IMPROVED TOPIC DISCRIMINATION OF BROADCAST NEWS USING A MODEL OF MULTIPLE SIMULTANEOUS TOPICS

Toru Imai, Richard Schwartz, Francis Kubala, and Long Nguyen*

*NHK (Japan Broadcasting Corp.) Sci. & Tech. Res. Labs., Tokyo 157, Japan
BBN Systems and Technologies, Cambridge, MA 02138, USA
imai@strl.nhk.or.jp

ABSTRACT

This paper presents a new method of topic spotting that attempts to retrieve detailed multiple simultaneous topics from broadcast news stories, each of which has about four different topics out of several thousand different topics. A new topic model uses a simple HMM where each state of the HMM represents one topic and the topic state emits topic-dependent keywords probabilistically. The model allows (unobserved) transitions among topics, word by word. These characteristics improve the discriminative ability between keywords and general words in a topic model and decrease the probabilistic overlap among the topic models more than the conventional topic models (such as a simple multinomial probability model). In addition, the model is not confused by words from multiple topics within one story. We applied the new method to topic spotting from manually transcribed texts of news shows. The new method showed better results in precision and recall rates than the conventional method.

1. INTRODUCTION

To retrieve topics automatically from broadcast news is useful for skimming, categorizing, or information retrieval. This paper presents a new method of topic spotting which attempts to retrieve detailed multiple topics from broadcast news stories. The transcriptions we deal with are produced by Primary Source Media as Broadcast News CD-ROM [1]. While previous research on topic classification [2-7] focused on selecting one topic out of a short list of tens of topics, each news story is manually labeled with about four different topics out of several thousand different topics. For example, a story about U.S. policies on loans to Mexico is labeled with four topics: "Clinton, Bill", "Mexico", "Money", and "Economic assistance, American".

The most conventional method of topic spotting is a simple multinomial probability model combined with some keyword selection methods [2, 6, 7]. Usually the topic model is obtained by counting the number of words in training stories after selecting keywords in the task by some methods. However, since each topic model is trained and modeled independently and all words in a story are assumed to be relevant to the topic, the model causes three problems. One is low discriminative ability between keywords and general words in the topic model. The second is the probabilistic overlap among the topic models. Third, the model is confused by words from other topics within the same story.

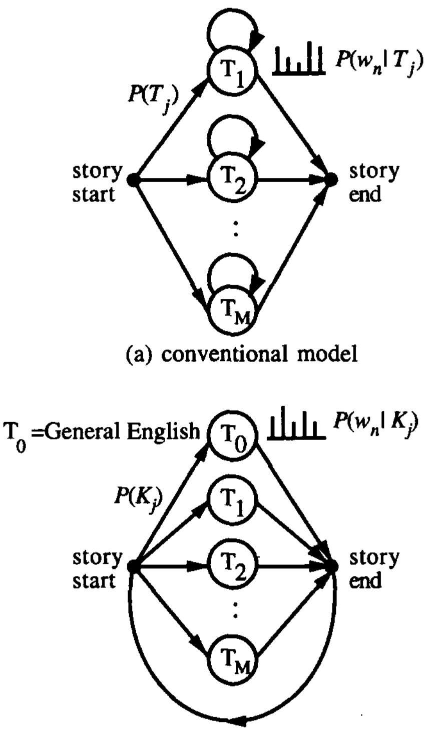
The proposed method uses a mixture model, in which we acknowledge that a story has several topics, and any particular word need not be related to all of the topics. In fact, most words are just general English. These characteristics are expected to overcome the problems of the conventional method. The probabilities of relevant keywords are increased, even if keywords occur only a few times, while the probabilities of general words are absorbed by a general English model. This results in decreased overlap between topics, and increased robustness, because keywords from one topic are no longer treated as negative evidence for another topic.

Section 2 describes the proposed new model. How to train the model is discussed in Section 3. In Section 4, we describe a multiple-topic spotting method. Section 5 presents topic spotting experiments from broadcast news stories. Section 6 concludes with some remarks.

2. TOPIC MODEL

In order to estimate multiple topics from a story, the conventional model [2] can be expanded to a finite state network of topic models with *a priori* probability, $P(T_j)$ (Figure 1a). Each state of a topic, T_j, emits all words in a story according to their probabilities, $P(w_n|T_j)$. The probabilities are obtained by counting the numbers of words or stories about the topics. The assumption that each word is related to *all* relevant topics for a story results in poor discrimination between keywords and general words, and the probabilistic overlap among the topic models, because topics that occur in the same story share the words in that story.

In a story with multiple topics, however, some words are related to one topic, while others are related to another topic, and most are related to none of the topics. The proposed method allows each word in a story to take different topics statistically. Thus, it is a mixture model of topics (a simple HMM: Hidden Markov Model [8]) where the sequence of topic states can not be observed and it is assumed that the state transitions are independent from previous topic states. The proposed model is shown in Figure 1b. There is a special topic of "General English" which is supposed to produce general words like "go" or "think". The network loops back after generating each word so it can transit from topic to topic with each word.



(a) conventional model

(b) proposed model

Figure 1: Topic models.

3. TRAINING

To train the proposed model, the EM (Expectation and Maximization) [9] algorithm is applied. What can be observed from a training story is a sequence of words and a set of unordered labeled topics. What can not be observed is the sequence of topics according to the word sequence. The EM algorithm attempts to maximize the expected likelihood of the training stories. The objective function is

$$\prod_i P\left(Story_i | Set_i\right)=\prod_i \prod_t P\left(w_{it} | Set_i\right), \tag{1}$$

where Set_i is a set of labeled topics of story i, $w_{it} \in \{w_n\}$ is the t-th word in story i, and word independency is assumed.

We define $P(K_j)$ as the probability that any particular word in a story about topic j will be a word that is directly related to that topic (a keyword). Then we define $P(w_n| K_j)$ as the probability that a keyword for topic j will be word n. Note that one of the topics is always general English, T_0, and the probability that it will generate a keyword is quite high.

Therefore the probability of a word w_{it} given the correct topic set is

$$P\left(w_{it} | Set_i\right)=\sum_{j \in Set_i} P\left(K_j\right) \cdot P\left(w_{it} | K_j\right) \Big/ \sum_{j' \in Set_i} P\left(K_{j'}\right). \tag{2}$$

To train this model, we distribute the count for each word in the training, w_n, to topics, and sum the count over the training stories.

$$C\left(w_n, K_j\right)=\sum_i \sum_t \delta_{w_n, w_{it}} \frac{P\left(K_j\right) \cdot P\left(w_{it} | K_j\right)}{\sum_{j' \in Set_i} P\left(K_{j'}\right) \cdot P\left(w_{it} | K_{j'}\right)}, \tag{3}$$

where $\delta_{l, m}$ denotes the Kronecker delta function.

The models, $P(w_n| K_j)$ and $P(K_j)$, are reestimated from $C(w_n, K_j)$ iteratively.

$$P\left(w_n | K_j\right)=C\left(w_n, K_j\right) \Big/ \sum_n C\left(w_n, K_j\right) \tag{4}$$

$$P\left(K_j\right)=\sum_n C\left(w_n, K_j\right) \Big/ \sum_n \sum_j C\left(w_n, K_j\right) \tag{5}$$

When the likelihood (1) converges after some EM iterations, the training stops.

Since (5) gives a keyword probability about topic j given "any" story, we convert it to a keyword probability about topic j given "a story about topic j" according to

$$P\left(K_j\right) \leftarrow \frac{P\left(K_j\right) \cdot \text{\#word in all stories}}{\text{\#word in stories about topic } j}. \tag{6}$$

4. SPOTTING

In the conventional model, each topic was independently measured and the topic that gives the highest score was considered as the retrieved topic.

$$P\left(T_j | Story\right)=P\left(T_j\right) \cdot P\left(Story | T_j\right) / P(Story) \tag{7}$$

By assuming word independency, the log score can be represented as

$$\log P\left(T_j | Story\right)=\log P\left(T_j\right)+a \sum_t \log\left(P\left(w_t | T_j\right) / P\left(w_t\right)\right), \tag{8}$$

where a is a weighting parameter for the assumption.

Our task in spotting is to determine which of several thousand different topics, in addition to general English, "produced" the words in a new story. We expect only a small percentage of the words to be directly attributable to each of the topics.

In principle, to recognize the most likely set of topics, Set, for a new story, we must consider all possible sets of topics, and for each one, compute

$$\begin{aligned} P(Set | Story) &= P(Set) \cdot P(Story | Set) / P(Story) \\ &\approx P(Set) \prod_t \frac{\sum_{j \in Set} P\left(K_j\right) \cdot P\left(w_t | K_j\right)}{\sum_{j' \in Set} P\left(K_{j'}\right) \cdot P\left(w_t\right)}. \end{aligned} \tag{9}$$

To calculate (9) for all sets of topics out of several thousand is not realistic. Instead, we first evaluate each topic independently using (10) and then rescore all subsets of the top-N topics using (9). The log score for each topic is calculated by the equation,

$$\begin{aligned} \log P\left(T_j | Story\right) = & \log P\left(T_j\right) + \\ & \sum_t f\left(\log\left(P\left(K_j\right)^b \cdot P\left(w_t | K_j\right) / P\left(w_t\right)\right)\right), \end{aligned} \tag{10}$$

where b is a weighting parameter, and f is a filter, $f(x)=x$ (if $x > \theta$) or θ (if $x \leq \theta$), which extracts positive information only in order to avoid the negative effects of words from other topics in the story. In spotting, the model trained by the EM algorithm responds to its own keywords and yields low scores for general words and words from other topics. The filter is used to take the higher scores for keywords and avoid the lower scores.

After the individual evaluation and sorting, to find the best topic set among the top-N topics, all possible 2^N-1 sets from the topics are evaluated by (9) (for small N, this is not expensive). Then the topic set that gives the best score is regarded as the retrieved topics from the story.

The probability of a topic set in (9) is approximated by the cooccurrence probabilities of each pair of topics in the training stories with smoothing by padding.

$$P(Set)=\prod_{l \in Set} \prod_{m \in Set\ (m>l)} P\left(T_l, T_m\right)^{1/{}_N C_2} \qquad (11)$$

5. EXPERIMENT

5.1. Transcriptions

A topic spotting experiment was performed by using manually transcribed text data with hand-labeled topics produced by Primary Source Media [1]. They are radio and TV news of CNN, ABC, NPR, and so on. The corpus had a total of 4.5 years of training data (from Jan. '92 to Jun. '96), covering 9,062 different topics. The test data consisted of 989 stories from the first half of July '96. The test stories have new topics which are out of trained topics (OOT), so the unlimited topics in the test are a problem in this task. Table 1 shows the number of unique topics and the percentage of OOT for 1 year and 4.5 years, if we limit the topics to those that occurred two or more times. In this paper we report an experiment trained from one year of data, because a preliminary experiment revealed better performance with one year of training (despite the higher OOT rate when using fewer topics) due to the older data being less relevant.

We used 42,502 training stories from July '95 to June '96. They had 4,627 different topics (plus a special topic of "general English") that appeared in more than one story to avoid over-training on topics with only one story. The number of topics in the test stories ranged from 1 to 13 with 4.5 on average.

Function words and pronouns like "the" or "he'll" which obviously don't relate to any topics were eliminated from our vocabulary according to 215 stop words modified from a list [10]. Suffixes like "ing" or "es" were also eliminated by a method like Porter's algorithm [11]. After all, the number of unique words in the training stories was 95,597. Each topic had about 3K unique words in the training stories. In the following experiments, correct story boundaries were given.

Table 1: Number of unique topics.

training term	min. story	#topic	OOT
4.5 year (92.1-96.6)	1	9,062	0.64%
4.5 year (92.1-96.6)	2	7,131	1.14%
1 year (95.7-96.6)	2	4,627	2.45%

5.2. Training

The topic models were trained using the EM algorithm. Initial distributions of the probabilities were set identically. After four iterations, the likelihood on the training stories converged, and we observed the predicted sharpening of the distributions. Tables 2a to 3b show example statistics from the conventional model and the proposed model. The conventional model estimates were obtained by counting stories or words for each topic and being smoothed [12]. Table 2a shows the prior probabilities that some topics will occur in a story. Table 2b shows, for the conventional method, the conditional probability for several words, given that one of the topics is "Clinton, Bill". Note that, while "president" and "Clinton" are likely, so are many general words like "go" and "think". Table 3a, for the new method, shows the probability that each topic will produce a keyword given a story that is related to the topic. Table 3b shows the probability of the words given that the topic "Clinton, Bill" produced a keyword. The relevant keywords are 8 to 10 times more likely, while the relatively general words like "go" or "think" have much lower probabilities than their word probabilities given by the conventional model. These statistics indicate greater discriminative ability of the proposed model between keywords and general words.

Table 2a: *a priori* probability.

	Topic	$P(T)$
1	Politics and government	0.145
2	Clinton, Bill	0.125
3	Administration	0.101
4	Presidents	0.092
5	Election	0.090

Table 2b: Topic-conditioned word prob.

	Word	$P(w\|T)$
1	president	0.013
2	go	0.011
3	think	0.010
4	Clinton	0.009
5	say	0.008

T="Clinton, Bill"

Table 3a: Probability of producing a keyword.

	Topic	$P(K_T)$
1	General English	0.935
2	Music, Black	0.085
:	:	:
476	Cinton, Bill	0.020
599	Politics and government	0.018
1,170	Administration	0.012

Table 3b: Word prob. given a topic keyword.

	Word	$P(w\|K_T)$
1	president	0.104
2	Clinton	0.096
3	house	0.036
4	white	0.034
:	:	:
36	go	0.003
44	think	0.003

T="Clinton , Bill"

5.3. Spotting Result

We tested both methods on 989 test stories. The values of the parameters in the proposed and conventional models were determined by preliminary experiments on training and spotting from 1995's transcriptions. We found that the parameters of θ=0.0 and b=0.35 in the proposed method and a=0.25 in the conventional method gave good performance in spotting. The filter f or keyword selection by chi-square test [2] was not used in the conventional method because they did not help the performance once the probabilities were smoothed.

Figure 2 shows the precision and the probability that at-least-one topic is correct when scoring topics independently. Precision is the fraction of correctly retrieved topics over all the retrieved topics. The proposed method achieved a precision of 0.757 for the first choice while the conventional method produced a precision of 0.636. We also evaluated the methods by recall (not shown). Recall is the fraction of correctly retrieved topics over all the labeled topics in the test stories. The proposed method yielded a recall of 0.507 for the top 5 choices while the conventional method yielded 0.468. At all top numbers of retrieved and sorted topics, the proposed method showed better results in precision, recall, and at-least-one accuracy than the conventional method.

The precision and recall generally vary inversely with each other. Besides how to find suitable topics, it is also difficult to find the best number to be retrieved. In order to decide how many topics should be answered from each story, after the individual topic evaluation, we rescored the top 5 topics (plus general English) by finding the best topic set among them using (9). By the rescoring, for example, a story about hurricanes could eliminate an unsuitable topic of "Discrimination in education" from the top 5. The final topic sets for the test stories resulted in different numbers of retrieved topics for each story. The precision was 0.758 for the first choice and 0.490 for the top 5 choices (if retrieved) that is 6.1% relatively higher than retrieving topics independently (Figure 2). The result suggests that the rescoring method can effectively eliminate some incorrect topics.

6. CONCLUDING REMARKS

A simple HMM was applied to retrieve detailed multiple topics from broadcast news stories. The EM algorithm improved the discriminative ability between keywords and general words in a topic model and decreased the probabilistic overlap among the topic models. The new method showed better results in precision, recall and at-least-one accuracy than the conventional method of a multinomial probability model.

In order to better understand the behavior of the system, we labeled topics on several stories ourselves. We merged the original labels with the top 10 answers from both methods. In general we felt that 6 to 8 of the topics were specifically relevant to the story, rather than the 4 labels. The original annotators had to make choices without examining the full list, while we had the benefit of seeing a short list of all possible relevant topics. In the majority of the cases where the system chose a topic that was not among the labels, we felt that the system was actually correct. Thus, the true precision of the top choice is probably much higher.

In this paper we described an experiment from manually transcribed texts. Experiments using transcriptions produced automatically by continuous speech recognition are also being performed, but we do not have enough speech data with hand-labeled topics to be reported here. The results from the transcriptions with errors will be reported in the near future.

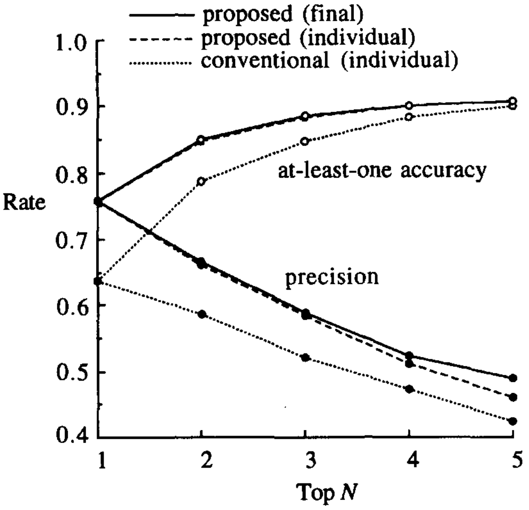


Figure 2: Topic spotting result.

REFERENCES

[1] http://www.thomson.com/psmedia/bnews.html

[2] L. Gillick, et. al., "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification Using Telephone Speech," Proc. ICASSP-93, Vol. II, pp. 471-474, 1993.

[3] B. Peskin, et. al., "Improvements in Switchboard Recognition and Topic Identification," Proc. ICASSP-96, Vol. II, pp. 303-306, 1996.

[4] J. McDonough, et. al., "Issues in Topic Identification on the Switchboard Corpus," Proc. ICSLP-94, pp. 2163-2166, 1994.

[5] J.H. Wright, et. al., "Improved Topic Spotting through Statistical Modelling of Keyword Dependencies," Proc. ICASSP-95, pp. 313-316, 1995.

[6] R.C. Rose, et. al., "Techniques for Information Retrieval from Voice Messages," Proc. ICASSP-91, pp. 317-320, 1991.

[7] Y. Yamashita, et. al., "Next Utterance Prediction Based on Two Kinds of Dialog Models," Proc. Eurospeech-93, pp. 1161-1164, 1993.

[8] X.D. Huang, et. al., *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[9] L.E. Baum, et. al., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," Ann. Math. Stat., vol. 41, pp. 164-171, 1970.

[10] http://www.perseus.tufts.edu/Texts/engstop.html

[11] M.F. Porter, "An Algorithm for Suffix Stripping," Program, 14(3), pp. 130-137, 1980.

[12] P. Placeway, et. al., "The Estimation of Powerful Language Models from Small and Large Corpora," Proc. ICASSP-93, Vol. II, pp. 33-36, 1993.