LIMITED RESOURCE TERM DETECTION FOR EFFECTIVE TOPIC IDENTIFICATION OF SPEECH

Jonathan Wintrode and Sanjeev Khudanpur

Center for Language and Speech Processing, Johns Hopkins University, Balitmore, MD
jcwintr@cs.jhu.edu,khudanpur@jhu.edu

ABSTRACT

We consider the task of identifying topics in recorded speech across many languages. We identify a statistically discriminative set of topic keywords, and examine the relationship between overall word error rate (WER), keyword-specific detection performance, and topic identification (Topic ID) performance on the Fisher Spanish corpus. Building increasingly constrained systems—from copious to limited training LVCSR to limited-vocabulary keyword spotting—we show that neither high WER (>60%) nor low-precision term detection (<40%) are necessarily impediments to Topic ID. By using deep neural net acoustic models for keyword spotting, we can double recall and ranked retrieval performance over comparable PLP-based models and achieve Topic ID performance on par with well-trained LVCSR or human transcripts.

Index Terms— Automatic speech recognition, spoken term detection, topic identification, deep neural networks

1. INTRODUCTION

With the spread of new communications technologies across the globe, such as smart phones and social media, there is an explosion of user-generated multimedia content in all languages. For example, users upload 100 hours of video to YouTube every minute, and 70% of YouTube traffic comes from outside of the United States [1]. Yet in spite of this wealth of language-rich content, little of the content itself is used in organizing, analyzing, or accessing the information contained therein.

One factor limiting indexing of this content is the lack of linguistic resources to build automatic systems in many languages. A second is the volume of data on large sites like YouTube. We examine and address both concerns in the context of supervised *topic identification* (Topic ID) of informal speech. We demonstrate the viability of the Topic ID task even under resource-limited conditions, by using *limited training* large vocabulary continuous speech recognition (LVCSR) and *limited vocabulary* keyword spotting with deep neural net (DNN) acoustic features. The goal of this work is to quantify the constraints inherent in currently available techniques with an eye towards developing targeted approaches in the future.

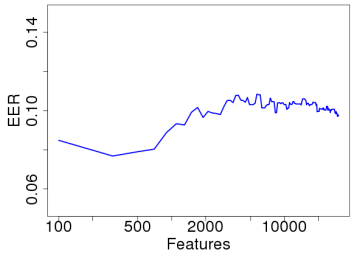

Fig. 1. EER by vocabulary size (chosen by χ^2 metric), Topic ID on Fisher Spanish human transcripts.

The nature of the Topic ID task makes it amenable to extracting information from spoken content in a language-rich digital environment. To identify topics in speech, typically some form of LVCSR is followed by categorization based on the extracted word or subword *tokens*. Most algorithms for text categorization (e.g., spam filters, document routing, author attribution) operate on bags-of-words, which are simply *accumulated token counts*. As a consequence, one need not be constrained by the accuracy of particular token instances, i.e. by the word error rate (WER), and good performance may be possible even with very limited resources.

To address high data volumes, related work on feature selection ([2], [3]) suggests that the use of fewer, more discriminative words may result in equal or better performance than using the entire vocabulary. We demonstrate this effect in a 25-class Topic ID task on the Fisher Spanish corpus. We selectively increase the vocabulary used for Topic ID based on a χ^2 statistic (cf [2]). Figure 1 shows we achieve the best error rates using only a small fraction (2-3%) of the vocabulary. This suggests that very *limited vocabulary* solutions may be used in a high volume setting, and one need not extract a large bag of all possible word types.

We will therefore focus on the top 1000 keywords according to the χ^2 metric in this paper. Given the result above, we will show to what extent we can limit the training or vocabulary requirements while still maintaining acceptable Topic ID performance.

1.1. Related Work in Resource-Rich Topic ID

Early work on Topic ID (cf [4]) demonstrated topic classification error rates using LVCSR transcripts that were comparable to what can be achieved using manual transcripts. The 0.8% error on the 10-topic task on the Switchboard corpus using 43.9% WER transcripts was so low that until recently, little follow-up work has been done on this corpus.

Later work focused on the larger Fisher English corpus, with 40 labeled topics. Hazen and Margolis [3] reported 8.2% ID error using manual transcripts, suggesting a more difficult task than Switchboard, irrespective of WER. Their work demonstrated the utility of indexing speech lattices from a good LVCSR system, achieving a 9.6% ID error rate.

Unlike Switchboard, however, the gap between high- and low-resource settings is not negligible on Fisher. Trading decode speed for accuracy, [5] found a significant degradation in ID error (from 10% to 19%) when WER fell to 47%, and even this WER is higher than the typical 60-70% WER observed for the limited-resource "Limited LP" training condition in the IARPA Babel program [6]. *This makes the low-resource Topic ID task worth revisiting using the Fisher corpus.*

1.2. Limited Resource Approaches to Topic ID

Low or zero-resource techniques encompass supervised and unsupervised methods. Typical supervised approaches train phonetic or subword recognition systems, on the assumption that such systems require less training data and avoid vocabulary limitations. However in [3], using tokens from English phonetic recognizers, ID error on Fisher more than doubles from from 9.6% to 22.9%. Using non-English phone recognizers, the error more than doubles again to 53%. Discriminative training of Topic ID feature weights reduces error on English and non-English phonetic tokens to 19.2% and 47.7% respectively [7]. But this still represents a significant degradation in performance from a word-based approach.

Unsupervised acoustic modeling tokenizes speech without transcribed training data. In their work on self-organizing units (SOUs), Siu et al., achieved 45.9% error on Fisher using HMMs and Segmental GMMs to discover word-like units from 4 hours of untranscribed English data [8]. Dredze et al., reported 7.5% ID error on the Switchboard task, counting clusters of repeated acoustically similar segments or "pseudoterms" in the corpus of interest [9]. Both SOU and pseudoterm techniques exhibit higher Topic ID error rates than word-based recognition but on par with subword approaches, which do require some transcribed data to implement.

1.3. Proposed Work

In this paper, we work on the Fisher Spanish conversational speech corpus [10], performing Topic ID experiments under increasingly limiting conditions. We first build a full vocabulary LVCSR system with the Kaldi speech recognition toolkit, limited to 14 hours of acoustic and language model data.

Second, using the 1000 most discriminative unigrams from Figure 1 and using the full-vocabulary system, we examine the relationship between Topic ID performance and term detection accuracy, as opposed to overall WER.

Third, we construct an HMM-based keyword spotting system, also using the Kaldi framework. We train on the same 14 hour audio corpus, but assume that only the instances of the top 1000 keywords are annotated for acoustic training.

Each of these constrained systems will provide successively stronger evidence for the robustness (or lack thereof) of the Topic ID "signal" in the presence of increasingly errorful word-token streams.

2. EXPERIMENT DETAILS

For all Topic ID experiments, we divide the Spanish Fisher corpus into two sets, 643 conversations for classifier training and 176 for evaluation, was was done in [11]. Restricting ourselves to the Spanish Call Home training vocabulary, we find that 99 of the top 1000 keywords are out of vocabulary and thus unavailable to the keyword spotter and topic classifiers.

Our classification setup is the same as [3] and [5]. We train 1-vs-all classifiers for each of the 25 topics and report results averaged over all topics. We record a variety of performance metrics, related to different application uses. For consistency with previous work we focus on: **Error**, the percentage of incorrect topic labels assigned, **AUC**, the area under the precision-recall curve, and **EER**, the (equal) error rate at the detection threshold at which $P(FalseAlarm) = P(Miss)$.

Construction of feature vectors for classification is a rich area of research unto itself, and a good overview may be found in [12]. For the SVM classifiers we use, we construct bag-of-words feature vectors using TF-IDF weights estimated either from human transcripts or decoder confidence scores in the same manner as [5]. The vector weights are all computable from the token counts $\mathrm{TF}_{i,d}$: the estimated number of of times a term t_i occurred in document d.

For Naive Bayes classifier, we use a log likelihood formulation described in detail in [11]. The scores for this classifier are also computed from bag-of-words features. The class-specific probabilities $P(t_i|c)$ are estimated from token counts on the classifier training data. The advantage of a bag-of-words approach, in the low resource setting, is that for either classifier we can use any method of tokenization, not just LVCSR, from which we may obtain token counts.

A summary of baseline classification results from manual transcripts is given above in Table 1. Results with and without feature selection are given for completeness. Subsequent sections will show that *one can approach or equal this baseline Topic ID performance with very limited training LVCSR, and even with limited vocabulary keyword spotting*.

Classifier	Vocab	Error	AUC	EER
Naive Bayes	30k	25.5%	0.835	0.122
	1k	22.1%	0.856	0.073
SVM	30k	17.1%	0.903	**0.071**
	1k	**15.9%**	**0.904**	0.072

Table 1. Topic ID performance using human transcripts.

3. LIMITED TRAINING LVCSR

In the limited training data condition, we use the Kaldi speech recognition toolkit [13] to train a 45K word Spanish LVCSR system on only the 14 hour Spanish Call Home data [14]. The vocabulary and pronunciations are also restricted to the Call Home lexicon. The Spanish Fisher corpus has 178 hours of speech containing 33.8K word types and 1.6M word tokens, with out-of-vocabulary (OOV) rates resp. of 47% and 5.7%.

We use the Kaldi training recipe for the IARPA Babel Limited LP condition described in detail in [6]. The acoustic models use 13-dimensional PLP features. PLPs are used to train both speaker-independent and speaker-adapted triphone models (the *tri5* models), using typical state-clustered HMM's with GMM output densities. The recipe then trains Subspace GMM's [15] (SGMM) on the output densities (*sgmm5* models). The SGMM parameters are boosted with a maximum mutual information (MMI) criterion (*sgmm5_mmi* models). All models use a trigram language model estimated on the training transcripts.

We also use Kaldi's CPU-based deep neural net (DNN) acoustic features in a hybrid HMM-DNN configuration [16] (denoted as *tri6_nnet*). However for small training sets (∼10hrs) Kaldi uses a smaller network configuration of only 2 hidden layers and 879 input and output dimensions. The DNN features had little impact on the full vocabulary LVCSR results, but large impact on the keyword spotting results, as will be evidenced in a subsequent section.

With the Fisher corpus as a test set, we observe WERs between 53.3 and 62.9% for the different acoustic models, almost a 10% absolute difference between the best and worst transcript output. As shown in Figure 2, there is a modest correlation (0.5) between WER and EER, the most stable Topic ID metric. Yet the total effect on Topic ID performance is minimal. The dashed lines denote the best and worst performance from the human transcript baseline, and with few exceptions, performance of the LVCSR systems falls between these bounds.

4. KEYWORD RETRIEVAL AND TOPIC ID

Using the lattice-based keyword search module of Kaldi, we use our 1000 keywords as input to a retrieval experiment on the Fisher Spanish corpus. We treat the unfiltered search results over word lattices as confidence weighted token counts, train classifiers as described previously, and measure both keyword retrieval efficacy (via term weighted value) and Topic ID performance.

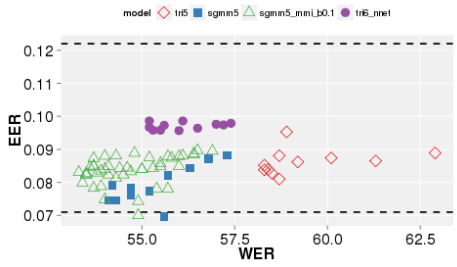

Fig. 2. EER v/s WER (obtained by varying LM weights in the LVCSR system)

Model	TWV	Error	EER	AUC
tri5 + NB	0.340	**0.186**	0.085	0.843
tri6_nnet + NB	0.389	0.193	**0.070**	0.846
sgmm5_mmi + SVM	0.391	0.216	0.087	**0.877**

Table 2. Retrieval and Topic ID performance for top-scoring systems (Full vocab LVCSR plus 1K keyword retrieval)

In this case, the classification vocabulary is limited, the decoder's is not. Table 2 shows the top scoring systems when topic classifiers are built from retrieval results. The term-weighted value (TWV) scores, as defined by NIST for the 2006 Spoken Term detection evaluation[17], are roughly half of what was observed on the 2006 English eval, yet still sufficient to achieve ID performance near the transcript baseline. In the next section, in a keyword spotting configuration, the recognizer vocabulary will be limited as well.

5. KEYWORD SPOTTING VIA KALDI

We also construct a limited vocabulary keyword spotting system in the Kaldi framework. We restrict the vocabulary to our top 1000 keywords, plus one "garbage" word token comprised of 3 "garbage" phones. At training time, we simply replace non-keywords with the "garbage" token in all transcripts and proceed with the LVCSR recipe described in the previous section. We did not attempt to tune the HMM or acoustic model structure in any manner, as our goal was to measure the impact of an out-of-the-box system on the Topic ID task, not to optimize keyword spotting behavior.

Effectively, the non-garbage acoustic phonetic models are only exposed to speech from the keyword examples in the Call Home training data. All other acoustic training examples are mapped to the garbage phone. The effective decoding graph this produces is shown in Figure 3.

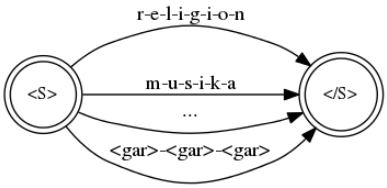

Fig. 3. Example decoding graph for keyword spotting

This decoding graph is typical in keyword spotting approaches, with what we call our "garbage" phone is similar to the "filler" model in the monophone keyword spotter from BUT [18]. However, we do not explicitly model or decode any triphones that occur outside the reduced vocabulary.

Except for reduced lexicon and decoding graph, we trained the acoustic models using the same recipe described in the previous section. This includes periodic forced aligning of the training data to the models. *We do not assume to have word-level alignments of our training examples, only an utterance-level segmentation of the training transcripts.*

5.1. Results

We present the keyword retrieval and Topic ID results in Table 3. In contrast to the full-vocabulary systems, the (weakest) keyword spotting models do not exhibit Topic ID performance on par with the transcript baseline, *except for the DNN-based models*. Rather than be disappointed by these results, we use this opportunity to look at the retrieval results and identify causes for the degradation.

The most noticeable difference between the DNN and PLP-based models is the increase in recall. On average, the *tri6_nnet* DNN models recalled nearly **twice as many** keyword instances as the other models. By contrast, the *tri5* PLP keyword spotters had the highest precision on the search task, but the lowest overall topic performance. A higher false alarm rate does not by itself inhibit Topic ID performance.

Based on the recall and precision of the keyword spotters, (top portion of Table 3) it is tempting to argue that recall by itself is sufficient for reasonable Topic ID performance. However, the ranked retrieval performance reveals something more nuanced.

Figure 4 shows the keyword spotting retrieval results in terms of the mean search AUC (MAUC) of all keywords plotted against Topic EER. The keyword spotting systems, even with DNN features, are at least 50% lower in terms of search accuracy than full-vocab LVCSR performance. The DNN system, however is twice as accurate in terms of ranked retrieval than all other keyword spotters.

Ranked retrieval metrics reflect the order of results. Higher AUC implies that correct keyword detections are more likely at the top of the term detection results. As we generate counts for Topic ID from the results, detections at the top contribute more to our bag-of-words model. We conjecture that for the DNN models, retrieval is good enough, given sufficient recall of topic-relevant words, that false alarms that obscure the topic signal do not occur high up in the result list.

Keyword Spotting Systems					
Model	EER	Recall	Prec.	MAUC	TWV
tri5	0.39	0.084	**0.641**	0.043	**-0.004**
sgmm5	0.27	0.206	0.512	0.080	-0.017
sgmm5_mmi	0.30	0.137	0.578	0.078	-0.007
tri6_nnet	**0.12**	**0.379**	0.338	**0.154**	-0.038
Full Vocabulary LVCSR					
tri5	0.09	0.278	0.464	0.395	0.342
sgmm5	0.08	0.309	0.478	**0.428**	**0.386**
sgmm5_mmi	0.08	**0.327**	**0.486**	0.427	0.381
tri6_nnet	**0.07**	0.269	0.458	0.433	0.384

Table 3. Naive Bayes Topic ID and retrieval performance. Paired t-test gives $p < 2 \times 10^{-16}$ between AM EER results.

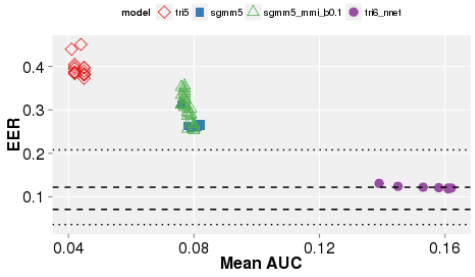

Fig. 4. EER versus keyword spotting metrics. Error bars on EER using human transcriptions are shown as dotted lines.

6. CONCLUSIONS

We have presented a number of limited resource approaches to Topic ID of spoken content. We have added to the body of evidence that Topic ID for speech is robust to high WER, and presented insight into causes of this robustness by measuring the Topic ID performance concurrently with keyword retrieval performance on the most topic-relevant words.

For the Spanish Fisher corpus, we achieved Topic ID performance rivaling classifiers based on manual transcripts by using either *limited training* LVCSR or *limited vocabulary* keyword retrieval. Even in the most limited keyword spotting configuration, DNN acoustic models spotted enough topic-relevant words to yield moderately good Topic ID.

For future work, we plan to examine if the relationship between keyword retrieval and Topic ID performance extends to zero-resource or unsupervised tokenization techniques, and further expand the set of viable alternatives to LVCSR.

7. REFERENCES

[1] YouTube, "Statistics - YouTube," http://www.youtube.com/yt/press/statistics.html, Aug 2013.

[2] Yiming Yang and Jan O. Pedersen, "A comparative study on feature selection in text categorization," 1997, pp. 412–420.

[3] Timothy J Hazen, Fred Richardson, and Anna Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 659–664.

[4] Barbara Peskin et al., "Improvements in Switchboard recognition and topic identification," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference-Volume 01*. IEEE Computer Society, 1996, pp. 303–306.

[5] Jonathan Wintrode and Scott Kulp, "Techniques for rapid and robust topic identification of conversational telephone speech," in *Proc. of Interspeech*, 2009.

[6] Guoguo Chen, Sanjeev Khudanpur, Daniel Povey, Jan Trmal, David Yarowsky, and Oguz Yilmaz, "Quantifying the value of pronunciation lexicons for keyword search in low resource languages," in *Acoustics, Speech and Signal Processing, 2013. ICASSP 2013. IEEE International Conference on*, 2013.

[7] Timothy J Hazen and Anna Margolis, "Discriminative feature weighting using mce training for topic identification of spoken audio recordings," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4965–4968.

[8] Man-hung Siu, Herbert Gish, Arthur Chan, William Belfield, and Steve Lowe, "Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, 2013.

[9] Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church, "Nlp on spoken documents without asr," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 460–470.

[10] David Graff et al., "Fisher Spanish - Transcripts," 2010.

[11] Jonathan Wintrode, "Leveraging locality for topic identification of conversational speech," in *INTERSPEECH*, 2013.

[12] Timothy J Hazen, "Topic identification," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 319–356, 2011.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[14] Barbara Wheatley, "Callhome Spanish Transcripts," 1996.

[15] Daniel Povey, Lukas Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek, Nagendra K Goel, Martin Karafiát, Ariya Rastrow, et al., "Subspace gaussian mixture models for speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4330–4333.

[16] Karel Veselỳ, Arnab Ghoshal, Lukáš Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," 2013.

[17] NIST, "The Spoken Term Detection (STD) 2006 Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf, 2006, [Online; accessed 28-Feb-2013].

[18] Igor Szoke, Petr Schwarz, Pavel Matejka, Lukas Burget, Martin Karafiát, Michal Fapso, and Jan Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Proc. of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.