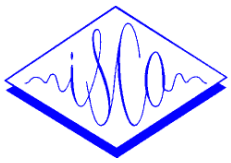

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ISSUES IN TOPIC IDENTIFICATION ON THE SWITCHBOARD CORPUS

John McDonough *Herbert Gish*
BBN Systems and Technologies
70 Fawcett Street 15/1c, Cambridge MA 02138 USA

ABSTRACT

Topic identification (TID) is the automatic classification of speech messages into one of a known set of possible topics. The TID task can be view as having three principal components: 1) event generation, 2) keyword event selection, and 3) topic modeling. Using data from the Switchboard corpus, we present experimental results for various approaches to the TID problem and compare the relative effectiveness of each. In particular, we examine issues in topic modeling and keyword selection.

1. INTRODUCTION

Topic identification (TID) is a classification problem where the task is the assignment of the correct topic label to a speech message known to be about one of a fixed number of possible topics. This classification is performed using only features extracted from the speech message itself. Several approaches to TID have been investigated and described previously [7, 11, 13]. In [7] a paradigm based on a large vocabulary word recognizer combined with unigram topic models was investigated. In [11] a speech recognizer was used as one component of a larger system that sorted air traffic control transmissions, and in [13] a word spotter was used to generate events for a message classifier.

In past work [2], we have presented the identification problem as being composed of three main sub-tasks—i.e., event detection, keyword selection, and topic modeling—and have investigated each in turn. While we have settled, at least for the present, the question of event detection, the other tasks remain open issues, and hence are the subject of our current investigations. In particular, we shall propose a new topic modeling scheme based on mixtures of multinomials. We will use this new model for both topic classification and as a component of a keyword selection paradigm. In an attempt to capture dependencies among keyword occurrences, we will also investigate the use of discriminative topic model training for the single multinomial case. We shall propose two alternate keyword selection criteria, one each based on the well-known Kulback-Liebler distance and the Mutual Information measure, and derive keyword sets of various sizes for these criteria as well as for the χ^2 hypothesis test considered previously [2]. Finally, we will compare the performance obtained using the several methods of topic modeling and keyword selection on a topic identification task derived from the Switchboard Corpus.

2. THEORETICAL DEVELOPMENT

In this section, we present the theoretical development and topic-spotting paradigms relevant for the remainder of the paper. We first describe the mixture multinomial model which plays in important role in keyword selection and topic classification. We then propose and discuss two new keyword selection metrics. Finally, we provide some motivation for an alternate topic model training scheme intended to account for keyword dependencies.

2.1. The Mixture Multinomial Model

Initially consider a sequence of words $\mathbf{w}$—hereafter to be referred to as a *conversation*—whose elements are drawn from some finite vocabulary $\mathbf{V} = \{v_l\}_{l=0}^{N_v}$ where $N_v = \| \mathbf{V} \|$ is the vocabulary size. To provide for the possibility of a non-exhaustive keyword set, we specify that v_0 is to represent the out-of-vocabulary (OOV) keyword; ie, it corresponds to "none of the above." Upon assuming that each of the elements of $\mathbf{w}$ is independent and identically distributed (iid), we immediately arrive at the *multinomial* model:

$$p(\mathbf{w}|\vec{\theta}) = \prod_{l=0}^{N_v} \theta_l^{n_l} \tag{1}$$

where $n_l = n_l(\mathbf{w})$ is the number of times keyword v_l occurred in sequence $\mathbf{w}$, and the parameter vector $\vec{\theta} = \{\theta_l\}_{l=0}^{N_v}$ represents the set of word occurrence probabilities corresponding to keyword vocabulary $\mathbf{V}$. Although useful for topic identification as written, the model of Eqn. (1) can be generalized to a *mixture* of multinomials according to

$$p(\mathbf{w}|\Lambda) = \sum_{k=1}^{K} q_k p(\mathbf{w}|\vec{\theta}_k) \tag{2}$$

where $\Lambda = \{q_k, \vec{\theta}_k\}_{k=1}^{K}$ is the complete set of model parameters, including both mixture-dependent word occurrence probabilities and *a priori* mixture weights.

Application of either of Eqns. (1) or (2) to topic identification is easily accomplished by stipulating that any relevant model parameters be topic-dependent, and then estimating these parameters from some training set. In the case of Eqn. (1), the word occurrence parameters $\vec{\theta}$ can be obtained as maximum likelihood frequency estimates. The parameters appearing in Eqn. (2) can be estimated by the Estimation-Maximization (EM) Algorithm [1]. As the development of the appropriate re-estimation formulae is straightforward, we omit the details and state the final result as follows: For some set $\mathcal{X} = \{\mathbf{w}^{(i)}\}_{i=1}^{N}$ of N training sequences, the E-step consists of calculating the posterior probabilities

$$\hat{z}_k^{(i)} = \frac{q_k p(\mathbf{w}^{(i)}|\vec{\theta}_k)}{\sum_{j=1}^{K} q_j p(\mathbf{w}^{(i)}|\vec{\theta}_j)} \tag{3}$$

where the sequence likelihoods $p(\mathbf{w}^{(i)}|\vec{\theta}_k)$ are as in Eqn. (1). The M-step, in turn, requires

$$\hat{q}_k = \frac{1}{N} \sum_{i=1}^{N} z_k^{(i)} \tag{4}$$

$$\hat{\theta}_{kl} = \frac{\sum_{i=1}^{N} z_k^{(i)} n_l^{(i)}}{\sum_{i=1}^{N} z_k^{(i)} M^{(i)}} \quad (5)$$

where $M^{(i)} = \sum_{l=0}^{N_v} n_l^{(i)}$ is the length of training sequence $\mathbf{w}^{(i)}$. Implicit in the above is that the parameter and posterior probability estimates are updated in an iterative manner, as consistent with the EM paradigm.

2.2. Keyword Selection

In previously published work, we have considered keyword selection based on contingency table analysis [2]. This approach can be viewed as testing the null hypothesis that two discrete empirical densities are *equivalent*, where the empirical densities are histogram estimates (see, for example, Duda and Hart [3]) of word distributions for two or more different topics. While it is not our primary concern here, this approach will be retained for the sake of comparison.

By way of developing our new keyword selection methods, consider again the sequence likelihood model of Eqn. (2), to which we wish to make the following modifications:

1. We consider only a single keyword at a time; hence the individual mixture parameters $\vec{\theta}_k$ consist of only two probabilities—one corresponding to the occurrence of a prospective keyword and one to the occurrence of the OOV word.
2. As previously alluded to, we specify that the parameters Λ are topic-dependent and to be determined from some labelled training data.

Using the *binomial* mixtures obtained based on these assumptions, it is possible to calculate statistics indicating differences in keyword occurrence patterns across various topics, and—by extension—the utility of any given keyword in identifying a desired topic. Two such statistics we will consider in some detail are the Kullback-Liebler (KL) distance and the Mutual Information measure.

Kulback-Liebler Distance

The *symmetrized* KL distance (see, for example, Cover [4]) is defined only for the two-topic case and can be expressed as

$$D(T \parallel \tilde{T}) = \int_{\text{all } \mathbf{w}} [p(\mathbf{w}|T) - p(\mathbf{w}|\tilde{T})] \log \frac{p(\mathbf{w}|T)}{p(\mathbf{w}|\tilde{T})} \, d\mathbf{w} \quad (6)$$

where, in keeping with our second rule, we implicitly assume that $p(\mathbf{w}|T) = p(\mathbf{w}|\Lambda(T))$. Although conceptually simple, the KL distance as stated can be computationally intractable for even moderately complex density functions. Thus, we are led to consider an empirical KL distance derived from some labelled training set $\mathcal{X} = \{(\mathbf{w}^{(i)}, c^{(i)})\}$, where $c^{(i)} \in \{T, \tilde{T}\}$, given by

$$D(T \parallel \tilde{T}) \approx \frac{1}{N(T)} \sum_{\mathbf{w} \in T} \log \frac{p(\mathbf{w}|T)}{p(\mathbf{w}|\tilde{T})} - \frac{1}{N(\tilde{T})} \sum_{\mathbf{w} \in \tilde{T}} \log \frac{p(\mathbf{w}|T)}{p(\mathbf{w}|\tilde{T})} \quad (7)$$

where $N(T)$ and $N(\tilde{T})$ are the numbers of training samples from topics T and $\tilde{T}$, respectively.

Mutual Information Measure

The ensemble mutual information (MI) measure [4] is given by

$$I(W;T) = \sum_{\text{all}\mathbf{W},T} p(\mathbf{w},T) \log \frac{p(\mathbf{w}|T)}{p(\mathbf{w})} \quad (8)$$

where $p(\mathbf{w}) = \int_{\text{all } \mathbf{W}} p(\mathbf{w}|T)p(T)d\mathbf{w}$ is the complete likelihood of sequence $\mathbf{w}$. It too admits an empirical version given by

$$I(W;T) = \frac{1}{N} \sum_{\mathbf{w}^{(i)} \in \mathcal{X}} \frac{p(\mathbf{w}^{(i)}|c^{(i)})}{p(\mathbf{w}^{(i)})} \quad (9)$$

We note that the MI measure can be defined for a multiplicity of topics. In the results to be presented subsequently, however, we will find it expedient to restrict ourselves to the two-topic case, as with the KL distance measure. In particular, when selecting keywords using either the KL distance or MI measure, we designate a given topic as *wanted* and the remaining topics as *unwanted*. We then train one mixture model for the wanted topic and one mixture model for *all* unwanted topics—the relevant keyword statistic is calculated for each keyword singly, as if there were only two topics. A subset of keywords pertaining to a particular topic is chosen by ranking all words considered according to the score obtained from a given metric, then taking the top-scoring words. By repeating these steps once for each topic of interest and taking the union of all the equally-sized keyword subsets thereby obtained, the final keyword set is chosen. This is done to ensure that a small number of topics with very high scoring keywords are not disproportionately represented in the final keyword set.

2.3. Linear Classifier Training

To revisit Eqn. (1), the closed-set topic classification problem under the simple iid assumption[1] involves selecting that topic T^* which, for sequence $\mathbf{w}$, maximizes the log-likelihood $\sum_{l=0}^{N_v} n_l \log \theta_l(T^*)$. Hence, if we were to consider only the two-topic case—designating the topics by T and $\tilde{T}$—optimal classification could be achieved through a threshold test on the log-likelihood ratio

$$\gamma(\mathbf{w}) = \sum_{l=0}^{N_v} n_l \log \frac{\theta_l(T)}{\theta_l(\tilde{T})} \quad (10)$$

From (10), it is apparent that the two-topic classifier is partitioning the feature space into decision regions separated by a hyperplane; i.e., that it is a *linear* classifier. In particular, it is a linear classifier in which the parameters are chosen without regard for the dependencies that may exist among the frequencies of occurrence of the keywords. For the multi-class problem, wherein a single multinomial is trained for each topic, similar arguments reveal the optimal classifier partitions the feature space into convex, disjoint decision regions with hyper-planar boundaries, whose parameters also do not take into account keyword occurrence dependencies. In consideration of the above, we speculated that the inadequacies of the iid assumption in the present application (ie, those stemming from the fact that conversational speech is decidedly *not* a source of iid symbols or words) might be partially overcome through the use of discriminative topic model training. To investigate this hypothesis, we chose to train a one-layer neural net as a topic discriminator, and considered only the two-class case in which a given topic was to be compared to all others in a finite set. This choice was made both for its ease of implementation and for its retaining the character of the linear classifier. The resulting network took as input features occurrence frequencies for a vocabulary of previously chosen keywords; it was trained to maximize the likelihood of the training set of conversations. Such models are known by various names in the literature, including *logistic discrimination models* which are members of the family of *generalized linear models* [14]. Such models are able to adjust its parameters to take into account keyword occurrence dependencies.

[1] We are actually also assuming that all topics are equally likely.

3. EVENT DETECTION

The goal in event detection is to extract relevant features or events from the speech message that can be used for topic modeling and identification. The events we consider in this study are the number of occurrences for each member of a set of keywords; each correct keyword detection provides information about the topic under discussion in the speech message.

In the past we have examined two methods of event detection: word recognition and word spotting. Recognition attempts to transcribe all words spoken in the speech message and requires the use of a vocabulary and language model sufficiently large and general to model all the topics. Word spotting presupposes that only a certain subset of words is relevant to identifying the topics, and attempts to find only these keywords in the speech message. As topic identification systems based on a word spotter front end have demonstrated performance superior to recognizer-based systems in prior research [2], we consider only word spotter event detection in the present work.

For the word spotting approach, we obtain an *expected number of occurrences* for each keyword by summing up the score associated with each putative keyword event detected by our HMM word spotter. In [10], we introduced a posterior probability scoring algorithm for HMM-based keyword spotting and extended it in [12] and [9]. Briefly, we compute the posterior probability $p(w,t)$ that keyword w has ended at time t given the observations from time 1 to T according to:

$$p(w,t) = p(s_t = e_w | O_1, ..., O_T) = \frac{\alpha(e_w,t)\beta(e_w,t)}{\sum_{\text{all } s} \alpha(s,t)\beta(s,t)} \quad (11)$$

where e_w is the ending state of keyword w, and $\alpha(s,t)$ and $\beta(s,t)$ are the forward and backward scores, respectively, for state s at time t as defined in the Baum-Welch algorithm [5]. A putative hit for keyword w is declared whenever $p(w,t)$ reaches a local maximum and its score is the value of $p(w,t)$ at the peak. To estimate the total number of occurrences n_w of keyword w in a speech message, we sum the scores of all putative events $p(w,t)$ in the message: $n_w \approx \sum_t p(w,t)$. These expected number of occurrences, determined by running the word spotter on actual Switchboard conversations, are used throughout this work for the purposes topic modeling, keyword selection, and classification performance testing.

4. EXPERIMENTS

The results reported here pertain to experiments performed on the same development test set used in prior topic identification work. In brief, we considered the closed-set topic identification problem for ten topics drawn from the Switchboard telephone corpus [8]. For each topic, the available conversation sides are partitioned into a training set and a development-test set. The training set is speaker disjoint with respect to the development-test sets. Each conversation side is half of a full-duplex recording and consists of speech from one speaker, along with possible cross-talk from the other. In all our experiments, each conversation side is independently scored for topic content. A break-down of the two data sets by topic is shown in Table 1.

The training set consists of 311 conversation sides and is used to train the acoustic and language models for the word spotter. The development-test set, which consists of 507 sides, is used to build the topic models, select the keywords, and tune the topic identification system.

4.1. Experimental Paradigm

The testing paradigm used in our developmental research involved a 10-fold cross-validation (CV) procedure to ensure unbiased results. It consisted of the following steps:

		# of Sides	
Topic	Description	Train	Dev-Test
302	Air Pollution	16	36
308	Music	31	51
312	Crime	30	53
314	Gun Control	30	38
320	Buying a Car	45	59
327	Public Service	27	27
351	Pets	24	69
353	Public Education	35	61
356	Exercise & Fitness	39	55
358	Family Life	34	58
	TOTAL	311	507

Table 1. Number of training and development-test conversation sides for the ten selected topics.

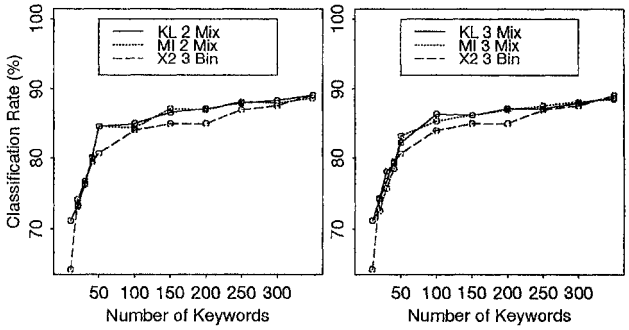

Figure 1. Topic identification performance for various types of keyword selection criteria, including symmetrized KL distance, MI measure, and hypothesis testing. In all cases, the topic classifier is based on a single multinomial model per topic.

1. The initial multinomial mixtures to be employed for keyword selection were trained using nine-tenths of the data in the Development Test set. Using the same subset of the development data, keywords were chosen based on one of the criteria discussed previously.
2. Based on the chosen keyword set, topic models with either single or multiple mixtures per topic were trained from the same nine-tenths of the data used in keyword selection.
3. The remaining tenth of the conversations, the held out set, are scored against each of the topic models and the percent correctly identified are tallied.

These steps are repeated ten times, once for each CV-fold, such that all conversations are tested exactly once.

4.2. Experimental Results

In the first series of experiments, we wished to gauge the effectiveness of the newly proposed methods of keyword selection. As such, the final topic classification scheme, involving a single multinomial for each topic, was fixed and the various selection methods were compared. Keyword sets of various sizes N_v were obtained by ranking prospective keywords based on one of the three metrics—KL distance, MI measure, or χ^2-significance score—then choosing those $N_v/10$ keywords for each *topic* with the maximum score. These words were subsequently used to build a topic classifier. The results of this experiment are shown in Figure 1. As apparent from Figure 1, both the KL distance and MI measure showed a slight improvement in classification accuracy over the previous contingency table test, especially for smaller keyword set sizes.

The second set of experiments examined the effect of using multiple mixtures per topic in the classification phase of the test-

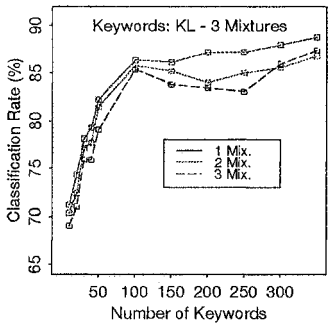

Figure 2. Performance comparison of single vs. multiple mixture topic models. Keyword set—obtained through three mixture TAO KL distance selection criterion—is identical in all three cases.

Topic	Description	MN	GLM
302	Air Pollution	0.9487	0.9625
308	Music	0.9763	0.9724
312	Crime	0.8915	0.9191
314	Gun Control	0.9625	0.9645
320	Buying a Car	0.9566	0.9625
327	Public Service	0.9507	0.9586
351	Pets	0.9527	0.9310
353	Public Education	0.9586	0.9487
356	Exercise & Fitness	0.9546	0.9487
358	Family Life	0.8738	0.9112
Avg.		0.9426	0.9479

Table 2. Comparison of discriminatively-trained generalized linear model (GLM) versus standard multinomial (MN) in the topic-against-other classification problem.

ing procedure. Hence, we fixed the keyword selection method—ie, the KL distance criterion calculated from three-multinomial mixtures—and then determined the performance of various numbers of mixtures in the final classifier. The results are shown in Figure 2. The results in this case indicate that the single multinomial per topic classifier is superior to both the two and three multinomial mixtures for keyword sets of all sizes tested. This is thought to be related to the paucity of training data; the more detailed models have more parameters that must be estimated and hence require more training if the final estimates are to be robust.

Linear Model Training

The results of our experiments concerning logistic discrimination model training are summarized in Table 2. For most topics both the mutlinomial generalized linear models give very comparable performance. For a few topics, however, including Crime (312) and Family Life (358), the balance tilts in favor of the discriminatively trained model. This is thought to be due to dependency among keywords that are captured by the logistic discrimination model and not by the multinomial. We also note that each keyword, due to false alarms, appears to have been spotted a number of times even when it hasn't occurred in the speech. The occurrence of such false alarms tends to decrease the dependencies in the observed word counts.

REFERENCES

[1] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, 39, pp. 1–38, 1977.

[2] J.W. McDonough, K. Ng, *et al*, "Approaches to topic identification on the switchboard corpus," *Proc. ICASSP*, Volume I, pp. 385–388, Sydney, May 1994.

[3] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.

[4] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[5] L.E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains" in *Ann. Math. Stat.*, 1966, vol. 37, pp. 1554-1563.

[6] F. Kubala, Y. Chow, A. Derr, et. al., "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-word Resource management Database." in *IEEE ICASSP*, 1988, pp. 291-294.

[7] L. Gillick, J. Baker, et. al., "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification using Telephone Speech" in *IEEE ICASSP*, 1993, Volume II, pp. 471-474.

[8] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research Development" in *IEEE ICASSP*, 1992, Volume I, pp. 517-520.

[9] P. Jeanrenaud, M. Siu, K. Ng, R. Rohlicek, H. Gish, "Phonetic-based Word Spotter: Various Configurations and Application to Event Spotting." in *ESCA Eurospeech*, 1993, Volume II, pp. 1057-1060.

[10] J.R. Rohlicek, W. Russell, S. Roukos and H. Gish, "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting." in *IEEE ICASSP*, 1989, pp. 627-630.

[11] J.R. Rohlicek, D. Ayuso, et. al., "Gisting Conversational Speech" in *IEEE ICASSP*, 1992, Volume II, pp. 113-116.

[12] J.R. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, et. al, "Phonetic Training and Language Modeling for Word Spotting" *IEEE ICASSP*, 1993, Volume II, pp. 459-462.

[13] R.C. Rose, E.I. Chang, R.P. Lippmann, "Techniques for Information Retrieval from Voice Messages" in *IEEE ICASSP*, 1991, Volume I, pp. 317-320.

[14] T.J. Hastie and D. Pregibon, *Statistical Models in S*, Wadsworth & Brooks, Pacific Grove CA, 1992.

[15] K. Farrell, R.J. Mammone, and A.L. Gorin, "Adaptive language acquisition using incremental learning," in *Proc. ICASSP*, 1993, pp. Volume I, pp. 501-504.