Cedex 9 - France

ABSTRACT

The use of cache memories and symmetric Kullback-Leibler distances is proposed for topic classification and topic-shift detection. Experiments with a large corpus of articles from the French newspaper "*Le Monde* show tangible advantages when different models are combined with a suitable strategy.

Experimental results show that different strategies for topic shift detection have to be used depending on whether high recall or high precision are sought. Furthermore, methods based on topic independent distributions provide complementary candidates with respect to the use of topic-dependent distributions leading to an increase in recall with a minor loss in precision.

1. INTRODUCTION

Statistical methods for topic spotting and the detection of topic shifts are now the object of research efforts because of their potential in important applications [6]. It is also possible to use topic-dependent language models (LM) for improving Automatic Speech Recognition (ASR) [3].

Topic spotting is a decision process applied to a document made of a sequence of words W_a^b whose result is a classification into a sequence of segments. Following classical game theory and decision models (see, for example, [5]), such a classification can be based on different strategies. The novelty proposed in this paper is a *group decision* with two contributions. A *pattern of preference* describes these contributions. For topic spotting, he first contribution proposes a topic preference for each sequence

W_1^{i-1} of the first *(i-1)* words $\{w_1,...w_{i-1}\}$ of a document using a unigram LM involving all the words of the vocabulary . The second one is based on a different computation which makes use of a set of keywords automatically selected for each topic. For each topic, a statistical distribution of topic keywords is obtained from a training corpus. Such a static distribution is continuously compared with the time-varying distribution of the content C_i of a cache memory [4] when word w_i is read in. The comparison is performed introducing, a symmetric Kullback-Leibler (KL) distance introduced in Section 2, which varies in time as new words are considered.

Another novel aspect is that an *intensity of preference* is associated to each contribution to the decision. This intensity depends on how reliably the contributing proposal to decision is made.

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 lexical entries (e.g. the singular and plural of the same noun). Let *L* be the size of such a vocabulary. Topic recognition is discussed in Section 3. For each topic, it is possible to rank the *L* words according to their frequency in a training corpus and select the first *K<L* words of this ordered sequence excluding 1800 words of a stop list. These first *K* words (4000 words for the examples described in this paper) are called *topic keywords*. The comparison of the cache content is limited only to topic key-words. Section 4 discusses methods and results for topic shift detection using topic-dependent and topic-independent distances.

2. DISTANCE COMPUTATION BETWEEN PROBABILITY DISTRIBUTIONS

Data sparseness is an unavoidable problem when statistical methods are used in Information Retrieval. A solution to the problem can be the introduction of methods for estimating the probability of occurrence of unseen words. Back-off and interpolation methods, described in the LM literature, can be used for this purpose.

These concepts are further developed in this paper by considering the cache content when a word is observed as the *context* for that word and by assigning the same default probability in the static and dynamic distribution to the *non-keywords* so that they do not influence a comparison. Assuming that there are *m* words in the cache, *G* of which are different, then the cache probability of a word w in the cache is given by the fraction of the relative frequency of appearances of that word in the cache. All the words not in the cache have probability proportional to their general unigram probability

$P_g(w)$ in all topics. The cache probability of each word w is given by:

$$P_{cache}(i,w) = \begin{cases} \beta \dfrac{n_i(w)}{m_i} & \text{if} \quad w \in C_i \\ \alpha P_g(w) & \text{if} \quad w \notin C_i \end{cases} \qquad (1)$$

$n_i(w)$ is the number of occurrences of word w in the cache at time i , m_i is the total number of items in the cache at the same time, α and β are normalization coefficients. Notice that the sum of all the $n_i(w)$ for the words in the cache is equal to m_i.

Let

$$P_{jK} = \sum_{k=1}^{K} P(w_k \mid T_j)$$

be the sum of the *static* unigram probabilities in topic T_j for the key-words in the topic.

The normalized probability for a word in topic T_j is expressed as:

$$P_j(w) = \begin{cases} \gamma_j P_1(w \mid T_j) \text{ if w is a keyword} \\ \alpha P_g(w) \text{ if w is not a keyword} \end{cases} \qquad (2)$$

and γ_j is another normalization coefficient. Coefficients α,β,γ are subject to constraints that probabilities properly sum to one:

$$\beta + \alpha F_i = 1; \qquad F_i = \sum_{w \notin C_i} P_g(w);$$

$$\beta = 1 - \alpha F_i \quad \gamma_j P_{jK} + \alpha P_{jg} = 1;$$

$$P_{jg} = \sum_{w \in NKJ} P_g(w); \qquad \gamma_j = \frac{1 - \alpha P_{jg}}{P_{jK}}$$

NKJ is the set of all the non-keywords for topic T_j. In principle F_i is time-dependent because the content of the cache varies in time. In practice, this term is always close to 1 since the cache contains a small number of words compared to L. It is possible to approximate F_i by its upper-bound F_{MAX} and consider

$\beta = 1 - \alpha F_{MAX}$. Once α has been set, β and γ are obtained with the above relations.

The KL symmetric distance between a topic distribution and the distribution of the cache content can be computed as follows :

$$d_j(i) = \sum_w \left\{ P_{cache}(i,w) - P_j(w) \right\} \log\left(\frac{P_{cache}(i,w)}{P_j(w)} \right)$$

Contributions to distance computation can be grouped into four cases:

1 $(w \in C_i) \wedge (w \in KJ)$ 2 $(w \notin C_i) \wedge (w \in KJ)$

3 $(w \in C_i) \wedge (w \notin KJ)$ 4 $(w \notin C_i) \wedge (w \notin KJ$

KJ is the set of keywords for topic T_j. Because of definitions (1) and (2), the fourth contribution is always zero and the computation is reduced to the union of words in the cache and topic keywords. As the cache size *m* is small (100 in our examples), distance computation involves a maximum of *K+m* addends as opposed to L, which is less than 10% of L in our case. Furthermore, absence of topic keywords in the cache is properly taken into account, resulting in a large difference between keyword and cache distributions.

The normalized distance

$$d_j^*(i) = \frac{d_j(i)}{d_j(0)}$$

is used to propose candidate topics for classification and candidate segments for the detection of topic shifts.

The just proposed formulation is an improvement of a previous one [1] and has produced better results described in the next section.

3. TOPIC CLASSIFICATION RESULTS

Three years of articles from the French newspaper "*Le Monde*" were considered for a total of 60 Mwords. 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 considered as the topic of the article, resulting in the set of Table 1.

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

Table 1: Topics

A test set of 1021 segments taken from articles was extracted from the corpus and not used for training. Table 2 summarizes topic identification results obtained for the test set and for different recognition strategies by setting the free parameter α = 0.1.

The first row refers to the case in which topic classification with the cache (C) and unigrams (U) agree with the classification based on the page of the journal (S). In the second row, an additional rule is added for which unigrams agree with S and the cache does not contain enough data to be reliable. The third row corresponds to decision made with the unigrams only. The fourth row corresponds to the use of the cache only except when the cache does not have enough data. The final row corresponds to a decision strategy with *a pattern of preference* that uses 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 on the first candidate and has been determined using the training set. The third and fourth columns in table 2 contain results obtained after replacing each word with its lemma. From a comparison of the first two columns, it appears that, when words are used, the number of cases in which the unigrams correctly agree with the cache is higher, the unigrams lead to better performance with lemmas but the use of the cache leads to better performance with words. This suggests combining words and lemmas which leads to the results shown in columns 5 and 6.

There, the same strategy is used with the unigrams computed with lemmas and cache probabilities with words.

	WORDS		LEMMAS		COMBINED	
Strategy	N	%	N	%	N	%
U=C=S	586	57,39	561	54.95	595	58.28
+U=S, Cout	644	63.08	650	60.66	652	63.86
+ U=S	752	73.65	754	73.85	754	73.85
1,2 + Cache	757	74.14	741	72.58	756	74.05
Comb strat	*825*	*80.8*	*819*	*80.22*	*828*	*81.1*

Table 2 : Strategy combination results in number of segments using words and lemmas, with unigrams and the cache

4. DETECTION OF TOPIC SHIFTS WITH TOPIC MODELS

In principle, a topic shift can take place at the end of each sentence indicating the end of a *segment* of text belonging to a topic.

A first type of methods locally compare adjacent blocks of sentences and use local rules to decide whether or not a block contains a topic shift.

Such an approach is followed in [2] where three methods are proposed. *Block comparison* considers running windows involving blocks with one or two sentences. Blocks are compared on the basis of a score s(i) involving the normalized inner product of vectors of word weights of two blocks one preceding and the other following the i-th sentence boudary. Valleys of s(i) are considered as cues for topic boundaries. *Vocabulary scores* are then considered to take into account the introduction of new words in each block. *Lexical chains* of words are also used.

A second type of methods considers segment boundary decision as a global optimization problem. An optimal sequence of segments between candidate boundaries is found as a search process in which candidate boundaries are scored and scores are combined leading to a cumulative score for every possible segmentation. Search can be based on Dynamic Programming (DP) for comparing a model for topic sequences with the text to be segmented. Statistical topic models are required for this computation and can be acquired with a training corpus.

In analogy with Automatic Speech Recognition (ASR), scores can be likelihoods, a sequence of segments is a sequence of recognized topics and the break points *e1, e1+1=b2,…, ex* are those for which the following probability is maximum:

$$P(T_1^X \mid W_1^N) = \frac{\prod_{x=1}^{X} P(W_{bx}^{ex} \mid T_x) \prod_{x=1}^{X} T_x}{P\left(W_1^N\right)} \qquad (3)$$

The probability $P(W_{bx}^{ex} \mid T_x)$ can be computed by using an LM trained with text belonging to topic T_x and $P(T_1^X \mid W_1^N)$ can be computed using a statistical *Topic Model* such as the one proposed by [3].

The probability $P\left(W_1^N\right)$ does not affect segmentation and can be ignored. Such an approach is proposed in [7] where 100 topics were obtained by clustering unigram LMs, one for each story of a training set.

The problem with such an approach is that the search process gives little importance to the context of a word in the text to be segmented, only one n-gram is used to score the relevance of that word, arbitrary pruning thresholds are introduced during search making it non-admissible with unpredictable effects, penalties have to be introduced to prevent oversegmentation with effects not yet completely evaluated.

The approach proposed in this paper emphasizes context comparisons by considering the whole content of the cache memory to represent contexts. Local comparisons are made by comparing the cache content with a topic LM or the cache content of two blocks. Global minimization of distances between keyword distributions is performed with DP matching and various strategies can be used for combining block-to-block or block-to-topic distances.

A first solution has been investigated that uses local rules for proposing candidate boundaries followed by DP matching to find a sequence of segments which may contain candidate boundaries but have to begin and end at the proposed boundaries. The purpose is that of assessing the suitability of the proposed distances for segmentation and to study the combined effect of local rules and global optimization. Let us consider:

$$j = \arg\min_k \left\{ d_k^*(i) \right\} \qquad B_j(i) = \frac{d_j^*(i)}{\sum_{k=1}^{J} d_k^*(i)}$$

Candidate boundaries are generated whenever

$\delta_j(i) = B_j(i) - B_j(i-1) \geq \vartheta$.

The cache is emptied at each candidate break-point leading to a more accurate account of the context describing each potential segment.

A corpus of 1400 articles in equal number for each topic and not used for training was selected from the year 1987. A segment with more than 300 words was extracted from each article; Segments were randomly combined into sequences of three segments each. DP matching is performed by considering only segments between successive pairs of adjacent candidate break-points. As it is possible that two successive segments are

now labeled with the same topic, topic shifts are detected when two different topic labels are found for two adjacent segments.

The solution is the one that maximizes the numerator of the (3) by assuming that all topic histories are equiprobable and:

$$\log P(W_{bx}^{ex} | T_x) = -\log B_j(ex)$$

By using the following definitions:

precision (p): the ratio between the number of correctly detected topic shifts and the total number of hypothesized shifts

recall(r): the ratio between the number of correct topic shifts and the total number of real topic shifts in the test corpus,

the values of *precision* and *recall* shown in Table 3 are obtained for two different values of θ.

Recall	0.931	0.9612	0.7457	0.8003
precision	0.2076	0.1586	0.7948	0.75

Table 3 – detection performance with rules (first two columns) and with DP (last two columns)

These results clearly show that different strategies have to be used depending on whether high recall or high precision are sought.

In order to evaluate the performance of distances that do not depend on topic classification, a different method was used to identify break-point candidates.

At the end e(i) of each sentence, the content of two cache memories CP(i) and CF(i) of 50 words each were considered. CP(i) captures the context just *before* e(i) while, CF(i) is the cache associated to a word following e(i) such that the first keyword following e(i) is the least-recently inserted word into the cache.

The distance KL(i) between the contents of CP(i) and CF(i) is computed and its behavior as function of *i* is considered. In particular, pairs of successive relative *minima* and *maxima* are considered together with their absolute distance values and their difference. Pairs of sufficient difference with maxima higher than a threshold are considered as *break-point events*.

If these events appear in a window centered on the real break-point, then the event is considered as a correct topic shift detector, otherwise it is considered as a false alarm.

Let *p1* and *r1* be respectively the precision and recall for the highest recall obtained with topic dependent distributions and *p2* and *r2* be the same indicators obtained with the just introduced topic independent break-point candidates. Table 4 shows a comparison of these indicators in the first two columns. If now, break-point candidates obtained with the two methods are merged if they appear within a given time window and the indicators *p3* and *r3* are evaluated using the resulting set of candidates, the last two columns of Table 4 are obtained.

r1	p1	r2	p2	r3	p3
0.9526	0.1625	0.6796	0.3925	0.9799	0.1543

Table 4: Precision and maximum recall for various methods

5. CONCLUSION

Various uses of KL distances and cache memories have been presented showing tangible advantages when different models are combined with a suitable strategy.

Different strategies for topic shift detection have to be used depending on whether high recall or high resolution are sought.

The results in Table 4 show that methods based on topic independent distances are not suitable for maximum recall, but they provide complementary candidates with respect to distances using topic-dependent distributions with a minor loss in precision.

6. REFERENCES

1. Bigi B., De Mori R., El-Beze M. and Spriet T., “Combined models for topic spotting and topic-dependent language modeling.” 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, Edited by S. Furui, B.H. Huang and Wu Chu, IEEE Signal Processing Society Publ., NY., pp.535-542, 1997

2. Hearst M.A. , “Texttiling: Segmenting text into multi-paragraph subtopic passages” Computational linguistics. 23(1):33-64, 1997.

3. Imai T., Schwartz R., Kubala F. and Nguyen L., "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, 1997

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

5. Luce R.D. and Raiffa H., *Games and Decisions*, Dover publ. New York, 1957.

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

7. Yarmon J.P. Yarmon, I. Carp, L. Gillick, S. Loweand P. van Mulbregt , “Event tracking and text segmentation via Hidden Markov Models” 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, Edited by S. Furui, B.H. Huang and Wu Chu, IEEE Signal Processing Society Publ., NY., pp.519-526.