# DISCRIMINATIVE FEATURE WEIGHTING USING MCE TRAINING FOR TOPIC IDENTIFICATION OF SPOKEN AUDIO RECORDINGS

*Timothy J. Hazen and Anna Margolis*

MIT Lincoln Laboratory
Lexington, Massachusetts, USA

**ABSTRACT**

In this paper we investigate a discriminative approach to feature weighting for topic identification using minimum classification error (MCE) training. Our approach learns feature weights by optimizing an objective loss function directly related to the classification error rate of the topic identification system. Topic identification experiments are performed on spoken conversations from the Fisher corpus. Features drawn from both word and phone lattices generated via automatic speech recognition are investigated. Under various different conditions, our new feature weighting scheme reduces our classification error rate between 9% and 23% relative to our baseline naive Bayes system using feature selection.

***Index Terms***— Audio document processing, topic identification, topic spotting.

## 1. INTRODUCTION

In this paper we investigate a discriminative approach to feature weighting for topic identification (or *topic ID*) using minimum classification error (MCE) training. This work extends our previous work in topic ID of audio files where we explored the use of features extracted from speech recognition hypothesis lattices created from both word-based and phone-based recognizers [1]. The primary goal of our previous work was to examine the application of existing approaches to topic ID to audio files of spoken human-human conversations. A particular focus of our work is topic ID using only the output of phonetic recognition systems (and potentially phonetic recognition systems from mismatched languages). An important aspect of this type of topic ID is the automatic discovery of phonetic features (e.g., phonetic n-gram sequences) which are relevant for the prediction of the topic.

Past research has largely viewed the topic ID problem as consisting of two primary stages: *feature selection* and *classification*. In feature selection, the goal is to reduce the large space of potential features to a smaller set which possesses the most relevant or discriminative features for topic ID. For example, in word-based systems this typically involves discovery of the content words which are most likely to be used during the discussion of a particular topic. Previous feature selection techniques have utilized statistical measures which capture correlations between features and topics, e.g. the mutual information between features and topics, the maximum a posteriori probability of topics given features, or $\chi^2$ statistics [1, 2].

Given a set of features, the second stage of topic ID is classification. The use of naive Bayes classifiers is popular throughout much of the topic ID research. Because these classifiers use generative models, their training can be performed efficiently, their parameters can be learned and adapted in an on-line fashion, and their accuracy is often sufficient for many tasks [3, 4]. There are two obvious potential drawbacks to the standard naive Bayes approach. First, because naive Bayes systems are based on generative models, their parameters are generally estimated statistically instead of being trained in a discriminative fashion. Second, the processes of feature selection and model training are generally performed independently instead of jointly.

In this work, we attempt to address the shortcomings of the traditional naive Bayes classifier by applying a discriminative procedure commonly called minimum classification error (MCE) training [5, 6] to the topic ID problem. In this paper, we describe the application of MCE training to feature weights in a naive Bayes topic ID system. We will present experimental evidence detailing the usefulness of our technique when applied to data from the Fisher Corpus of conversational human-human telephone speech. Results based on both word-based and phone-based speech recognition of the audio data are provided.

## 2. EXPERIMENTAL TASK DESCRIPTION

### 2.1. Corpus

For the data set for our experiments we have used the English Phase 1 portion of the Fisher Corpus [7, 8]. This corpus consists of 5851 recorded telephone conversations. During data collection, two people were connected over the telephone network and given instructions to discuss a specific topic for 10 minutes. Data was collected from a set of 40 different topics. The topics were varied and included relatively distinct topics (e.g. "Movies", "Hobbies", "Education", etc.) as well as topics covering similar subject areas (e.g. "Issues in Middle East", "Arms Inspections in Iraq", "Foreign Relations"). Fixed prompts designed to elicit discussion on the topics were played to the participants at the start of each call. For our experiments the corpus was subdivided into four subsets:

1. Recognizer training set (3104 calls; 553 hours)
2. Topic ID training set (1375 calls 244 hours)
3. Topic ID development test set (686 calls; 112 hrs)
4. Topic ID evaluation test set (686 calls; 114 hrs)

### 2.2. Speech Recognition Systems

#### *2.2.1. Overview*

In our topic ID experiments, the first stage of processing is to apply automatic speech recognition (ASR) to each segment of audio in each audio document. The ASR system is used to generate a network, or *lattice*, of speech recognition hypotheses for each audio

This work was sponsored by the Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

segment. While topic ID systems generally use word-based speech recognition [9, 10], for some applications knowledge of the important lexical keywords for a task may be incomplete or entirely unavailable. Under these situations, it is possible to instead perform topic ID based on phonetic strings (or phonetic lattices) while retaining the same topic ID framework [11, 12].

In this work we explore the use of both word-based and phone-based speech recognition. Within each lattice we can compute the posterior probability of any hypothesized word (or phone sequence), and an *expected count* for each word can be computed by summing the posterior scores over all instances of that word over all lattices from both sides of the conversation within each call.

#### *2.2.2. Word-Based Speech Recognition*

For word-based ASR we have used the MIT SUMMIT speech recognition system [13]. The system's acoustic models were trained using a standard maximum-likelihood approach on the full 553 hour recognition training set specified above without any form of speaker normalization or adaptation. For language modeling, the system uses a basic trigram language model with a 31.5K word vocabulary trained using the transcripts of the recognizer training set. Because this recognizer applies very basic modeling techniques with no adaptation, the system performs recognition faster than real time (on a current workstation) but word error rates can be high (typically over 40%).

#### *2.2.3. Phone-Based Speech Recognition*

For phonetic recognition we use a phonetic ASR system developed at the Brno University of Technology (BUT) [14]. Two versions of the system were trained, one which uses an English phone set and one which uses a Hungarian phone set. The English recognizer was trained using 10 hours from the Switchboard Cellular Phase 1 conversational telephone speech corpus [8]. The Hungarian recognizer was trained using the Hungarian portion of the SPEECH-DAT corpus [15]. This corpus contains read speech collected over the Hungarian telephone network.

## 3. PROBABILISTIC TOPIC IDENTIFICATION

### 3.1. The Naive Bayes Formulation

In a probabilistic approach to topic identification, the goal is to determine the likelihood of a document being of topic $t$ (from a set of topics $T$) given the document's string of words $W$. In audio-based topic ID the true string of spoken words is not known and must be determined automatically. In this case the variable $W$ represents a set of words (or word-like features such as phonetic n-grams) that are extracted from an audio document via ASR.

For topic ID scoring we use a hypothesis testing likelihood ratio approach. For closed-set topic ID, an audio document (as represented by $W$) will be determined to belong to topic $t_i$ if the following expression holds:

$$\forall j \neq i \quad \frac{P(W|t_i)}{P(W|\bar{t}_i)} > \frac{P(W|t_j)}{P(W|\bar{t}_j)} \tag{1}$$

Here $P(W|t)$ represents the likelihood that $W$ is produced given the topic is $t$, while $P(W|\bar{t})$ represents the likelihood that $W$ is produced given the topic is not $t$. The same scoring approach could also be used for open set topic detection by comparing this likelihood ratio against a pre-determined score threshold.

When modeling $P(W|t)$, if we assume $W$ is a word string, we can expand it into its underlying sequence of $N$ words, $w_1, \ldots, w_N$. In the naive Bayes approach to the problem, statistical independence is assumed between each of the individual words in W. Under this assumption, the likelihood of $W$ given $t$ is approximated as:

$$P(W|t) \approx \prod_{i=1}^{N} P(w_i|t) \tag{2}$$

The expression above assumes a sequence of $N$ individual words. When using expected counts from lattices, this expression can alternatively be represented with a counting interpretation as follows:

$$P(W|t) \approx \prod_{\forall w \in V} P(w|t)^{C_{w|W}} \tag{3}$$

In this interpretation, the occurrence count $C_{w|W}$ within $W$ of each word $w$ in the system's vocabulary, $V$, is used to exponentially scale the score contribution of that word. Under this interpretation non-integer values of the counts $C_{w|W}$ are allowed, thus providing the system the ability to incorporate expected count estimates from lattices generated by a recognition system.

The likelihood function $P(W|\bar{t})$ is generated as follows:

$$P(W|\bar{t}) = \frac{1}{N_T - 1} \sum_{\forall t_i \neq t} P(W|t_i) \tag{4}$$

Here $N_T$ is the total number of known topics. This expression assumes a uniform prior distribution over all topics.

In practice the score for topic $t$ given words $W$, expressed as $\mathcal{F}(t|W)$, is implemented in the log domain as the following sum:

$$\mathcal{F}(t|W) = \sum_{\forall w \in V} C_{w|W} \log \frac{P(w|t)}{P(w|\bar{t})} \tag{5}$$

This naive Bayes approach is utilized as our baseline system.

### 3.2. Parameter Estimation

The likelihood function $P(w|t)$ in our system is estimated from training materials using maximum *a posteriori* probability (MAP) estimation with Laplace smoothing as follows:

$$P(w|t) = \frac{N_{w|t} + N_V P(w)}{N_{W|t} + N_V} \tag{6}$$

In this expression, $N_V$ is the total number of words in the vocabulary used by the system, $N_{w|t}$ is the number of times word $w$ occurs in training documents of topic $t$, and $N_{W|t}$ is the total number of words in the training documents of topic $t$. The term $P(w)$ represents the prior likelihood of word $w$ occurring independent of the topic. This likelihood function is also determined using MAP estimation with Laplace smoothing as follows:

$$P(w) = \frac{N_w + 1}{N_W + N_V} \tag{7}$$

In this expression, $N_w$ is the number of occurrences of the specific word $w$ in the training corpus and $N_W$ is the total count of all words from the $N_V$ word vocabulary in the training corpus.

### 3.3. Feature Selection

As discussed earlier, it is typically the case that a small number of topic specific features contribute heavily to the determination of the topic (e.g., content words), while many other features (e.g., function words) contribute nothing to the decision. For this reason, topic identification systems typically employ a feature selection process in which only a subset of the possible features are actually used.

In our previous work we examined multiple measures for feature selection including the information gain measure and the $\chi^2$ statistic. We had the most success using the maximum topic posterior probability measure [1]. In this approach, we select the top $N$ words per topic which maximize the posterior probability of the topic, i.e. the words which maximize the value of $P(t|w)$, where $P(t|w)$ is determined using MAP estimation as follows:

$$P(t|w) = \frac{N_{w|t} + 1}{N_w + N_T} \tag{8}$$

We use this feature selection method in the experiments in this paper.

### 3.4. MCE-Based Feature Weighting

Feature selection can be viewed as a specific case of feature weighting, where each feature receives either a weight of one or a weight of zero. In the more general case, we can allow the weights of each feature to be of any value (or at least any positive value). To express this, let us first use the following simplifying notational substitution:

$$f(t|w) = \log \frac{P(w|t)}{P(w|\bar{t})} \tag{9}$$

The basic naive Bayes expression in Equation 5 can now be generalized to include variable valued features weights as follows:

$$\mathcal{F}(t|W) = \sum_{\forall w \in V} \lambda_w C_{w|W} f(t|w) \tag{10}$$

Here, $\lambda_w$ is the feature weight associated with word $w$.

Our goal is to learn values for the collection of feature weights which minimize the topic ID error rate. We utilize the discriminative minimize classification error (MCE) training approach to this problem [5, 6]. In this approach, a misclassification measure is first defined as:

$$\mathcal{M}(W) = \mathcal{F}(t_I|W) - \mathcal{F}(t_C|W) \tag{11}$$

Here, $t_C$ represents the correct topic for the audio document and $t_I$ represents the best scoring incorrect topic. If the document is correctly classified the misclassification measure $\mathcal{M}(W)$ will be negative, while incorrect classification will result in a positive value.

The misclassification measure is then mapped by a sigmoid loss function onto the $[0, 1]$ continuum as follows:

$$\ell(W) = \frac{1}{1 + e^{-\beta \mathcal{M}(W)}} \tag{12}$$

Here, $\beta$ represents the slope of the sigmoid function. The loss function will be close to zero for documents with large negative values of $\mathcal{M}(W)$ and close to one for documents with large positive values of $\mathcal{M}(W)$. The average value of the loss function over all documents approximates the actual topic ID error rate (becoming exact as $\beta \rightarrow \infty$). As such, minimizing the average value of the loss function should also minimize the classification error rate. Because the loss function is a smooth monotonic function over $\mathcal{M}(W)$, it can be differentiated with respect to the individual features weights and optimized via an iterative gradient descent algorithm.

The partial derivative of the loss function $\ell(W)$ with respect to a specific feature weight $\lambda_w$ is:

$$\frac{\partial \ell(W)}{\partial \lambda_w} = \beta \ell(W) \left(1 - \ell(W)\right) \left(f(t_I|w) - f(t_C|w)\right) C_{w|W} \tag{13}$$

The partial derivatives over all $W$ in the training data can be averaged and an update of the feature weights can be expressed as:

$$\lambda'_w = \lambda_w - \epsilon \frac{1}{N_D} \sum_{\forall W} \frac{\partial \ell(W)}{\partial \lambda_w} \tag{14}$$

Here $N_D$ is the total number of training documents and $\epsilon$ is a learning rate parameter used during the iterative MCE training. In our experiments we do not allow $\lambda_w$ to go negative, but we do not constrain or normalize the values of each $\lambda_w$ in any other way.

Ideally, when computing the partial derivatives of $\ell(W)$, the log likelihood ratios used in $\mathcal{F}(t|W)$ should be estimated from data which excludes $W$. This can be accomplished via a *jack-knifing* training process over the training data. In our experiments, the topic ID training data was subdivided into ten partitions, and for each training document $W$ the functions $f(t|w)$ were estimated from the nine partitions of the training data that did not include $W$.

### 3.5. Relationship to Previous Work

The MCE approach in this paper is similar in nature to prior work in discriminative training for automatic call routing tasks [16, 17]. The basic naive Bayes formulation can be mapped into the vector space call routing formulation presented by Kuo and Lee [16]. Specifically, the log likelihood ratios $f(w|t)$ can be viewed as entries in an $N_T$ by $N_V$ matrix which is referred to as a routing matrix $\mathbf{R}$. Viewed in matrix form, a vector of topic scores $\vec{s}$ is generated from the product of the routing matrix $\mathbf{R}$ and the vector of features counts $\vec{c}$ as follows:

$$\vec{s} = \mathbf{R}\vec{c} \tag{15}$$

In their work, discriminative training of all terms of the routing matrix $\mathbf{R}$ is performed. In our work, the set of feature weights can be viewed as a diagonal $N_V$ by $N_V$ matrix $\mathbf{\Lambda}$ whose diagonal entries are the feature weights $\lambda_w$. When adding $\mathbf{\Lambda}$ into the formulation, the topic ID vector is produced from this expression:

$$\vec{s} = \mathbf{R}\mathbf{\Lambda}\vec{c} \tag{16}$$

Viewed in this matrix formulation, our MCE approach trains the diagonal entries in $\mathbf{\Lambda}$ instead of the full matrix $\mathbf{R}$. Because $\mathbf{\Lambda}$ contains far fewer trainable parameters than $\mathbf{R}$, it is less susceptible to overfitting by the discriminative training process.

## 4. EXPERIMENTAL RESULTS

For our experiments, we report results on 40-topic closed-set topic ID (though our hypothesis testing formulation is equally valid for topic detection tasks as well). Figure 1 demonstrates the effect of training the feature weights with up to 125,000 MCE iterations for our word-based system. Results are shown for both the topic ID error rate and the average value of the loss function for both the training set and the held-out development test set. The experiment used all 30,373 words from recognizer's vocabulary observed in the lattices of the topic ID training set. The feature weights were initialized to 1.0 for all words. The training used values of 0.1 for $\beta$ and 1.0 for $\epsilon$.

In Figure 1, the topic ID error rate on the training set begins at 14.9% (with an average loss function value of 0.167) and, after

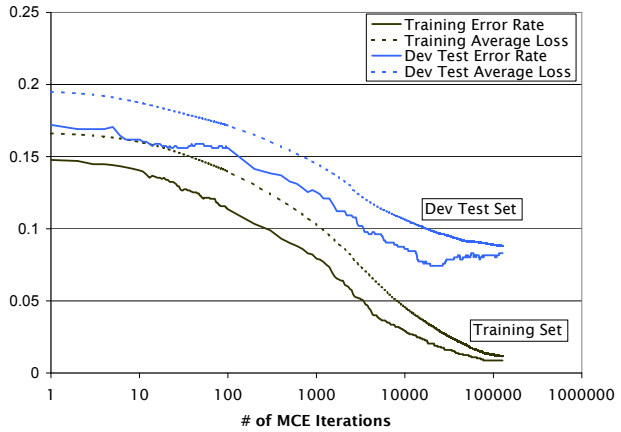


**Fig. 1**. Topic ID error rates and average loss function values for MCE feature weight learning using a 30373 word vocabulary from the output of our word-based ASR system.

| Experimental Conditions | | Topic Error Rate(%) | |
|---|---|---|---|
| Recognition Type | Features | Pre-MCE | Post-MCE |
| English Words | 30373 Words | 16.9 | 7.4 |
| English Words | 3155 Words | 9.6 | 7.9 |
| English Phones | 13899 3-grams | 30.0 | 19.2 |
| English Phones | 3363 3-grams | 22.2 | 21.0 |
| Hungarian Phones | 14413 3-grams | 65.0 | 48.5 |
| Hungarian Phones | 3494 3-grams | 53.0 | 47.7 |

**Table 1**. Closed-set topic ID error rate of the naive Bayes classifier before and after MCE training of the feature weights under various conditions.

125,000 iterations, is reduced to 0.9% (with an average loss function value of .012). On the held out development set, the error rate begins at 17.2% (with an average loss function value of 0.196) and, after 125,000 iterations, is reduced to 8.3% (with an average loss function value of 0.088). Despite the large number of iterations, the loss function on the held-out development data is still decreasing which indicates the training is not over-fitting the training data. A similar trend was observed in our other experiments as well.

Table 1 shows the topic ID performance on our evaluation test set using the recognition outputs from the English word recognition system, the English phonetic recognition system, and the Hungarian phonetic recognition system. In these experiments the final feature weights were determined by selecting the set of weights which minimized the average of the loss function on the development test set. In each case, the table shows the results before and after the MCE feature weight training. Results are also shown for different levels of initial feature selection.

In the case of word recognition, we compare the MCE training using the full recognition vocabulary against using only 3155 preselected word features. Without MCE training, feature selection provides an obvious benefit to the naive Bayes system. However, when using MCE training of the feature weights, better performance was achieved when no preselection of the feature set was used. A similar trend was observed using the English phonetic system where a larger set of preselected trigrams outperformed a smaller set after MCE training was applied. For the Hungarian phonetic system, the larger trigram set did not perform as well as the smaller set, but the benefit of preselecting a reduced set of trigram features was dramatically reduced by the MCE training.

## 5. SUMMARY

In this paper we have presented a discriminative MCE-based approach to feature weight training for topic ID. We have applied this approach to topic ID for human-human telephones conversations using both word-based and phone-based recognition. When tested under various different constraints, relative reductions in topic ID error rates between 9% and 23% were achieved over our baseline system.

## 6. REFERENCES


[1] T. Hazen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Proc. ASRU*, Kyoto, December 2007.

[2] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. Int. Conf. on Machine Learning (ICML)*, Nashville, TN, July 1997.

[3] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proc. ACM SIGIR Conf. on Research and Development in Info. Retrieval*, Dublin, July 1994.

[4] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *Proc. AAAI Workshop on Learning for Text Categorization*, Madison, WI, July 1998.

[5] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, no. 12, Dec. 1992.

[6] E. McDermott and S. Katagiri, "Prototype-based minimum classification error/generalized probabilistic descent training for various speech units," *Computer Speech & Language*, vol. 8, no. 4, pp. 351–368, 1994.

[7] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generation of speech-to-text," in *Proc. Int. Conf. on Lang. Resources and Eval.*, Lisbon, May 2004.

[8] Available from: http://www.ldc.upenn.edu/

[9] J. McDonough, *et al*, "Approaches to topic identification on the Switchboard corpus," in *Proc. ICASSP*, Adelaide, Apr. 1994.

[10] R. Schwartz, T. Imai, L. Nguyen, and J. Makhoul, "A maximum likelihood model for topic classification of Broadcast News," in *Proc. Eurospeech*, Rhodes, September 1997.

[11] J. Wright, M. Carey, and E. Parris, "Statistical models for topic identification using phoneme substrings," in *Proc. ICASSP*, Atlanta, GA, May 1996.

[12] R. Kuhn, P. Nowell, and C. Drouin, "Approaches to phoneme-based topic spotting: an experimental comparison," in *Proc. ICASSP*, Munich, April 1997.

[13] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 137-152, 2003.

[14] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Proc. Int. Conf. on Text, Speech and Dialogue*, Brno, Czech Republic, Sep. 2004.

[15] Available from: http://www.fee.vutbr.cz/SPEECHDAT-E

[16] H.-K. J. Kuo, and C.-H. Lee, "Discriminative training of natural language call routers," in IEEE Trans. on Speech and Audio Processing, vol. 11, no. 1, pp. 24–35, Jan. 2003.

[17] P. Liu, H. Jiang, and I. Zitouni, "Discriminative training of naive Bayes classifiers for natural language call routing," in *Proc. Interspeech*, Jeju Island, Korea, Oct. 2004.