Confidence-Based Techniques for Rapid and Robust Topic Identification of Conversational Telephone Speech

Jonathan Wintrode[1], *Scott Kulp*[2]

[1] U.S. Department of Defense
[2] Department of Computer Science, Rutgers University
jcwintr@tycho.ncsc.mil, sckulp@cs.rutgers.edu

Abstract

We investigate the impact of automatic speech recognition errors on the accuracy of topic identification in conversational telephone speech. We present a modified TF-IDF feature-weighting calculation that provides significant robustness under various recognition error conditions. For our experiments we take conversations from the Fisher corpus to produce 1-best and lattice outputs using one recognizer tuned to run at various speeds. We use SVM classifiers to perform topic identification on the output. We observe classifiers incorporating confidence information to be significantly more robust to errors than those treating output as unweighted text.
Index Terms: topic identification, speech recognition, feature selection

1. Introduction

The primary task we explore in this paper is the association of topic labels with a spoken audio conversation - topic identification. We also look at the related task of deciding whether or not a given topic occurs within the conversation - topic detection. Having a putative topic label for an audio document can be useful in more *accurately* organizing, sorting, clustering, prioritizing, filtering, or searching spoken audio archives. We focus as well on the *rapid* processing of large audio corpora. As the volume of digital media increases, the need for the ability to process that media in a timely manner increases in importance. We assume that we can process the audio with some form of automatic speech recognition (ASR) and that each audio document is represented by noisy ASR output. Given that we can make a tradeoff between speech recognition accuracy and recognition speed, we explore what, if any, degradations occur at higher speeds and lower accuracy.

While previous work has looked at topic identification (topic ID) performance under degraded conditions, in this work we also consider techniques to ameliorate those degradations. In particular we look at using confidence information from higher speed ASR systems to recover from higher word error rates (WER) and topic ID error rates. We will demonstrate that we can improve topic ID performance on the higher WER systems by 25-40%.

1.1. Related Work

Previous related work on topic identification, classification, or detection tasks has addressed both broadcast and conversational telephone speech genres and has looked at various ASR capabilities: word-based and phonetic-based recognition, in-language and out-of-language recognition, and true transcripts, 1-best hypotheses, and ASR lattice output.

Early work on the Switchboard conversational telephone corpus by Peskin et al in [1,2] suggested that topic ID from 50-60% accurate ASR output can do as well as topic ID from human transcripts [2]. Their test set consisted of 120 conversations, evenly split among 10 topic labels. The topic models (as differentiated from the ASR acoustic models) were trained on the human transcripts, rather than recognition output.

In 1997, NIST defined a "Topic Detection and Tracking" task in the Broadcast News domain [3]. This task now includes 3 corpora (TDT1, 2, and 3) which contain 25, 100, and 60 topics, respectively, as well as roughly 116,000 audio documents from both English and Mandarin Chinese news sources. The TDT corpora represent a more difficult task in terms of number of documents and number of topics than the Switchboard task. McCarley and Franz, working on the first and English-only TDT corpus, looked specifically at the impact of speech recognition errors on the topic detection task [4]. They did find significant degradations in topic detection performance when comparing systems using 65-70% accurate ASR output to systems using human transcripts. As with the work on the Switchboard corpus, there was only a comparison between true transcripts and one set of ASR transcripts, not ASR systems of varying accuracy. The difference in topic detection performance could only be attributed to the presence of errors, rather than any specific level of errors in the automatic transcripts.

Most recently and most closely related to the work described in this paper, is work by MIT Lincoln Laboratory performing topic ID on the Fisher conversational telephone corpus [5,6]. We describe the corpus in more detail in the next section. The size of the task in terms of topics and conversations falls in-between the Switchboard task and the TDT tasks. The portion of the Fisher corpus used for topic ID experiments consists of roughly 2000 conversations divided into 40 topics.

The work at MIT looked at topic ID under a variety of ASR systems, including word-based and phonetic-based ASR systems, as well as out-of-language recognition (using BUT's Hungarian phonetic recognizer applied to the English audio). Their results show no significant degradation in topic ID accuracy when comparing topic classifiers built from English word-based ASR output to classifiers built from the human transcripts (8-10% ID error rate). However, the topic classifiers built from English and Hungarian phonetic recognition output more than doubled the ID error rate for each degradation in training quality from 8% (English words) to 23% (English phones) to 53% (Hungarian phones). Their subsequent work using discriminative feature selection [6] improved performance across all types of recognition systems. The relative error rate reduction for the two poorer performing

systems was only 16% for English phones and 10% for Hungarian phones.

2. Experiment Description

For this paper we performed experiments in both topic identification and detection using the Fisher audio corpus. To perform identification we took the maximum scores of individual detectors. For the sake of brevity, unless otherwise noted, we will refer to the combination of the two tasks as topic ID. We trained and evaluated topic ID systems using recognition output from multiple configurations of the same recognizer, described in Section 2.2. For the purposes of establishing an upper limit of our topic ID algorithm with no ASR errors we also trained a system from the Fisher human-generated transcripts. Our topic ID systems are built using SVM classifiers with linear kernels. The overall topic ID training and evaluation procedure is described in detail in Section 2.3.

2.1. Corpus

For our experiments, we used the English Phase 1 section of the Fisher audio corpus, which is described in more detail in the literature [7]. Of the 5851 conversations in the corpus, we use the 1375 conversation topic ID training partition for building our topic classifiers, and we use only the 686 conversation evaluation partition for that purpose. These are the same sets defined in [5] so as to facilitate comparison with prior work, only we do not make use of the development partition or the recognizer training partition.

For the topic ID task, each 2-sided conversation is treated as a single document and assigned one of 40 topic labels. These topic labels were used to prompt the collection of data for the Fisher corpus and range from “Movies” to “Sports” to “Time Travel.” Our task, given the set of training audio documents and associated training labels, is to build models and predict the topic labels for the documents in the evaluation set.

2.2. Speech Recognition System

We decoded the topic ID training and evaluation sets of the Fisher corpus using three configurations of BBN’s Byblos automatic speech recognition (ASR) system. Our configurations represent three operating points in terms of recognition speed and accuracy. For convenience, we define them in terms of their decoding speed relative to the number of hours of audio processed (xRT). A brief description of the configurations follows:

- 10xRT – An 8 pass system, 4 passes without speaker adaptation (forward/backward with and without cross-word context), followed by 4 passes with speaker adaptation (forward/backward, with and without cross-word context) [8].
- 1xRT – A 2 pass system, forward and backward, with no speaker adaptation [9].
- 0.1xRT – Same as the 1xRT system, but with more aggressive beam pruning during recognition [9].

All three configurations use 5-state clustered HMM’s. The systems differ in the number and type of passes, pruning parameters, and the use of cross-word context (only used by the 10xRT system). All configurations also share a common acoustic and language model, trained from 378 hours of the Switchboard corpus. For each configuration we output both 1-best transcripts, with word-level confidences, and word lattices with word-level posterior probabilities.

2.3. SVM Topic Detection and Identification

Once the training and evaluation cuts were decoded, we trained a set of SVM classifiers for topic detection and then identification. We used Thorsten Joachims' SVM-Light package available at [10]. During training we built a 2-class classifier for each topic *T*, in which cuts labeled with topic *T* were labeled as ‘Target’ for that classifier and all others were labeled as ‘Non-target’. Building classifiers for each topic allowed us to easily perform both topic detection and identification tasks.

Our training procedure consists of the following:

- Decode all training cuts.
- For each document in the training corpus, calculate term frequency for all words in the document.
- For all words in the training corpus, calculate the document frequency over the corpus.
- For each document, calculate TF-IDF scores for each word possibly occurring in the document.
- Generate a feature vector for each document using the top *M* scoring words, ranked by TF-IDF.
- For each topic, train a 2-class SVM classifier using these feature vectors.

For evaluation, our procedure was similar to training:

- Decode all evaluation cuts.
- For each document in the evaluation corpus, calculate term frequency for all words in the document.
- For each document, calculate TF-IDF scores for each word possibly occurring in the document, using the IDF values from the training corpus. Words that do not occur in the training corpus are ignored.
- Generate a feature vector for each document using the top M scoring words, ranked by TF-IDF.
- For detection, score each topic classifier against the corpus using these feature vectors.
- For identification, use the highest scoring detector on a cut as the hypothesized label for that cut.

In the following section, we describe how we generated and selected features for training and evaluating our SVM classifiers. To measure performance of the detection task we calculated the equal error rate (EER) for individual detectors, and for the 40-topic identification task we measured the overall error rate.

3. Feature Vector Generation

SVM training and classification requires we transform each document into a vector of numerical features. The SVM algorithm takes these vectors in high-dimensional space, along with a category value {+1,-1} and seeks to draw a hyperplane in that space that best separates the two classes.

In terms of the Vector Space model from information retrieval, we would like to generate a document vector V_d which captures both intra-document and inter-document similarities [11]. We begin with TF-IDF (term frequency, inverse document frequency) weights, which have been developed precisely for such a purpose. Likewise, document frequency by itself has been shown to be an effective method of feature selection [12] in text categorization, and TF-IDF weights have been used in both topic extraction and detection for NIST’s Topic Detection and Tracking task (TDT) [13].

3.1. Feature Weighting

For each document d, we define a feature vector V_d with weights $w_{d,i}$, as input to the SVM classifier as follows:

$$V_d = \lfloor w_{d,1}, w_{d,2}, ..., w_{d,k} \rfloor \quad (1)$$

$$w_{d,i} = tf(i,d) \cdot \log\left(\frac{N}{df(i)}\right) \quad (2)$$

where k is the number of words occurring in the corpus, N is the number of documents in the corpus, and *tf* and *df* are term frequency and document frequency respectively.

Normally, term frequency and document frequency are obtained by counting occurrences of words in the corpus. For the case in which we treat ASR transcripts as text, this is precisely what we do. However, we also considered topic ID under two additional cases where 1) there is a confidence associated with each word in the transcript, and 2) we consider the entire lattice of words hypothesized by the recognizer, each of which has an associated posterior probability. In both cases, we would like to discount words in the transcript the recognizer deems unlikely as actually having occurred, and in the second case we would also like to allow for those additional hypotheses that do not occur in the 1-best transcript.

For this purpose we approximate *tf* and *df* calculations probabilistically as the expected term frequency (*etf*) and estimated document frequency (*edf*), We use these approximations for the SVM vector weights $w_{d,i}$:

$$tf(i,d) \approx etf(i,d) = \sum_{j=1}^{n} P(t_j = i \mid o_j) \quad (3)$$

$$df(i) \approx edf(i) = \sum_{d} \min(1, etf(i,d)) \quad (4)$$

$$w_{d,i} = etf(i,d) \cdot \log\left(\frac{N}{edf(i)}\right) \quad (5)$$

We take the *etf* of a term to be the confidence-weighted sum of the n hypothesized occurrences of term i in the document. For lattices, this corresponds to the expected count of hypothesized words in the lattice, which we obtain using the *lattice-tool* from SRI's LM toolkit [14]. For 1-best transcripts, the expected term frequency is the term frequency weighted by the word-level confidence for each hypothesized word. Because we use SVM classifiers in which each word in the training constituted a unique dimension in the feature space, we ignored words in the evaluation data that did not occur in training. This guaranteed that we never used terms with an *edf* of 0.

An advantage of this approach is that we can estimate document frequency entirely in terms of our term frequency estimates, without the need to convert to intermediate data structures, such as confusion networks, as has been proposed in [15-16] in the context of spoken document retrieval. Considering topic ID from ASR lattices, it is particularly important to approximate the document frequency as well as the term frequency, because any unweighted calculation on the lattices will overestimate counts.

3.2. Feature Selection

Rather than limit the overall number of words K as candidates for feature selection, we instead chose to limit the number of nonzero weights in the vector to a fixed value M. For each document we selected the top M terms with the greatest feature weights $w_{d,i}$. In all of the experiments described in the following section, we set M=500. For some conversations this effectively included all words in the conversation. In some cases there were small gains in reducing M to 200 or 100, effectively reducing our vocabulary size K. However, our techniques incorporating word-level confidences and lattice posteriors into the TF-IDF calculation proved to be more effective, and required less tuning. This presumably drives certain weights near zero, providing the same effect as reducing M or K.

4. Experimental Results

For our first experiment we generated SVM feature vectors from 1-best transcripts treating those transcripts as unweighted text and calculating TF-IDF using Equation 2. For each of our three systems (10xRT, 1xRT, 0.1xRT) we also measured WER on both topic ID training and evaluation sets. For the sake of brevity, we only report the WER measured on the evaluation data, as the WER on the training was within 1-2% of the value given. As an upper bound, we also build classifiers using the ground truth, human generated transcripts.

Table 1. WER and Topic ID performance.

System	WER (%)	ID Error (%)	Avg. EER (%)
Truth	0	10.2	1.2
10xRT	34	10.1	2.6
1xRT	45	19.1	4.8
0.1xRT	47	19.2	4.6
Adapt. BW pass	36	10.2	2.7
Adapt. FW pass	42	10.2	2.6
Unad. BW pass	40	11.1	2.7
Unad. FW pass	48	11.1	2.8

When looking only at the full system configurations we observed degradation in topic ID performance at higher WER of the 1xRT and 0.1xRT system a expected. At 34% WER, the 10xRT system performs almost identically to the true transcripts on the identification task. We don't presume the 0.1% difference in ID error rate between the truth and 10xRT transcripts to be statistically significant, and we *do* see evidence of the effect of the 10xRT system's 34% WER in the performance of the individual topic detectors,. Their average performance more than doubles from the ground truth baseline.

However, when we consider 4 passes of the more accurate 10xRT system, we see WER degradation in the initial unadapted passes without the same degradation in topic ID performance observed in the 1xRT and 0.1xRT configurations. The primary difference between these unadapted passes and the 1xRT and 0.1xRT systems is that these passes consider cross-word context whereas the fast systems do not. This discrepancy suggests that there is enough information in the early passes to perform the topic ID task at high speeds as well as if we had true transcripts.

To verify this, our second set of experiments involved generating feature vectors from the ASR lattices of the 1xRT and 0.1xRT systems. The vector weights were calculated using the expected TF-IDF calculation described in Equation

5, setting M=500. Table 2 shows a 40% relative decrease in topic ID error rate for the 1xRT configuration using recognition lattices, and a 35% relative decrease in error rate for the 0.1xRT configuration.

Table 2. Topic ID performance with transcript and lattice-derived features.

System	ID Error (%)	Avg. EER (%)
Truth	10.2	1.2
10xRT	10.1	2.6
1xRT	19.1	4.8
1xRT (lattice)	**11.4**	**3.1**
0.1xRT	19.2	4.6
0.1xRT (lattice)	**12.5**	**3.7**

We then tried two approaches to apply the technique to confidence-weighted 1-best transcripts. If the lattice information from the faster systems was sufficient for performance approaching that using ground truth, perhaps confidence weighting transcripts would be sufficient as well. First we approximated TF-IDF using the confidences to only estimate *etf*, while using the unweighted *df* treating the transcripts as text. Secondly, we applied the confidence weights to both *etf* and *edf* calculations, as described in Equation 5. As the results show, weighting document frequency as well as term frequency by word confidences is essential for the higher WER 0.1xRT system.

Table 3. Topic ID performance of various confidence-weighted features applied to transcripts.

System	ID Error (%)	Avg. EER (%)
Truth (unweighted)	10.2	1.2
10xRT (unweighted)	10.1	2.6
1xRT (unweighted)	19.1	4.8
1xRT (etf, no edf)	13.7	4.2
1xRT (etf, edf)	**13.7**	**4.2**
0.1xRT (unweighted)	19.2	4.6
0.1xRT (etf, no edf)	17.2	5.0
0.1xRT (etf, edf)	**14.4**	**3.9**

5. Conclusions

When topic ID classifiers are built using ASR output treated as text, we observe a significant increase in both detection and identification error as recognition speeds increase to 0.1xRT. Yet, the lack of a significant increase in topic ID error for the unadapted passes of the 10xRT system using cross-word context, which exhibit similar WER to the 0.1xRT system, suggests that this degradation in topic ID performance is less attributable to the overall decrease in word error rate but rather to the types of errors.

On the other hand, in spite of the higher topic ID error observed in our fastest (0.1xRT) system, we see that at nearly 50% WER, there is sufficient information in either the word recognition lattice or the 1-best confidence scores to reduce the topic ID performance degradation by 25-40%. This result is relevant to the amount of audio data a topic ID system can process accurately. Our estimated TF-IDF feature-weighting calculation allows us to leverage confidence information to achieve topic ID accuracy comparable to the 10xRT system with an ASR system that is running 100 times faster.

6. Future Work

Our conclusions suggest the need for a more detailed error analysis of the 10xRT unadapted pass word errors as compared to the 0.1xRT word errors. Also, in this work we present a simple but effective feature selection mechanism. We would expect to extend our analysis to other feature selection techniques.

7. Acknowledgements

We are indebted to TJ Hazen from MIT Lincoln Labs for his sharing of the Fisher topic corpus definitions, and also to Wade Shen for early feedback on these experiments. This work could not have gotten off the ground without their help.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2007-ST-104-000006. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

8. References

[1] Peskin B. et al., "Topic and Speaker Identification via Large Vocabulary Continuous Speech Recognition," in *Proc. ARPA Workshop on Human Language Technology*, Princeton, March 1993.

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

[3] Wayne, C., "Multilingual Topic Detecting and Tracking: Successful Research Enabled by Corpora and Evaluation," in *Proc. LRE* 2000.

[4] McCarley, J., Franz, M. "Influence of Speech Recognition Errors on Topic Detection", in *Proc. SIGIR 2000,* pp. 342-344.

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

[6] Hazen, T., Margolis, A., "Discriminative Feature Weighting using MCE Training for Topic Identification of Spoken Audio Recordings," in *Proc. ICASSP*, Las Vegas, April 2008.

[7] Cieri, C., et al., "The Fisher Corpus: A resource for the next generation of speech-to-text,"in *Proc. Of Int. Conf. on Language Resources and Evaluation*, Lisbon, Portugal, May 2004

[8] Prasad, R., et al., "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech System," in *Proc. Rich Transcription Workshop*, Palisades, NY, Nov. 2004.

[9] Colthurst, T., et al., "Parameter Tuning for Fast Speech Recognition," in *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007.

[10] http://svmlight.joachims.org

[11] Baeza-Yates, R., Ribiero-Neto, B., *Modern Information Retrieval*, 1999, pp. 27-30

[12] Yang, Y, Pederson, J., "Ä comparative study on feature selection in text categorization," in *Proc. ICML*, Nashville, TN, July 1997.

[13] Sista, S., et al., "Unsupervised Topic Discovery Applied to Segmentation of News Transcription", in *Proc. Eurospeech*, Geneva, 2003.

[14] Stolke, A., "SRILM – An Extensible Language Modeling Toolkit," in *Proc. International Conference on Spoken Language Processing*, 2002.

[15] Weschler, M., Schäuble, P., "Speech Retrieval Based on Automatic Indexing," In *Working Notes IJCAI Workshop: Intelligent Multimedia Information Retrieval*, Montreal, Canada, 1995, pp. 59 – 69

[16] Mamou, J., Carmel, D., Hoory, R. "Spoken document retrieval from call-center conversations," in *Proc. SIGIR*, 2006, pp. 51-58.