A Hierarchical Approach for Topic Identification

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

A Hierarchical Approach for Topic Identification

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

LORIA - Campus scientifique - BP 239 - 54506 Vandœuvre-lès-Nancy
phone: +33/0 3 83 59 20 97- fax: +33/0 3 83 41 30 79 - e-mail: {bigi, brun, smaili, jph}@loria.fr

Abstract: This paper focuses on language model adaptation, and more especially on topic identification (TID) for Automatic Speech Recognition (ASR). The structure of a set of topics is redefined by the introduction of a hierarchy. TID models may then make use of the semantic relationships between parent and son nodes of the topic-tree. The originality of the approach presented in this article lies in the allocation of a unique vocabulary to brother nodes, which rests on the use of two backing-off levels. In comparison with TID performance when using a non-hierarchical approach, results encourage us to carry on in this way.

INTRODUCTION

Statistical language modeling (LM) attempts to capture regularities of natural language. It is a crucial part of a large variety of language technology applications and, among other things, of Automatic Speech Recognition (ASR). In our research, we are involved in *topic adaptation* which is driven by identifying the topic of new data and adapting the language model toward that topic. In this context, a topic represents a subset of the language. The topic adaptation often consists in performing a linear combination (5, 1) of a general language model and the appropriate topic language model. Statistical methods for designing topic language models in ASR systems have been extensively studied (4, 11, 3). Recent researches have proved that topic adaptation of language models in ASR improves perplexity (7, 9, 6).
The main objective of *topic identification* (TID) is to assign one or several topic labels to a flow of data. Labels are chosen from a set of topics fixed *a priori*. Thus, the number of topics can vary in a wide range, from 8 (12) to more than 5,000 (11). We assume that coping with the problem of granularity difference of topics should induce an improvement of TID performance. This paper redefines the structure of a set of topics, by introducing a hierarchy allowing an exploitation of the semantic relationship between parent (topics) and son nodes (sub-topics), and particularly brother relations. These semantic relations can become an invaluable advantage for assigning a topic more precisely.
The originality of this paper lies in the way to create vocabularies and their probability distributions. During the training phase, common vocabularies are alloted to brother nodes, in order to usefully take into account their brotherhood. In a second time, words probability distribution in each topic have to be estimated. This model rests on two backing-off levels. In the test phase, one or several topics are assigned to new textual data. In this application, the aim is to find the best newsgroups in which a specific message has to be posted. To evaluate the improvement of performance induced by the use of the topic-tree dependent model, results are compared with a classical TID model using the same topics without hierarchical structure. Experiments are carried out on data extracted from the French newsgroups of UseNet.

HIERARCHICAL TOPIC IDENTIFICATION

Motivations

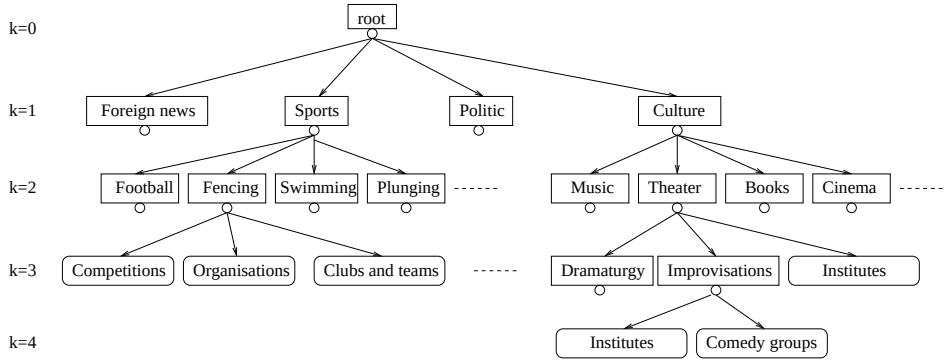
This Section presents the topic identification problem when exploiting semantic relations between topics. Intuitively, it seems interesting to specify that *football*, *fencing*, *swimming* or *diving* are subtopics of *Sport*. Some words have the same chance to be observed (the same probability) in these subtopics (for example: *tournament*, *match*, *meets*, *referee*, etc, which all relate to *Sport*). However, these words have different probabilities in other topics. Other words, as for them, are specific to one sub-category (for example: *goal*, *sword*, *shirt*, *snorkel*). Their probability differs in these sub-categories. Given a document, determining in a first step, that *sport* topic is the main topic, and in a second one choosing which specific sport is related, has apparently two advantages. At first, it is possible to penalize some "rival" categories of the level, which may avoid some semantic ambiguities. As an example, the French word *tuba* represents a musical instrument (topic *Culture*) and a *snorkel* (topic *Sport*). Second, by scanning the hierarchy breadth-first, topics of a given level are treated in the same step. This type of scanning may cope with the classification problem of differences of granularity between topics, present in a non-hierarchical structure.
Hierarchical TID can be used in several areas: information retrieval (8,10), speech recognition (11), etc. In (8), the hierarchical topic structure is used to decompose the problem into a set of simpler problems, which induces the

opportunity to develop more complex models. In (11), the text being decoded is classified into multiple topics and topic language models are interpolated, obtaining a new language model. However, the hierarchy is not exploited for TID but only for language model adaptation. So, it seems that hierarchies are becoming an increasingly prevalent way for many domains. But to take advantage from these structured data, developing original models is becoming necessary.

Suggested Solution

The hierarchy of topics can be represented by a tree (Figure 1). At each level, groups of brother nodes (noted b) are defined, brother nodes are nodes with a common father. Let us denote T_{jbk}, the j^{th} topic of the b^{th} brother group at the k^{th} level, and T_{bk} a group of brothers. One advantage of this representation is that all topics of a given level have a comparable granularity. Consequently, it seems natural to explore the tree breadth-first to find the best topics of a given document at each level. In the test phase, most probable topics of a text $W_1^N = \{w_1, w_2, ..., w_N\}$ have to be found: at each level of the hierarchy, the probability of each node (or leaf) is evaluated. Afterwards, an algorithm will be used to find the most relevant topics in the hierarchy of the text W_1^N.

FIGURE 1. Example of a topic-dependent tree



Let us denote V_{jbk} the vocabulary of T_{jbk}. As a matter of fact, we take advantage from the strong semantic relationship between brother topics in order to share the same vocabulary. This vocabulary is denoted V_{bk}. This common vocabulary makes it possible to set up an adapted backing-off scheme between brother nodes. For instance, in level $k = 2$ of Figure 1, differences between *swimming* and *music* are boosted in comparison to *swimming* and *fencing* which are not.

Vocabularies and Probability Distributions in Topics

During the training phase, topic vocabularies V_{jbk} probability distributions have to be estimated. In classical models, each topic-vocabulary is made up of all words occurring in the corresponding topic training corpus. The probability of a word w given a topic T_j is evaluated as follows:

$$P(w \mid T_j) = \begin{cases} \gamma_j f(w \mid T_j) & if\ (w \in V_j) \\ \varepsilon & else \end{cases} \quad (1)$$

where $\sum_{w_i \in V_j} \gamma_j f(w \mid T_j) + \varepsilon = 1$.

In this paper, the model proposed defines topic vocabularies extracted from each cell of the tree containing data. Topic vocabularies T_{jbk} are made up of each word occurring in the training corpus. Then, vocabularies are modified to assign identical vocabularies to brother nodes T_{bk}. Moreover, vocabularies of father nodes are made up of the union of the vocabularies of their sons, plus the vocabulary of their own training corpus.

The main problem to solve is the attribution of a probability distribution to each topic. First, topic frequency $f(w \mid T_{jbk})$, where $w \in V_{jbk}$ is collected. The corresponding probability is evaluated as:

$$P(w \mid T_{jbk}) = \begin{cases} \gamma_{jbk} f(w \mid T_{jbk}) & if\ (w \in V_{jbk}) \\ \omega_{bk} & if\ (w \in V_{bk}) \\ \varepsilon & else \end{cases} \quad (2)$$

where γ_{jbk} is a normalization coefficient. ω_{bk} represents the first backing-off level, which takes into account all vocabulary words which have a zero-frequency. ε is the probability of the unknown word (UNK), which corresponds to the second backing-off level. The estimation of ε, ω_{bk} and γ_{jbk} is remaining and now discussed.

The model has to respect the following constraint: $\sum_{w_i \in V} P(w_i \mid T_{jbk}) = 1$, where $\{V = V_{bk} + UNK\}$.
This implies:

$$\sum_{w_i \in V_{jbk}} \gamma_{jbk} f(w_i \mid T_{jbk}) + \sum_{w_i \in V_{bk}; w_i \notin V_{jbk}} \omega_{bk} + \varepsilon = 1$$

As there are no UNK words in the training, $\sum_{w_i \in V_{jbk}} f(w_i \mid T_{jbk}) = 1$, consequently:

$$\gamma_{jbk} = 1 - \varepsilon - \sum_{w_i \in V_{bk}; w_i \notin V_{jbk}} \omega_{bk} \tag{3}$$

The ω_{bk} value will be set to a rate of the minimum brother probabilities. That means, for each word which does not belong to a particular topic but belongs to its brother topics, a uniform probability is alloted. The ε value, as for it, will also correspond to a low value (smaller than any ω_{bk}).

Retrieving Document Topics

In this article, a topic unigram (1) has been implemented. The probability of a topic T_{jbk} given a document W_1^N is evaluated as:

$$P(T_{jbk} \mid W_1^N) = \frac{P(T_{jbk}).P(W_1^N \mid T_{jbk})}{P(W_1^N)} \tag{4}$$

where $P(T_{jbk})$ is the *a priori* probability of topic T_{jbk}, and:

$$P(W_1^N \mid T_{jbk}) = \prod_{n=1}^{N} P(w_n \mid T_{jbk})$$

is the probability of the sequence of words $W_1^N = w_1, \ldots, w_n$, evaluated as the product of probabilities of each topic word T_{jbk}.

EXPERIMENTS

There currently exists many hierarchical bases which are continuously managed manually. So, it is possible to directly use these hierarchies. At the present we only aim to show that a hierarchy introduced in the training phase improves TID performance. In future works, a more suitable corpus for ASR will be used. In this experiment, we use newsgroups «fr» of UseNet. Newgroups are forums federated by topics, where, throughout one time given, all the messages sent are preserved. Each message of a newsgroup is named a *news* and includes a header and a body. The number of newsgroups is about 23,000 by the world. The French-speaking number of groups we use is over 300, covering a period of several month, which represents a total of 2 Go of data (more than 1 million news). These data are very corrupted (many errors, attached fi les, multi-languages, etc) which require many preprocessings. In this way, the most signifi cant «noises», like attached fi les or images, are cleaned from data. Thereafter, the text is tokenized into words and some words are gathered into classes by using preexistent lexicons (name, punctuation, country, town), or by using syntactic information (e-mail, net address, hour, price, smiley, number).
The test corpus is made up of several days of the newsgroups (110 Mo, about 60,000 news). Given a news, our aim is to propose the set of topics that fi t the best that news. Two cases are highlighted in this paper: when models propose one topic or several topics [1], noted *1_topic* and *10_topic*. Then, these topics are compared with the newsgroups where the news has been sent. Results are given in terms of recall and precision ratios, where:

$$Recall = \frac{Number\ of\ topics\ correctly\ detected}{Number\ of\ topics\ to\ find}$$

$$Precision = \frac{Number\ of\ topics\ correctly\ detected}{Total\ number\ of\ topics\ detected}$$

On our corpus, the classical unigram resulted in a recall of 0.35 and a precision of 0.38 for the *1_topic* experiment and a recall of 0.67 and a precision of 0.07 for the *10_topic* experiment. These values are slightly increased by the

[1] a value of 10 has been arbitrarily chosen.

use of the hierarchical topic unigrams, with a recall of 0.37 and a precision of 0.40 for the *1_topic* experiment. The signifi cance of these low values can be temperate. First, data are corrupted and a unigram model has been proved to be not robust for this kind of data (2). Moreover, training corpora are much disparate in terms of size (from 1 to 40,000 news). Unigrams are based on frequency computations which may not be reliable in some newsgroups. In Table 1, the *1_topic* hierarchical topic unigrams performance is presented. In this table, two sets of groups are studied: the fi rst one, which relates to linux os does perform well and the second one, which relates biology which does not perform well. Differences between these two groups can be explained by the size of their vocabularies compared to the number of news in the training corpus. We can remark that biology is of level $k = 2$ and linux os of level $k = 4$, which implies a difference of granularity. Future works will reduce this diffi culty by taking into account the hierarchy in the test phase.

TABLE 1. Topic Identification Performance of classical unigrams

Newsgroup	Recall	Precision	Nb news training	Nb news test	V_{jbk}
fr.comp.os.linux.annonces	0.83	0.83	124	7	7954
fr.comp.os.linux.confi guration	0.94	0.97	16111	931	4536
fr.bio.pharmacie	0.10	0.11	684	35	10363
fr.bio.medecine	0.34	0.46	14316	490	44970

CONCLUSION

This paper suggests the use of a topic hierarchy in order to improve topic identifi cation and language model adaptation tasks. The use of a hierarchy allows the exploitation of semantic relations between brother and parent nodes. The originality of this article rests on an original mean to create topic vocabularies. As a matter of fact, we take advantage from the strong semantic relationship between brother topics in order to share the same vocabulary. The fi rst results obtained are promising and confi rm that the hierarchisation of the topics can prove to be effi cient. It encourages us to continue in this way and to develop solutions to take into account the hierarchy in the test phase. Other models are being studied to better take into account the hierarchy, and then improve performance.

REFERENCES

(1) Bigi, B., De Mori, R., El-Bèze, M. and Spriet, T., «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) Bigi, B., Brun, A.,Haton, J.P., Smaili, K. and Zitouni, I., «A Comparative Study of Topic Identification on Newspaper and E-mail», *IEEE International Conference on String Processing and Information Retrieval*, (2001).

(3) Chen, S., Seymore, K. and Rosenfeld, R., «Topic adaptation for language modeling using unnormalized exponential models», *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, 681-684, (1998).

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

(5) Iyer, R. and Ostendorf, M., «Modeling long distance dependence in language: topic mixtures vs. dynamic cache models», *IEEE Trans. Speech Audio Process. SAP-7*, **1**, 30-39, (1999).

(6) Khudanpur, S.P. and Wu, J., «A Maximum Entropy Language Model Integrating N-Gram and Topic Dependencies for Conversational Speech Recognition», *IEEE International Conference on Acoustics, Speech and Signal Processing, Phoenix, Arizona*, I, 2192, (1999).

(7) Kneser, R. and Peters, J., «Semantic clustering for adaptive language modeling», *IEEE International Conference on Acoustics, Speech and Signal Processing*, II, 779-782, (1997).

(8) Koller, D. and Sahami, M. ,«Hierarchically classifying documents using very few words», *Proceedings of the Fourteenth International Conference on Machine Learning* , (1997).

(9) Martin, S.C., Liermann, J. and Ney H., «Adaptive Topic-dependent language modeling using word-based varigrams», *Proceeding of the European Conference On Speech Communication and Technology*, (1997).

(10) McCallum, A., Rosenfeld, R., Mitchell, T. and Ng, A.y., «Improving Text Classification by Shrinkage in a Hierarchy of Classes», *International Conference on Machine Learning*, (1998).

(11) Seymore, K. and Rosenfeld, R., «Using Story Topics for Language Model Adaptation», *Proceeding of the European Conference On Speech Communication and Technology*, (1997).

(12) Yamashita, Y., Tsunekawa, T. and Mizoguchi, R., «Topic recognition for news speech based on keyword spotting», *IEEE International Conference on Spoken Language Processing*, Sydney, Australia, paper 23, (1998).