

USING LATENT TOPIC FEATURES TO IMPROVE BINARY CLASSIFICATION OF SPOKEN DOCUMENTS

Jonathan Wintrobe

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD

ABSTRACT

In many topic identification applications, supervised training labels are indirectly related to the semantic content of the documents being classified. For example, many topically distinct emails will all be assigned a single broad category label of “spam” or “not-spam”, and a two-class classifier will lack direct knowledge of the underlying topic structure. This paper examines the degradation of topic identification performance on conversational speech when multiple semantic topics are combined into a single broad category. We then develop techniques using document clustering and Latent Dirichlet Allocation (LDA) to exploit the underlying semantic topics which improve performance over classifiers trained on the single category label by up to 20%.

Index Terms— topic identification, LDA, clustering

1. INTRODUCTION

As the amount of online multimedia containing informal speech increases with social media, online videos, voicemail, and other digitized repositories, the need for rapid searching, filtering, and browsing of spoken documents becomes paramount. As in text information retrieval, topic identification (ID) can be used to improve search results, enrich browsing, or provide filtering of documents (such as spam detection). Topic ID of spoken documents has been part of the repertoire of speech retrieval and browsing since work on the Switchboard corpus in 1993 [1].

Much of the previous work has considered the effects of a variety of automatic speech recognition (ASR) approaches [2] [3], feature selection techniques [4], and non ASR-based approaches [5], on topic ID. This paper examines an often implicit assumption that the label assigned to each document is semantically related to the words or word-based features of the spoken document.

We consider a genre of topic ID tasks such as spam detection, document recommendation, and email prioritization. We are given only a broad category label for training and evaluation such as “spam/not spam”, “like/don’t like”, “wanted/unwanted”. Our labels no longer correspond to the document semantic content. Without any other information,

a standard topic ID approaches will seek to classify these categorical labels directly using word-based features. Any information about latent semantic topic structure remains hidden and not explicitly considered by the model.

This paper examines the effects of broad category labels on the task of detecting topics in conversational speech. We consider a vector space clustering approach and Latent Dirichlet Allocation (LDA) to extract latent topic information in order to better model the target categories. We contrast these approaches using additional latent information with a traditional supervised learning approach that directly models the relationship between word-based features and the given category labels.

2. EXPERIMENTAL SETUP

We use the English Phase 1 section of the Fisher audio corpus, divided into training and testing partitions. We used 1374 conversations for training and 686 conversations for testing. Each conversation is assigned one of 40 topic labels, varying from sports to the Iraq war to time travel, about which the participants were instructed to speak. The corpus and train/test split is described in more detail in [4].

The experiments build on the work in [6] and use the same ASR output in addition to human generated transcripts. For this paper, we only consider the ASR lattice output from the 0.1 times real-time (0.1xRT), and 1xRT systems, where 0.1xRT implies that the recognizer will take 1 minute to process a 10 minute conversation. As described in [6], higher recognition speed results in higher word error rate (WER).

2.1. Experimental Trials

For each trial we uniformly sample $N > 1$ out of the 40 possible Fisher topics, merge them into a single *target* category, then build and evaluate a classifier using this modified labeling. This task is analogous to the spam detection task in that multiple semantic topics (personal, business, etc.) are assigned a single category label of “spam” or “not-spam”, or a recommendation task in which multiple genres are grouped together under the heading of “like” or “don’t like”. We distinguish the modified label, the target *category*, which we are

trying to detect, from the underlying N Fisher topics.

The standard approach would be to take the category labels and build a single classifier, relying on the learning algorithm to discover any hidden structure of the individual topics in order to better identify the broader category. This approach, which we refer to as the *merged* topic classifier, we take as the baseline for all subsequent experiments. All the classifiers described below or their components are support vector machines (SVM). We hold the classification algorithm constant across all trials, allowing comparison with previous work. As before, we used the `svm_light` package [7], with default settings as our classification package.

The baseline system uses the same TF-IDF weighted features described in [6]. The only difference is that we relabel the train and test examples according to the selected *target* and *non-target* categories. For each detection system, we consider N from 2 to 20 (as 21 or more target topics equates to 19 or fewer non-targets). We measure the average equal error rate (EER) across all trials with the same N . We see in Figure 1 that even for the ground truth transcripts, the EER more than doubles after merging only 2 topic labels, and continues to rise until asymptoting at about 7% EER. The baseline system exhibits similar behavior at higher error rates (1x and 0.1xRT ASR systems).

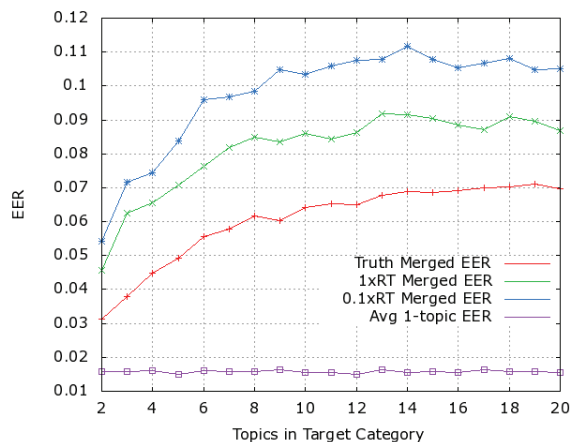


Fig. 1. Baseline SVM classifier with merged labels.

Can we do better than simply assuming the category label given to us is also the semantic topic label? In the following section we examine two approaches to expose the underlying topic structure and use that structure to improve detection of the target category over the single classifier baseline.

3. LATENT FEATURES

In order to build a better classifier for the target category in a mixed-topic scenario, we want to extract additional features that give information about the underlying semantic topic structure. We employ two different approaches for

augmenting the baseline bag-of-words features. We first use vector-space clustering in order to recover the underlying topic labels in an unsupervised manner. We then model the underlying topics as a mixture of hidden variables using Latent Dirichlet Allocation.

3.1. Clustering

We evaluate two clustering-based approaches to exploiting the underlying topic structure. Our first approach uses the `cluto` [8] clustering toolkit to partition the training documents in the target category C . Each cluster is assigned a label and these new labels are used instead of C to train N classifiers, which we denote $C_1 \dots C_N$, where N is the number of clusters. Each unknown document U is then scored against each classifier as a stacked detector, where the score output for U is the maximum score from the N individual classifiers. We will refer to this approach as the *stacked* classifier.

For the clustering step, we consider the feature vectors in category C . We cluster this subset in order to extract the underlying topics. We set the number of clusters to be N , the original number of topics labeled category C . In practice we would choose N based on the amount of training data in category C or using additional development data, which we did not have for these experiments. With perfect clustering we recover the underlying topics completely, whereas in practice we have an approximation of the latent topic labels.

The second approach also begins by partitioning the training documents in category C into N clusters and training classifiers $C_1 \dots C_N$. We then score all *training* documents against each C_i . Taking the score from each C_i plus the maximum model score, we have an additional $1374 N + 1$ -dimensional feature vectors. With these features we train a second fusion SVM, C_f to discriminate category C from \bar{C} . This approach of fusing SVM outputs into a second classifier is modeled after a simplified version of SVM fusion in [9].

Unknown test documents U are scored against each cluster classifier C_i . The scores from each classifier and the maximum model score are used as input to the fusion SVM, C_f . The score from C_f is the output score assigned to U for category C .

The fusion classifier C_f should allow us to model a broader range of underlying topic distributions. The system need not make a hard decision and select the score for a single topic. Instead the fused classifier C_f can consider the scores of the individual classifiers C_i as a weighted mixture of topics the fused classifier C_f can learn, analogous to Latent Dirichlet Allocation, described in the following section.

3.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [10] provides a generative approach to finding underlying topic distributions. LDA assumes that underlying each document is a mixture of latent

topic variables which together generate the observed words. This is a less restrictive assumption than we need for this task, as we assume each document in category C is drawn from only one of N topics, rather than a mixture. The more general LDA model ought to be more appropriate to conversational speech in which speakers naturally switch or go off-topic during the course of the 5-10 minute Fisher conversations.

We use the GibbsLDA++ [11] implementation of LDA parameter estimation to infer topic distributions for the training data. Similar to clustering, LDA requires specification of the number N of latent topic variables. Previous work has indicated that models estimated with higher values of N performs better on text classification [12]. However we have limited training data, relative to text corpora, to estimate models with higher values of N , so we may have a sub-optimal solution. We chose N empirically to be 100 for the results reported here. Setting N between 50 and 200 resulted in nearly identical performance.

Our LDA-based classifier C_{LDA} is trained using the 100 LDA topic weights for each training document, plus the score for each training document against the baseline merged classifier. Using the LDA topic weights by themselves resulted in a system performing much worse than the baseline, quickly climbing to 10% EER with 5 merged topics. The parameters of the LDA model are used to perform inference on an unknown document U , which is also scored against the baseline model. These $N+1$ features are then scored against the fused classifier C_{LDA} .

4. RESULTS

To evaluate classifier performance as the number of topics in the target category C increases, we varied the number of topics in C from 2 to 20. For each level N we evaluated our baseline, cluster-based, and LDA-based systems. Using the ground truth transcripts we ran 300 trials for each N . The results for each system are shown in Figure 2. We measured the equal error rate (EER) averaged over all trials with a given N . We chose EER over other metrics as we are concerned only with the detection of category C , often out of a significantly larger number of non-target documents (i.e spam).

The results indicate first that the cluster-based stacked detector does no better than merging and training on the single category label C . In fact, as N increases, it does much worse. However, by using a second classifier to fuse the clustered topic detectors, the underlying topic structure does improve performance from 2.5 to 15.3% relative to the baseline detector. However, for $N > 10$, the fused cluster-based detector performs increasingly worse.

Closer analysis shows that the both cluster-based systems are highly dependent on the error rate of the component classifiers. When considering the N topics in category C , the maximum single topic EER was strongly indicative of the overall EER of the cluster-based systems. When we plot the

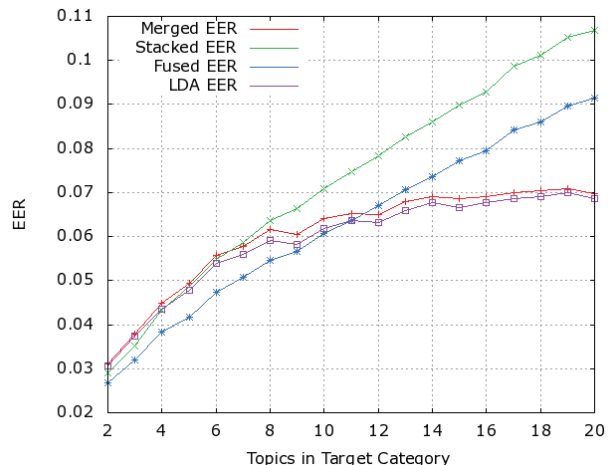


Fig. 2. Topic detection results for 2-20 merged topics.

average of the maximum single topic EER against the average EER of the fused detector for all values of N we see the strong linear relationship. Intuitively, as N increases, the likelihood increases of including a hard to detect topic in C .

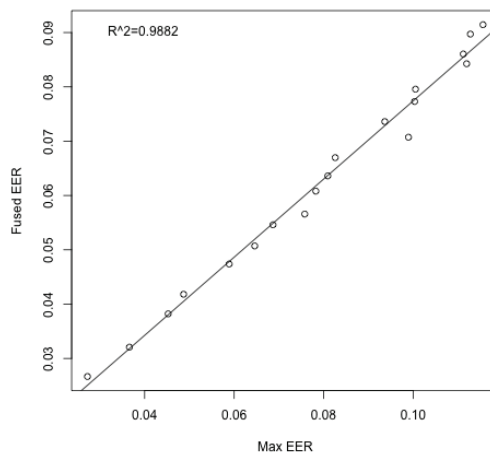


Fig. 3. Fused system linear dependence on maximum single-topic error in category C ($R^2 = 0.99$).

On the other hand, the LDA-based system gives an improvement over the baseline for all values of N . Adding the LDA topic weights improves performance from 1.2 to 4.1% relative to the baseline. The baseline merged detector and the LDA-based system do not exhibit the same strong linear relationship to the individual topic detectors ($R^2 = 0.89$). Both are more robust, on average, in the presence of individually harder to detect topics.

4.1. ASR Results

We evaluated the cluster-based approaches on ASR lattice output from the 1xRT and 0.1xRT systems. While overall the error rates were higher, we observed the same trends as for the human transcripts. Using the 1xRT lattice output, the fused classifier improved the EER from 1.2 to 11.1% relative to the baseline for $N \leq 10$. Using the higher error rate 0.1xRT output, the fused classifier also improved the error rate over the baseline from 3.8 to 20.9% for $N \leq 15$. Surprisingly the fused detector from the 0.1xRT system was only slightly worse than the fused detector from the 1xRT system, indicating robustness in the presence of higher word error rate (WER).

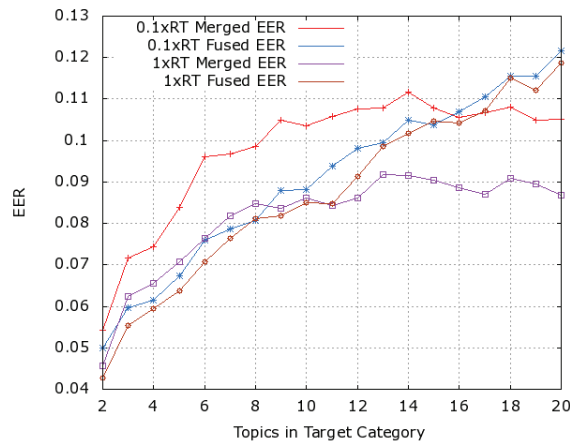


Fig. 4. Topic dection of ASR results.

5. CONCLUSIONS

When we look to identify documents from a target category C comprised of multiple underlying semantic topics, the accuracy of a single classifier decreases significantly as N , the number of different topics in C increases. Even for a target category of 2 topics, the average EER increases by 158%, 45%, and 46% when using human transcripts, 1xRT, and 0.1xRT ASR output respectively. By extracting information about the latent topic structure of the data, we have shown we can reduce the error identifying category C from 1 to 20% relative to the baseline.

Given the different performance characteristics, we believe the approaches are complimentary. We observe that the cluster-based systems give a larger magnitude decrease in EER than the LDA system but over a smaller range of N . The clustering techniques exploit the topic structure of the documents in category C , whereas the LDA system estimates the topic mixture using the entire training corpus. It is worth exploring in future experiments if a combination of the two sets of features can further reduce the error when given only broad

category labels.

In addition to considering classifiers other than SVMs, there has been recent work on supervised LDA [13] which gives a formal model for the relationship between observed categories and unobserved topics, which would be worth adapting to this particular task.

6. REFERENCES

- [1] B. Peskin et al., "Topic and speaker identification via large vocabulary continuous speech recognition," in *Proc. of ARPA Workshop on Human Language Technology*. ARPA, 1993.
- [2] F. Hazen, T. Richardson and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Kyoto, Japan, 2007.
- [3] R. Kuhn, P. Nowell, and C. Drouin, "Approaches to phoneme-based topic spotting: An experimental comparison," in *Proc. of ICASSP*, Munich, Germany, 1997.
- [4] T. Hazen and A. Margolis, "Discriminative feature weighting using MCE training for topic identification of spoken audio recordings," in *Proc. of ICASSP*, Las Vegas, NV, 2008.
- [5] M. Dredze, A. Jansen, G. Coppersmith, and K Church, "NLP on spoken documents without ASR," in *Proc. of EMNLP*, Cambridge, MA, 2010.
- [6] J. Wintrode and S. Kulp, "Techniques for rapid and robust topic identification of conversational telephone speech," in *Proc. of Interspeech*, Brighton, U.K., 2009.
- [7] "SVM light," <http://svmlight.joachims.org>.
- [8] "Cluto," <http://www-users.cs.umn.edu/~karypis/cluto/>.
- [9] L. Marvel, B. Henz, and C. Boncelet, "Fusing rate-specific SVM classifiers for ± 1 embedding steganalysis," in *42nd Annual Conference on Information Sciences and Systems*, Princeton, NJ, 2008.
- [10] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning*, 2003.
- [11] "Gibbs LDA++," <http://gibbslda.sourceforge.net/>.
- [12] T. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. of National Academy of Sciences*, 2004.
- [13] D. Blei and J. McAuliffe, "Supervised topic models," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.