Topic and Keyword Identification for Low-resourced Speech Using Cross-Language Transfer Learning

Wenda Chen[1,2], *Mark Hasegawa-Johnson*[1], *Nancy F. Chen*[2]

[1]Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2]Institute for Infocomm Research, A*STAR, Singapore

wchen113@illinois.edu, jhasegaw@illinois.edu, nfychen@i2r.a-star.edu.sg

Abstract

This paper studies topic and keyword identification for languages in which we have no transcribed speech data. We adopt a transfer learning framework to transfer what is learned from rich-resourced languages (RRL) to low-resourced languages (LRL). Specifically, we propose that a convolutional neural network (CNN) trained as a topic classifier in an RRL learns features (hidden layer activations) that can be used for the same purpose in an LRL. The CNN observes acoustic features, RRL phones, or segment clusters generated by an unsupervised phone clustering system; its hidden layers are retained, and its output layer re-trained from scratch on the LRL. Our results are compared with the state-of-the-art topic classification methods on cross-language ASR transcripts. We also discuss the successful detection of topic dependent keywords and the use of unsupervised learning based clusters in our approach for low-resourced language topic detection.

Index Terms: speech recognition, low-resourced languages, topic detection

1. Introduction

In the Ethnography and Field Linguistics study of endangered languages [1], a common method for eliciting complex sentences is to ask the Informant a series of questions about standard topics. The Linguist might, for example, ask the Informant to describe her favorite food as a child, to describe a typical day during her childhood, or to describe what she was doing when she first heard about a historical event that has been previously established to be of importance to the community. A free-form elicited corpus of this type results in speech that is not transcribed, but that is tagged for topic: the question asked by the Linguist can be treated as a marker of the topic of the Informant's response. Topic markers of this kind permit future students, historians, and linguists to access the data using structured search methods, even if the data are never completely transcribed. We propose, here, that topic labels of this type can be used to train a spoken language topic labeling system, even in a language for which no transcribed speech exists.

In most cases, there is too little topic-labeled data in the low-resourced language (LRL) to train a topic detector from scratch. Instead, we propose to train the topic detector first using speech in a rich-resourced language (RRL), like English or Mandarin. The words associated with each topic will, of course, not transfer very well from one language to the other, but the hidden layers of the neural net may encode features that can be transferred between languages with greater success: formant transitions, phones and phone sequences, syllables, perhaps even a few complete words that are shared in common between the two languages. We therefore propose to transfer knowledge from the RRL to the LRL by retaining the hidden layers, but re-training the output softmax layer from scratch.

Topic detection is a heavily studied problem, including methods specialized for both text [2] and speech [3] sources. Topic detection and tracking from speech is most accurately performed when one can first perform automatic speech recognition (ASR), then apply text-oriented topic detection methods such as latent Dirichlet allocation [2] or partial semantic parse [4]. It has been demonstrated that ASR-based topic detection outperforms methods without transcription, even when the ASR output has a relatively high error rate [5, 6]. In a language without transcribed speech, however, it may not be possible to train an ASR. When ASR is not available, methods used in the "Topic Detection and Tracking" competitions of the 1990s become relevant: methods that search for sequences or temporal patterns of phonemes [7, 8], frames [9], or parametric trajectory mixtures [10]. Recent studies of topic detection in the speech of under-resourced languages have revived the study of discriminatively extracted phonetic sequence information [11]. It has been demonstrated, for example, that topic ID can also be applied to the output of a phone recognizer constructed from self-organizing phone-like units learned in an unsupervised way from untranscribed speech[12, 13].

When there is no transcribed speech in an under-resourced language, a weaker form of transcription can be acquired using mismatched crowdsourcing [14, 15, 16]. In this approach, people who don't speak the LRL are asked to transcribe it as if it were a sequence of nonsense syllables. Their transcriptions are treated as a sort of noisy phone transcript, and can be used to train an automatic speech recognizer. We have previously demonstrated [17] that mismatched crowdsourcing is more useful when it is possible to acquire transcripts from more than one group of transcribers: even if neither group of transcribers understands the LRL, it is beneficial if the transcribers have distinct native languages, so that they are able to recognize different types of phonetic distinctions in the LRL. One of the technologies that becomes possible, in this situation, is a nullspace clustering approach [15] that permits us to infer the phone set of the LRL by observing the coincidence of different phonemes annotated by crowd workers with different native language backgrounds. In the work proposed in this paper we will not use mismatched transcripts, but instead, the nullspace clustering approach will be applied to the transcriptions generated by RRL phone recognizers in two different source languages (English and Mandarin).

Section 2 describes the features and model used for topic detection. Section 3 describes a multi-task learning framework that permits models to be optimized simultaneously for topic detection, and for the detection of keywords found to be salient for each topic, improving the topic detection accuracy. Section 4 describes experimental configuration and results, and

Section 5 concludes.

2. Topic Identification Models

This section describes the topic identification model from low-resourced speech data using clusters and phone level machine transcriptions. There are two categories of topic identification: classification and detection. This paper focuses on topic classification and keyword identification using only the phone level cross-language ASR results with no native transcriptions.

2.1. Observations

Topic detection algorithms in this work are convolutional neural networks (CNNs), trained to observe speech acoustic features, phones from a cross-language ASR trained on the RRL (rich-resourced language), or phonetic clusters generated by an unsupervised nullspace clustering algorithm [15] applied to RRL-phone transcripts generated by two different source-language speech recognizers.

The acoustic features used in this study are mel-frequency cepstral coefficients (MFCCs [18]). The feature-based CNN observes a sequence of MFCC vectors.

The RRL is English, and the RRL phone set is noted in ARPABET [19]. The RRL-phone CNN therefore observes sequences of ARPABET phones, for example, a sample utterance has the first input from the results of English Recognizer: "EH, K, IH, N, CH, AA, UW, B, EH, IH, K, AA, CH, AA, N, H, EH, IH, N, N, UW, L, AH."

The nullspace-clustering input is generated by analyzing RRL-phone transcripts of the LRL (Singapore Hokkien, in this paper) by ASRs trained to perform phone recognition in English and Mandarin. The ASR transcripts are clustered as described in [15], and phonetic labels are assigned to each cluster based on the set of phonologically distinctive features most frequently attested by transcriptions within each cluster. The resulting transcription contains a sequence of cluster labels, for example: /k,u,n,ts,o,b,i,t,ts,a,n,h,a:,n,j,l,a:/.

2.2. Model Training

Here we describe the transfer learning networks from rich-resourced language (RRL) corpus for low-resourced language (LRL) speech. The illustration is shown in Figure 1. The purpose is to classify the topics of the input speech and detect the keywords in the document. Here the keywords are defined as phone sequence patterns in the low resourced language that are linked with only one topic.

The transfer learning procedures are:

1. For both RRL and LRL speech, where each audio document has a topic tag, we generate features that will be used as input to the CNN: MFCC vectors, RRL-phones (ARPABET), or nullspace-clustered segment labels.
2. Documents in the training corpus are collected into subsets according to their topic labels. Then we collect all the phone sequence cluster patterns in the documents that occur more than once, with pattern length between 5-10 phones. Create a set of the phone sequence patterns for each topic.
3. Compare all the phone sequence patterns across the topics and delete the phone sequence patterns that exist for more than one topic. The remaining phone sequence patterns are called keywords from now on and each keyword is attached to one topic. Make a set of all keywords for both RRL and LRL.
4. Train the Convolutional Neural Network (CNN) neural network using audio waveforms in the RRL. Inputs to the CNN are MFCCs, RRL-phones, or clusters. There are 2 softmax layer outputs: topic class vector (length is number of topics) and keyword class vector (length is number of keywords in the full keywords set). For each input document, we have the corresponding topic and the set of keywords detected in the document as the labels.
5. Then replace the output layer to be the topic class vector and keyword class vector appropriate to the LRL. All weights in the network are retained except the output softmax weights, which are re-initialized to random values. Keep the input and middle layers of the network. Since the input format is the same in both LRL and RRL (either MFCCs, RRL-phones, or null-sequence clustered segments), we re-train the entire network (including both the softmax and all preceding layers) for the LRL.
6. In testing, given the input documents in LRL after using cross-language English phone recognizer, we can obtain the corresponding topic and keyword set in LRL that uniquely occur for the corresponding topic. These results are evaluated by the topic classification accuracy and keyword detection F1 score.

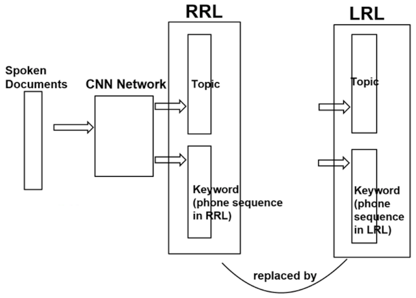
RRL
LRL
Spoken Documents
CNN Network
Topic
Keyword (phone sequence in RRL)
Topic
Keyword (phone sequence in LRL)
replaced by

Figure 1: *Transfer Learning step*

3. Topic Modeling and Multi-task Learning

A neural network can be trained to recognize only the topic of a document, or only its keywords. Training to recognize just one or the other of these outputs can achieve good accuracy on a rich-resourced language, but fails to transfer well from the RRL to the LRL. Far improved accuracy is obtained by training the CNN to detect both topic and keywords simultaneously, using methods borrowed from the field of multi-task learning.

Assumptions: for each document d, we have the topic t and keyword set k. Given the topic and keyword labels in target language, we can build the sequential modular neural network model using a multi-task learning framework [20, 16].

Training : Given d, t, and k, we seek to train the topic detection softmax weights λ, the keyword detection softmax weights θ, and their shared hidden layer weights W. The training objective (loss) $L(W, \lambda, \theta)$ is:

$$L = \alpha L_t(d, \lambda, W) + (1 - \alpha) L_k(d, \theta, W) \qquad (1)$$

where α is a convex combination weight that trades off the relative importance of topic detection versus keyword detection,

$L_t(d, \lambda, W)$ refers to the objective function for topic classification and $L_k(d, \theta, W)$ refers to the objective function for keyword detection.

At the output softmax layer for classification, we have

$$L_t(d, \lambda, W) = -\sum_t \sum_i \hat{y}_i(t) \log y_i(t; W, \lambda) \qquad (2)$$

where $y_i(t; W, \lambda) \in [0, 1]$ is the value of the i^{th} output of the softmax layer at time t, $\hat{y}_i(t) \in \{0, 1\}$ is the training label at time t and i is a target language topic/keyword index. The training objective consists a combination of topic identification and keyword detection with multi-task learning framework. The input document features are used to train a CNN network. The CNN applies a convolution filter W on the input features X before passing through the softmax output layer.

$$c_i = I(W^T X_{i:i+h-1} + b_v) \qquad (3)$$

$$y_i(t; W, \lambda) = \text{softmax}(\lambda max_i\{c_i\}) \qquad (4)$$

where I is the activation function, λ is the softmax weight matrix, and $y_i(t; W, \lambda)$ is the topic classification probability.

4. Experimental Settings and Results

Section 4.1 describes the speech corpus used in these experiments, and Section 4.2 describes the testing procedure. Section 4.3 gives the results of testing using acoustic features (MFCCs) and RRL-phones (ARPABET) as inputs to the CNN, for a truly low-resourced language (Singapore Hokkien) and a simulated low-resourced language (Spanish). Experimental results using the clustered mismatched transcripts generated using cross-language ASR systems are described in Section 4.4.

4.1. Corpus Description

Table 1 lists characteristics of the English, Spanish, and Singapore Hokkien databases used in this research. English is used in this study as the rich-resourced language (RRL). Singapore Hokkien is a low-resourced language spoken by about one million people in Singapore [21]; it has no native orthography, but phonetic orthographies have been developed for this language [22], and have been adapted for automatic speech recognition in previously published studies [23]. In order to have an LRL with more data than Singapore Hokkien, this study will also use Spanish as a simulated LRL.

Data in Singapore Hokkien consist of interviews between Mandarin-speaking Ethnologists and Hokkien-speaking Informants. Questions are asked in Mandarin, and are used as the topic labels for the Singapore Hokkien replies. There are 15 distinct topics, with a total of 396 question-answer pairs, with an average of 6 sentences per answer and 15 words per sentence.

Fisher English part 2 [24] has 5849 documents, each up to 10 minutes, and 50 topics. Informants were paired automatically, and were asked to discuss a topic chosen at random by the software. Switchboard-1 English [25] has 3638 5-minute telephone conversations involving 657 participants, and consists of approximately 260 hours of speech. About 70 topics were provided, of which about 50 were used frequently. Fisher Spanish consists of 819 telephone conversations of 10 to 12 minutes in duration from 136 speakers.

RRL-phones were generated for all three languages using an ASR trained in previously published work by other investigators [26]; experimental settings on Fisher and Switchboard English corpora are the same as in [6]. Unsupervised phone cluster strings were produced by clustering the ASR based mismatched transcripts, as described in [15, 14].

#docs	Training	Development	Test
Hokkien	320	38	38
Switchboard	3000	310	310
FisherEnglish	5000	424	424
FisherSpanish	650	70	70

Table 1: *Corpus Description*

4.2. Experimental Methods: Hokkien

The Singapore Hokkien corpus contains data in the format of Q&A where the interviewer is asking some oral history questions in Mandarin, and the interviewee is responding in Singapore Hokkien, as shown in Figure 2. All the response speech has been transliterated by human transcribers into representative words for Hokkien based on the pronunciation, using the transcription system described in [22], but these transliterations were not used for this research. Instead, we can define one topic for each answer response based on the question asked and find the keywords in each answer using the procedures described in Section 2.2. The total number of topics was 15.

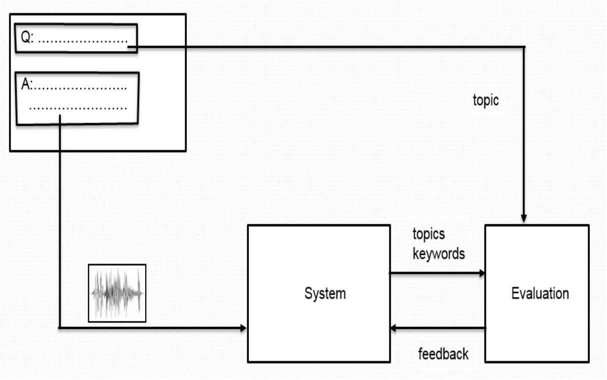

Figure 2: *Block diagram of the proposed system*

4.3. Topic Classification Results

For each document, we label it with the topic ID and keyword set. The hidden layers trained with English corpora (Fisher + Switchboard) are used as the initialization for the Spanish and Hokkien test corpora. In Tables 2-4, the results are presented with keyword F1 scores and topic classification scores.

The baseline system is a Support Vector Machine (SVM) classifier trained and tested as described in [6]. The SVM observes either MFCC (39 dimensions: MFCC+d+dd) or vectors showing the expected presence frequency of each English triphone. English triphone transcripts were generated by the ASR described in [26].

The proposed CNN outperforms an SVM baseline, even on the RRL English corpus (Table 2). We speculate that improvement on the English corpus results from one of two factors: (1) the CNN is able to learn about sequences longer than a triphone, and (2) the CNN is trained using a multi-task learning paradigm that tries to detect not only the topic, but also the set of keywords (phone N-grams for $5 \leq N \leq 10$) uniquely associated with each topic. Both the SVM and the CNN perform with greater accuracy using phone-sequence inputs instead of raw MFCC sequence inputs.

Tables 3 and 4 show topic detection results for Spanish and Singapore Hokkien. Because of the relatively small corpus sizes of both Spanish (650 documents) and Hokkien (320 documents), the SVM baseline performs rather poorly (54.8% accuracy on Spanish, 30.2% accuracy on Hokkien). The proposed transfer learning approach is able to outperform the baseline by a considerable margin. Keyword detection accuracy is not very high on the LRLs (28.4% for Spanish, 24.6% for Singapore Hokkien), but even though the keyword detector is learned with only limited success, it seems to be a useful secondary task for the multi-task training of the topic detector.

English	Proposed Model		SVM
Features	Keyword F1	Class. Acc.	Class. Acc.
MFCC	28.4%	67.3%	49.3%
Phone	37.5%	85.9%	82.1%

Table 2: *Results for English corpus*

Spanish	Proposed Model		SVM
Features	Keyword F1	Class. Acc.	Class. Acc.
MFCC	19.2%	43.1%	30.5%
Phone	28.4%	62.3%	54.8%

Table 3: *Results for Spanish corpus*

Hokkien	Proposed Model		SVM
Features	Keyword F1	Class. Acc.	Class. Acc.
MFCC	18.0%	25.4%	18.5%
Phone	24.6%	43.0%	30.2%

Table 4: *Results for Hokkien corpus*

4.4. Usage of Clusters

The clustering approach is an unsupervised learning method to obtain the target phone sequence labels for untranscribed speech in low-resourced languages [15, 14]. It uses the cross-language ASR results from 2 speech recognizers, usually in English and Mandarin, on any low-resourced speech. Then it iteratively clusters the frequently co-occurred pairs of English and Mandarin phones that are used to represent the same target phone. Then the clusters are optimized to agree with the target phones based on distinctive features. These clusters are used to convert the machine transcriptions into cluster sequence and then the target phone sequence to obtain the phone labels for the target language. Here we use these phone labels as input to the CNN system for topic classification. As observed in Table 5, we see that the clusters using the cross-language recognition results from two languages are consistently better in accuracy of the input, accuracy of the topic classification and resulting in better keywords detection score.

Language	Keyword F1	Class. Acc.
Spanish	30.9%	65.1%
Hokkien	27.2%	44.8%

Table 5: *Results using clusters as input features*

4.5. Discussion

This paper has described two key approaches for topic detection in an LRL. First, features learned on an RRL (hidden layers of the CNN) are transferred to the LRL, and then re-trained for topic detection on the LRL. Second, a multi-task learning framework is used to improve topic recognition results: the network is trained with two softmax outputs, one which classifies the topic of the utterance, and one which detects topic-specific keywords. The "keywords" are generated using an unsupervised clustering approach applied to RRL-phone transcripts of the audio segments in the training corpus: we identify any phone sequence in the LRL that occurs more than once for one topic, and that never occurs for any other topic.

Cross-language transfer learning of the hidden layers is effective for the LRL because (1) the phone sequence is generated using the same recognizer in both RRL and LRL — either an English phone recognizer, or a recognizer trained to detect phonetic clusters based on the clustering of English and Mandarin phone recognition transcripts, therefore (2) the longer-term features learned by the CNN represent frequent sequence patterns that may be discriminative in either language. Finally, (3) the final softmax layer is randomly re-initialized in the LRL, therefore the LRL is free to re-assign the detected hidden layer features to whatever topic in the LRL uses them most frequently and discriminatively.

5. Conclusions

This paper applies transfer learning to the topic identification problem of low-resourced languages where no native transcriptions are available. It also utilizes the speech acoustic features and unsupervised learning based transcriptions and clusters for the topic ID. Keywords (frequent and discriminative phone N-grams) are detected as a secondary task in a multi-task learning framework, improving the accuracy of the topic detection network in both the rich resourced source language and the low-resourced target language.

6. Acknowledgements

This work is funded by the Agency for Science, Technology and Research (A*STAR) Graduate Scholarship.

7. References

[1] W. J. Samarin, *Field Linguistics: A Guide to Linguistic Field Work*. New York: Holt, Rinehart and Winston, 1967.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Jan 2003.

[3] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," in *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Feb 1998.

[4] T. J. Hazen and F. Richardson, "Modeling multiword phrases with constrained phrase trees for improved topic modeling of conversational speech," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 222–227, 2012.

[5] M. Morchid, R. Dufour, and G. Linares, "A lda-based topic classification approach from highly imperfect automatic transcriptions," *LREC.*, pp. 1309–1314, 2014.

[6] T. J. Hazen, "Topic identification," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech (eds G. Tur and R. De Mori)*, 2011.

[7] R. Kuhn, P. Nowell, and C. Drouin, “Approaches to phoneme-based topic spotting: An experimental comparison,” *Proc. ICASSP*, Munich, April 1997.

[8] M. W. Theunissen, K. Scheffler, and J. A. du Preez, “Phoneme based topic spotting on the switchboard corpus,” *Proc. Eurospeech*, Aalborg, September 2001.

[9] E. Nöth, S. Harbeck, H. Niemann, and V. Warnke, “A frame and segment based approach for topic spotting,” *Proc. Eurospeech*, Rhodes, September 1997.

[10] W. Belfield and H. Gish, “A topic classification system based on parametric trajectory mixture models,” *Proc. Interspeech*, Geneva, September 2003.

[11] T. Hazen, M. Siu, H. Gish, S. Lowe, and A. Chan, “Topic modeling for spoken documents using only phonetic information,” *Speech Recognition and Understanding (ASRU)*, 2011.

[12] M.-H. Siu, H. Gish, A. Chan, and W. Belfield, “Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision,” *Proc. Interspeech*, Makuhari, September 2010.

[13] M.-H. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, “Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery,” *Comput. Speech Lang.*, vol. 28, no. 1, pp. 210–223, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2013.05.002

[14] W. Chen, M. Hasegawa-Johnson, and N. F. Chen, “Recognizing zero-resourced languages based on mismatched machine transcriptions,” *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[15] W. Chen, M. Hasegawa-Johnson, N. F. Chen, and B. P. Lim, “Mismatched crowdsourcing from multiple annotator languages for recognizing zero-resourced languages: A nullspace clustering approach,” *Proc. Interspeech*, 2017.

[16] M. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. Chen, P. Hager, T. Kekona, R. Sloan, and A. K. Lee, “ASR for under-resourced languages from probabilistic transcription,” *IEEE/ACM Trans. Audio, Speech and Language*, vol. 25(1), pp. 46–59, 2017.

[17] W. Chen, M. Hasegawa-Johnson, and N. F. Chen, “Mismatched crowdsourcing based language perception for under-resourced languages,” *Procedia Computer Science*, vol. 81, pp. 23–29, 2016.

[18] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[19] V. Zue, S. Seneff, and J. Glass, “Speech database development at MIT: TIMIT and beyond,” *Speech Communication*, vol. 9, pp. 351–356, 1990.

[20] V. H. Do, N. F. Chen, B. P. Lim, and M. Hasegawa-Johnson, “Multi-task Learning for Phone Recognition of Under-resourced Languages using Mismatched Transcription,” *IEEE/ACM Transactions on Audio, Speech and Language Processing, submitted*, 2017.

[21] A. Tien, “Chinese hokkien in singapore: evidence for an indigenised singapore culture,” *National University of Singapor*, August 2013.

[22] H. Y. Q. Amelia, “A phonological and phonetic description of singapore hokkien,” *B. A. thesis, Nanyang Technological University*, 2012.

[23] V. Lim, H. S. Ang, E. Lee, and B. P. Lim, “Towards an interactive voice agent for singapore hokkien,” *HAI ’16 Proceedings of the Fourth International Conference on Human Agent Interaction*, pp. 249–252, 2016.

[24] C. Cieri, D. Miller, and K. Walker, “The fisher corpus: a resource for the next generations of speech-to-text,” in *Proc. LREC*, May 2004.

[25] J. Godfrey, E. Holliman, and J. McDaniel, “SWITCHBOARD: telephone speech corpus for research and development,” in *Proc. ICASSP*, San Francisco, 1992, pp. 517–520.

[26] P. Schwarz, “Phoneme recognition based on long temporal context, phd thesis,” *Brno University of Technology*, 2009.