A comparative study of Topic Identification on Newspaper and E-mail

Brigitte Bigi, Armelle Brun, Jean-Paul Haton, Kamel Smaïli, Imed Zitouni

A Comparative Study of Topic Identification on Newspaper and E-mail

Brigitte Bigi, Armelle Brun, Jean-Paul Haton, Kamel Smaïli and Imed Zitouni
LORIA/INRIA-Lorraine 615 rue du Jardin Botanique, BP 101,
F-54600 Villers-lès-Nancy, France
e-mail: {bigi, brun, jph, smaili, zitouni}@loria.fr

Abstract

This paper presents several statistical methods for topic identification on two kinds of textual data: newspaper articles and e-mails. Five methods are tested on these two corpora: topic unigrams, cache model, TFIDF classifier, topic perplexity, and weighted model. Our work aims to study these methods by confronting them to very different data. This study is very fruitful for our research. Statistical topic identification methods depend not only on a corpus, but also on its type. One of the methods achieves a topic identification of 80% *on a general newspaper corpus but does not exceed* 30% *on e-mail corpus. Another method gives the best result on e-mails, but has not the same behavior on a newspaper corpus. We also show in this paper that almost all our methods achieve good results in retrieving the first two manually annotated labels.*

1 Introduction

With the amount of available textual data increasing exponentially, automatically processing these data has become of central importance. This automatic treatment can be realized using topic identification (TID), which is our aim in this paper. The main objective of TID is to assign one or several topic labels to a flow of textual data. Labels are chosen from a set of topics fixed *a priori*. Several approaches have already been proposed in the literature [5, 8] which are generally based on a specific metric. This paper deals with the evaluation problem of TID algorithms on two kinds of textual corpora: newspaper and e-mails. A comparison study of several TID methods is presented. In the literature, another comparison has been performed by Yang and Liu[9]. However, this study treats classical text categorization methods. In this article, methods studied are mainly issued from statistical language modeling in speech recognition domain, thus different from text categorization methods.

TID has various applications : documents categorization, speech recognition systems, selecting documents for WEB engines, etc. Another promising direct application of TID is e-mail routing. This application consists of dispatching e-mail messages in accordance with their content. For example, a hot-line which receives a large number of e-mails per day would like to dispatch them automatically to several boxes. Each box corresponds to a specific problem to be solved, which can be considered as a topic.

2 E-mail topic identification

Routing e-mail messages is a direct application of TID. It amounts to identify an e-mail topic and send it to the appropriate operator. An e-mail has specific features which make it different from a newspaper article. In opposition to newspapers, it is not easy to find special e-mail corpora, for obvious confidentiality reasons. E-mails are often noisy and it is difficult to process them automatically. Several questions arise for e-mail TID. Which part of information should be kept? Should all the headings which constitute the structure of an e-mail be removed? Some of them could be very useful for detecting the topic as: *subject*, *date*, *sender*. In addition, e-mails are often ungrammatical, punctuation is usually missed, abbreviations are widely used, foreign words, images, web pages may be present, etc. In order to compute statistics, and since corpora are not large enough, we have to take into account word mistakes by correcting errors.

3 Evaluated models

This section introduces the five statistical methods involved in the TID experiments detailed in the forthcoming sections. Only some of them share the same vocabulary. In what follows, a sequence $W_1^N = w_1, ..., w_N$ to be classified is made up of the first N words of a document. Each document is associated with one single label among J predefined topics.

3.1 Topic unigram language model

The topic unigram language model [5, 8] is one of the most classical and standard ones. It is based on a counting of the number of occurrences of each word for each topic and involves all words of each topic vocabulary. The posterior probability $P(T_j \mid W_1^N)$ is expressed as:

$$P(T_j \mid W_1^N) = \frac{P(T_j)P(W_1^N \mid T_j)}{\sum_{k=1}^{J} P(T_k)P(W_1^N \mid T_k)} \quad (1)$$

where $P(T_j)$ is the *a priori* probability of topic T_j, and $P(W_1^N \mid T_j) = \prod_{t=1}^{N} P(w_t \mid T_j)$ is the likelihood of sequence W_1^N given a topic T_j. When a word w_t is not encountered in the training corpus of T_j, the model assigns an ε-probability. With this model, words not related to any topic have the same importance than those which are specific to a topic (keywords).

3.2 Cache model

The Cache model [1] is based on a set of keywords automatically selected for each topic. These words, called *topic keywords*, have a statistical distribution obtained from the training corpora. This static distribution is continuously compared with the time-varying distribution of the content C of a cache memory by introducing the symmetric Kullback-Leibler (KL) divergence $d_j(K, C)$ which varies in time when new words are considered. A special type of back-off scheme [4] is introduced in this divergence. The resulting definition of cache probability is:

$$P_c(w) = \begin{cases} \beta \frac{N(w)}{N_C} & if \quad w \in C \\ \alpha P_g(w) & if \quad w \notin C \end{cases} \quad (2)$$

where $N(w)$ is the number of occurrences of word w in the cache C, N_C the total number of words in the cache, α and β normalization coefficients and $P_g(w)$ a probability associated with all words not in the cache. The topic word probability, to which the content of the cache is compared, is given by :

$$P_j(w) = \begin{cases} \gamma_j P(w \mid T_j) & if\ w \in V(T_j) \\ \alpha P_g(w) & if\ w \notin V(T_j) \end{cases} \quad (3)$$

where $V(T_j)$ is the T_j topic keywords vocabulary, $P(w \mid T_j)$ the unigram probability of w in T_j, γ_j a coefficient of normalization depending on T_j and $P_g(w)$ the same probability as in equation 2, associated with non-keywords of T_j. At each time t, and for topic T_j, the symmetric KL divergence $d_j(t)$ is computed between the two precedent probability distributions (2) and (3):

$$d_j(t) = \sum_{i \in V} (P_c(w_i) - P_j(w_i)) log \left(\frac{P_c(w_i)}{P_j(w_i)} \right) \quad (4)$$

where $V = \{w \mid w \in (T_j \cup C)\}$. This distance is normalized by its value on an empty cache:

$$d_j^*(t) = \frac{d_j(t)}{d_j(0)} \quad (5)$$

The decision for assigning one or two labels is made by taking the one or two lowest distances (eq. 5).

3.3 The TFIDF Classifier

The TFIDF classifier [7] represents topics as vectors. Each one is characterized by a set of distinct words $D_j = (w_{j1}, w_{j2}, \ldots, w_{jn})$ where n is the number of words of the topic j and w_{jk} the weight of word k. w_{jk} is defined as $w_{jk} = nf_{jk}.idf_k$, where nf_{jk} is the term frequency, i.e. the number of times the word w_k occurs in the topic j. Let DF_k be the number of documents in which word w appears and $|D|$ the total number of documents. idf_k, the inverse document frequency, is given by: $idf_k = log\left(\frac{|D|}{DF_k}\right)$. This weighting function assigns high values to topic-specific words, i.e. words which appear frequently in one topic and in a limited number of topics. Conversely, it will assign low weights to words appearing in many topics or rare. The similarity between a topic j and a document represented by a vector D_i is measured by the following cosine:

$$sim(D_j, D_i) = \frac{\sum_{k=1}^{n} w_{jk} w_{ik}}{\sqrt{\sum_{k=1}^{n} (w_{jk})^2 \sum_{k=1}^{n} (w_{ik})^2}}$$

The selected topic is the one of highest similarity.

3.4 The perplexity-based model

The perplexity is a measure issued from information theory [4] which is widely used in speech recognition [6], but rarely in TID. Perplexity reflects the ability of a language model in modeling a text. It is computed as the inverse geometric mean of the likelihood of a text:

$$PP(W_1^N) = \left(P(w_1) \prod_{k=2}^{N} P(w_k \mid w_{k-1}, \ldots, w_{k-n+1}) \right)^{-\frac{1}{N}}$$

where W_1^N corresponds to the text on which perplexity is evaluated, N is the size of this text and n is the order of the language model (n-gram language model). In our experiment n has been set to 2 (bigram model). A good language model assigns a low perplexity to actual texts. In a TID framework, a language model is built for each topic. Then, perplexity corresponding to each topic is evaluated on the text to classify. The topic corresponding to the lowest perplexity is the one assigned to the text.

3.5 The weighted model

This model results from an adaptation of the one introduced in [2], it is specifically conceived to handle noisy and sparse data. As the Cache model, it computes distances between the text and the topics. Each topic is represented by a unigram and word weights. A topic-weight is assigned to a word according to a function inversely proportional to the number of topic-vocabularies in which this word is present. For instance, a word w_i appearing in five topic-vocabularies will have a topic-weight of $\frac{1}{5}$. The topic weight of an unknown word will be the inverse of the number of topics (i.e. $\frac{1}{J}$). The probability of unseen events is estimated with respect to each topic. Probabilities are computed for words for which the number of occurrences of which exceeds a predefined threshold. Words belonging to all vocabularies and function words are removed. In this model, a score is computed for a text W_1^N and for each topic j:

$$T_j(W_1^N) = \beta_j \frac{\sum_{k=1}^{N} LP(w_k \mid T_j)\, \eta(w_k)}{N} \qquad (6)$$

where w_k denotes the k^{th} word of W_1^N, $LP(w_k \mid T_j)$ the log probability of w_k in topic j, and $\eta(w_k)$ the weight assigned to w_k. β_j is the weight assigned to topic j:

$$\beta_j = \sqrt{\frac{\alpha_j}{\sum_{k=1}^{J} \alpha_k}} \qquad \alpha_j = \sum_{k=1}^{N(V_j)} \eta(w_k) \qquad (7)$$

$N(V_j)$ represents the number of words of the j^{th} topic-vocabulary. The resulting topic is the one corresponding to the highest value of $T_j(W_1^N)$.

4 Experimental Results and Discussion

Two corpora are used in this article: the first corpus is made up of four years (1987-1991) of the French newspaper *Le Monde* (over 80 M words). This corpus has been divided into 7 topics. The different topics and their vocabulary size are: *Foreign News* (173 K), *History* (37 K), *Science* (66 K), *Sport* (16 K), *Business* (Business and Economy) (167 K), *Culture* (Culture, Art, Books and Media) (274 K), *Politic* (Politic and Ideas) (133 K). The test data are made up of more than 1500 paragraphs. These paragraphs have been manually labelled. Due to the ambiguity concerning the topic of several paragraphs, one or two labels have been assigned to each paragraph, without assigning a priority to these labels. The second corpus is made up of 5,000 e-mails which have been supplied by the French start-up MIC2. This corpus consists of the users requests of an Internet provider. It has been split manually into ten topics. Training data contain 90% of the corpus and left data are reserved for the test.

Each of the five methods have been evaluated at first on newspaper, and then on the e-mail corpus. Because of the ambiguity of some paragraphs, we carried out two different experiments on newspaper corpus. The first one concerns paragraphs having only one label (noted $\{t_1 = t_2\}$) and the second one concerns paragraphs having two different labels (noted $\{t_1, t_2\}$). Column 1 of Table 1 gives the results on paragraphs having only one label (representing 55% of the total number of paragraphs). The term $\{x_1\}$ is the label proposed respectively by a model. Results show that, on large corpora, the cache model provides the best results (82% of correct TID). Nevertheless, performance of basic methods like unigram or perplexity (79% and 80% respectively) is comparable to cache model. We think that these results are due to the large size of training corpus which enables a reliable estimation of the probabilities.

Columns 2 and 3 represent respectively the case where the two best labels identified by the methods are the two ones of the paragraph and the case where at least one of the two topics identified by the models belongs to the topics of the paragraph. We remark that unigram, cache and perplexity are still the best methods. We can notice that during the training phase our models have not been optimized to detect the two best topics, but only the best one, therefore results in column 2 are probably underestimated. All methods presented here give good results. In column 3, the identification rate of at least one label shows a performance range between 92% and 97%.

Models	$\{x_1\} = \{t_1 = t_2\}$	$\{x_1, x_2\} = \{t_1, t_2\}$	$(x_1 \vee x_2) \in \{t_1, t_2\}$
Topic U.	80.1%	45.7%	96.5%
Cache Model	82.0%	48.8%	96.6%
TFIDF	72.2%	29.9%	92.4%
Perplexity	79.0%	47.7%	97.1%
Weighted U.	74.1%	40.3%	95.0%

Table 1. TID performance on newspaper

Table 2 presents the identification rates on the e-mail corpus. As expected, results are worse than on the newspaper corpus. The weighted unigram is the only method giving good results, which springs from the fact that this method has been specifically designed for e-mails. The accuracy of the other methods drops on this difficult corpus. This may result from the fact that most of these methods are based on frequency computations which cannot be reliable in the case of e-mail corpora, due to noisy words, etc. To illustrate this phenomenon, suffice it to say that 66% of words in the e-mail corpus have been met only once, against 41% in the newspaper corpus. This rate can reach 74% in the first case, whereas it does not exceed 46% in the second one.

Models	TID	Models	TID
Topic U.	30.0%	Perplexity	32.5%
Cache Model	52.5%	Weighted U.	67.5%
TFIDF	42.5%		

Table 2. TID performance on e-mail corpus

Topic	word	DP	Topic	word	DP
F. News	War	14	Sports	Prost	43
F. News	Year	14	Sports	Season	14
History	War	14	Sports	Team	43
History	Politics	14	Sports	Senna	86
Politics	RPR	14	Culture	Movie	14
Politics	PS	14	Business	Billions	14
Politics	Mitterand	57	Business	rate	14
Science	Doctors	14			

Table 3. Most discriminant words

5 Effects of vocabularies

In this section, a brief analysis of the vocabulary is presented. Considering topic-vocabularies made up of the most frequent words of the topic-training corpora, we study the discriminant power $DP(w)$ of the ten most frequent words of each vocabulary. $DP(w)$ is evaluated as:

$$DP(w) = (1 - \frac{j(w)}{J}) * 100$$

, where $j(w)$ is the number of topics in which the word w appears. Table 3 presents the most discriminant words of each topic, among the ten most frequent ones. Some words are discriminant but others, in spite of their relative high discriminant power, are not semantically very representative of a topic. Words as *RPR* and *PS* (French political parties) should be more discriminant in identifying French politics. Their weights may be increased in order to be more characteristic of this topic. The noun "Senna" is very discriminant of Sports topic and can help efficiently the identification. We regret the little use of this very discriminant word. A word like *Billions* is not really characteristic of the Business topic, but its combination with other words could constitute a good business topic detector. We have to try to find how to combine words in order to have more discriminant expressions. An idea consists of the use of models based on typical word sequences.

6 Conclusion

In this paper, five methods for TID have been presented. These methods show tangible advantage when applied to two different corpora: newspaper articles and e-mails. It is also important to notice that, on one side, newspaper articles describe a very large variety of facts with a very large vocabulary. On the other side, e-mail corpora describe a very small variety of subjects with a restricted vocabulary, often insufficient to correctly learn statistical models. The methods presented have different behaviors depending on the type of the corpus. The weighted model has been developed especially for sparse corpora, which explains why it gives the best e-mail TID. The method based on cache achieves the best result for newspaper TID. Unfortunately, these methods were not sufficiently powerful on e-mails due to unreliable statistics computed on the corpus.

This paper also shows the importance of vocabularies. Some words could play a more important discriminant role, but they have been discarded by some methods which reduce their importance when they appear in several vocabularies. This work is very positive in terms of results, we need now to improve the choice of the vocabulary and to adapt the word weights in order to make some words more discriminant. One way is to introduce the concept of phrase units into the vocabularies. Future work will also be devoted to improve methods independent of the type of corpus.

References

[1] B. Bigi, R. De Mori, M. El-Bèze, and T. Spriet. A fuzzy decision strategy for topic identification and dynamic selection of language models. *Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal*, 80(6), 2000.

[2] A. Brun, K. Smaïli, and J. P. Haton. Experiment analysis in newspaper topic detection. In *String Processing and Information Retrieval - SPIRE, IEEE Computer Society*, Sep 2000.

[3] F. Jelinek. Self-organized language modeling for speech recognition. *Readings in Speech Recognition, A. Waibel and K-F. Lee editors*, pages 450–506, Morgan Kaufmann, San Mateo, Calif., 1990.

[4] F. Jelinek, R. Mercer, L. Bahl, and J. Baker. Interpolated estimation of markov source parameters from sparse data. *Pattern Recognition in Practice*, pages 381–397, 1980.

[5] J. McDonough and K.Ng. Approaches to topic identification on the switchboard corpus. In *Proceeding of the International Conference on Acoustics, Speech and Signal Processing*, pages 385–388, 1994.

[6] K. S. S. Matsunaga, T. Yamada. Task adaptation in stochastic language models for continuous speech recognition. In *ICASSP*, pages 165–168, 1992.

[7] G. Salton. Developments in automatic text retrieval. *Science*, 253:974–980, 1991.

[8] R. Schwartz, T. Imai, F. Kubala, L. Nguyen, and J. Makhoul. A maximum likelihood model for topic classification of broadcast news. In *Proceeding of the European Conference On Speech Communication and Technology*, Rhodes, Greece, 1997.

[9] Y. Yang and X. Liu. A re-examination of text categorization methods. In *International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1999.