UNSUPERVISED DATA SELECTION AND WORD-MORPH MIXED LANGUAGE MODEL FOR TAMIL LOW-RESOURCE KEYWORD SEARCH

Chongjia Ni, Cheung-Chi Leung, Lei Wang, Nancy F. Chen and Bin Ma

Institute for Infocomm Research (I^2R), A*STAR, Singapore
{nicj, ccleung, wangl, nfychen, mabin}@i2r.a-star.edu.sg

ABSTRACT

This paper considers an unsupervised data selection problem for the training data of an acoustic model and the vocabulary coverage of a keyword search system in low-resource settings. We propose to use Gaussian component index based n-grams as acoustic features in a submodular function for unsupervised data selection. The submodular function provides a near-optimal solution in terms of the objective being optimized. Moreover, to further resolve the high out-of-vocabulary (OOV) rate for morphologically-rich languages like Tamil, word-morph mixed language modeling is also considered. Our experiments are conducted on the Tamil speech provided by the IAPRA Babel program for the 2014 NIST Open Keyword Search Evaluation (OpenKWS14). We show that the selection of data plays an important role to the word error rate of the speech recognition system and the actual term weighted value (ATWV) of the keyword search system. The 10 hours of speech selected from the full language pack (FLP) using the proposed algorithm provides a relative 23.2% and 20.7% ATWV improvement over two other data subsets, the 10-hour data from the limited language pack (LLP) defined by IARPA and the 10 hours of speech randomly selected from the FLP, respectively. The proposed algorithm also increases the vocabulary coverage, implicitly alleviating the OOV problem: The number of OOV search terms drops from 1,686 and 1,171 in the two baseline conditions to 972.

Index Terms— Submodular optimization, keyword spotting, spoken term detection, active learning

1. INTRODUCTION

Due to the ever-increasing availability of multimedia archives, we need solutions which can index and search the data more efficiently. However, searching for keywords in spoken documents is still a challenging problem as spoken documents are usually not transcribed. A state-of-the-art keyword search system usually requires a large vocabulary continuous speech recognition (LVCSR) system. With the LVCSR system, a lattice is generated for each utterance in the archive, and an inverted index is obtained from the lattices. Keyword search is performed by searching for a given keyword on the inverted index.

Although spoken material can be easily found online, it is expensive and time-consuming to transcribe the material and obtain a pronunciation dictionary for the training of a LVCSR system. The situation becomes more serious if the system is for a low-resource language.

This paper considers an unsupervised data selection problem in a low-resource language as follows: Given a set of untranscribed audio data, we select a subset of the data for manual transcription and creating a pronunciation dictionary due to the constraint of budget. The objective is that the selected data provides a good performance contribution to the LVCSR system and the corresponding keyword search system. We would study whether the selection of data plays an important role to the system performance: the word error rate (WER) of the LVCSR system and the actual term weighted value (ATWV) of the keyword search system.

This paper also studies the effect of unsupervised data selection on the vocabulary coverage of the dictionary, which alleviates the out-of-vocabulary (OOV) problem in keyword search. Handling out-of-vocabulary search terms [1,2,3] is another important issue to be considered in keyword search. Given a low-resource setting, the pronunciation dictionary created from the selected data subset possibly cannot well cover the search terms. Even worse, in Tamil, as many morphological rich languages, many search terms cannot be seen in the selected training data. To deal with this issue, as in [2,3,4], unsupervised morphological segmentation and word-morph mixed language modeling are considered in our experiments. The smoother estimates of word-morph mixed language models are proposed to address the data sparsity issues from the data set.

In this paper, we propose to use Gaussian component index based n-grams as acoustic features in a submodular function for unsupervised data selection. This data selection approach does not require an initial LVCSR system, and the submodular function can provide a near-optimal solution in terms of the objective being optimized. Our companion work [5] shows that the acoustic feature selection approach can perform as good as the phonetic feature (derived from an initial LVCSR system) selection approach. The result is obtained in another LVCSR application scenario, that an initial ASR system is available, and an additional subset of untranscribed data is selected for manual transcription and added to retrain the acoustic model.

Although the effect of the unsupervised data selection on the performance of a keyword search system, to the best of our knowledge, has never been studied, different kinds of active learning techniques have been studied to address the data selection problem for phone recognition or LVCSR. In the work based on submodular optimization, in additional to the work in [5], Lin *et al.* [8] proposed to use submodular function optimization on a Fisher-kernel based graph over untranscribed

utterances. This approach requires to compute the similarity between all utterance pairs. Wei *et al.* [9] proposed a two-level feature-based submodular function for selecting subsets of untranscribed data in the TIMIT corpus for training phone recognizers.

For supervised data selection based on submodular optimization, Wei *et al.* utilized phone n-grams as phonetic features [10] to select a subset from the transcribed training data to build an acoustic model. Wei *et al.* used the string kernel submodular function based on hypothesized phone labels to select a data subset to build a phone recognizer [11]. Shinohara [12] considered the phone distribution in the submodular function closed to a desired (uniform) distribution. Apart from the work based on submodular optimization, many earlier works on data selection focus on selecting either informative or representative utterances. The confidence-based data selection [15-18] is the commonly used data selection approach for selecting informative utterances. Siohan *et al.* [13, 14] proposed to use i-vector and context dependent tri-phone state sequence represent utterance, and then use relative-entropy data selection algorithm to select data. Wu *et al.* proposed to choose data uniformly according to the distribution of the target speech units [19]. Itoh *et al.* [20] suggested that both informativeness and representativeness of the data should be assessed at the same time. The fusion of informativeness and representativeness was used in data selection. However, these approaches cannot provide any optimality guarantee as those based on submodular optimization.

2. UNSUPERVISED DATA SELECTION FOR ACOUSTIC MODEL TRAINING

2.1. Background

Submodularity is a property of set-valued function. Consider a finite set $V = \{1,2,\cdots,n\}$ and a set-valued function $f: 2^V \to R$ that assigns each subset $S \subseteq V$ to a real number $f(S)$.

The function $f: 2^V \to R$ is submodular function if for every subset $A, B \subseteq V$ with $A \subseteq B$ and every item $s \in V \setminus B$,

$$f(B \cup \{s\}) - f(B) \le f(A \cup \{s\}) - f(A). \quad (1)$$

The submodular function f is monotone non-decreasing if $f(A \cup \{s\}) - f(A) \ge 0$ for every item $s \in V \setminus A, A \subseteq V$. The submodular function f is normalized if $f(\emptyset) = 0$.

A subset selection problem can be formulated as the following equation:

$$\max_{S \subseteq V} \{f(S) : c(S) \le K\} \quad (2)$$

where $c(S) \le K$ is the constraint.

The optimal problem is NP hard, and it can be approximately solved by using greedy forward selection algorithm [21]. The greedy algorithm can provide a good approximation to the optimal solution, and it is possible to be the best we can do in polynomial time [22].

2.2. Utterance representation

2.2.1 Utterances as Gaussian component index sequences

Gaussian mixture model (GMM) is widely used to capture the acoustic characteristics of utterances. In text-independent speaker recognition, GMM is used as a universal background model to capture the general speech characteristics of a population of speakers [23]. The model captures not only speaker variation but also environmental variation. Moreover, GMM-based tokenization has been around for quite some time in other speech processing applications [6,7,25-27]. In zero-resource speech processing, each Gaussian component in the model can represent a phoneme class sharing similar acoustic characteristics [25]. This motivates us to represent each untranscribed utterance using a sequence of Gaussian component indices. This method is suitable for the situation where no transcription or initial LVCSR system is available.

Algorithm 1. Gaussian component index based utterance representation

Step 1: Extract the spectral features (MFCC or PLP) from each utterance.

Step 2: Apply the Voice Activity Detection (VAD) technique in [28], and the voiced frames are retained.

Step 3: Train a GMM using the voiced frames.

Step 4: Represent the voiced portion of each utterance using the GMM, and output the index of Gaussian component with the maximum posterior probability along the frame sequence of the utterance.

2.2.2 Utterances in vector space representation

By applying the above Algorithm 1, each utterance is now presented by a Gaussian component index sequence which is in text-like format. As the text-like representations of the utterances usually have different lengths, vector space based techniques such as term frequency-inverse document frequency (tf-idf) can be used to represent each utterance in a fixed-dimensional vector.

As a statistical measurement used for evaluating how important a word is to a text document, tf-idf is widely used to solve information retrieval and text mining problems. The Algorithm 2 describes the procedure of how to model each utterance in the vector space.

Algorithm 2. Utterances in vector space representation

Step 1: Count the n-grams of Gaussian component indexes by the using Gaussian component index sequence representation of each utterance.

Step 2: Compute the tf vector for each utterance by using Gaussian component index based n-gram as term.

Step 3: Compute the idf vector by using all utterances for each Gaussian component index based n-gram.

Step 4: Compute $tf(term, s) \times idf(term)$ vector for each utterance.

By applying Algorithm 2, each utterance can be represented by a tf-idf vector.

2.3. Unsupervised development set distribution matching based submodular data selection

To select the utterances which may contribute to build a better acoustic model for the potential application, data distribution in the application domain is examined.

Suppose $P = \{p_u\}_{u \in U}$ to be the probability distribution over the feature set U, which is utilized to characterize the

application domain, and can be computed from a development set. The modular function $m_u(S)=\sum_{s\in S} m_u(s)$ indicates the importance of feature u in subset S, where $S\subseteq V, u\in U$ and $m_u(s)=tf(u,s)\times idf(u)$ which is proposed in the above section. The normalized function $\overline{m}_u(S)=\frac{m_u(S)}{\sum_{u\in U} m_u(S)}$ can be viewed as the probability of importance of u over the feature set U, and $M=\{\overline{m_u}(S)\}_{u\in U}$ denotes the probability of distribution

To compare the two distributions P and M, KL-divergence $D(P\|M)$ can be evaluated as

$$D(P\|M)=const.+\log\left(\sum_{u\in U} m_u(S)\right)-\sum_{u\in U} p_u \log(m_u(S)). \quad (3)$$

A set-valued function is then defined as

$$f(S)=\log\left(\sum_{u\in U} m_u(S)\right)-D(P\|M)=\sum_{u\in U} p_u \log(m_u(S)). \quad (4)$$

From submodular function optimization theory [29], Eq. (4) is a normalized and monotone non-decreasing submodular function. In this paper, this function is used for submodular based data selection.

3. WORD-MORPH MIXED LANGUAGE MODEL

Tamil is an agglutinative language with rich concatenation of suffixes onto words, and it is a subject-object-verb (SOV) language with free word order. The traditional word-based n-gram language model (LM) does not work well due to the huge number of different word forms. This problem becomes more serious in a low-resource condition, and this leads to the augmentation of out-of-vocabulary (OOV) words. In order to reduce the number of OOV words,morphological segmentation approaches are used to produce sub-word units.

There are rule-based morphological segmentation and data-driven based morphological segmentation [3-4,30-32]. When building a rule-based morphological segmenter, it needs corresponding linguistic knowledge. The unsupervised data-driven morphological segmentation algorithms, which discover the sub-word units from a text corpus, are commonly used [2-4,30]. These algorithms are language independent, andsimply assume that each word is composed of a number of subword units. The subwords generated by using an unsupervised morphological segmenter are not true morphemes in linguistics. In this paper, these subwords are referred to as "morphs".

In order to improve the performance of the keyword search system, the word-morph mixed language model is proposed to generate word-morph mixed lattices, and then these word-morph mixed lattices are used for word-morph hybrid keyword search. Algorithm 3 lists the building procedure in detail.

Algorithm 3. **Building a word-morph mixed language model**

Step 1: Train a language model only by using the word-based training transcriptions.

Step 2: Train a morphological model based on Morfessor [33] by using the word list, and generate the word-to-morph sequence map for each word.

Step 3: Replace each word in training transcriptions by its morph sequence, and use the text to train a morph-based language model.

Step 4: Replace the words whose frequencies are less than a threshold in training transcriptions with their morph sequences, and then use the text to train a language model.

Step 5: Interpolate the language models obtained by using Step 1, Step 3 and Step 4, and get the word-morph mixed language model.

By using the algorithm, different word morph contexts and thus a smoother word-morph mixed language model can be obtained. Before decoding using word-morph mixed language model, the phonetisaurus G2P toolkit [34] is used to acquire the pronunciation of each morph. The OOV keywords or keyword phrases in a keyword list are replaced by their corresponding morph sequences. The word-morph hybrid keywords or keyword phrases list is used in keyword search. The word-morph mixed language model is also used in our KWS system [35] and keyword-aware KWS system [36].

4. EXPERIMENTS

4.1 Experimental setup

The Tamil speech provided by the IAPRA Babel program for OpenKWS14 is used in our keyword search experiments. There are two kinds of packs. One is full language pack (FLP), which contains all data resources (including 60 hours of transcribed audio) provided by the program for training keyword search systems. Another is limited language pack (LLP), in which only a subset of 10 hours of audio in FLP is provided with transcription. The audio data is conversational telephone speech. The telephone channels include landlines, cell phones, and phones embedded in vehicles. In order to improve the phoneme coverage, the scripted speech is also recorded, which is only included in FLP.

To evaluate our proposed data selection algorithm, we build two baseline systems (denoted as LLP and FLP-10h-random respectively): (1) using the 10 hour transcribed data in LLP for LVCSR training; and (2) randomly selecting 10 hours of data from FLP for LVCSR training. For reference, we also build a system (denoted as FLP) with using all the 60 hour transcribed data in FLP for LVCSR training. When building all different systems, the lexicon only contains the words which occur in the training transcriptions. 10 hours of development set *Dev10h* and 15 hours of evaluation part 1 *Evalpart1* are used for evaluation. When evaluating the keyword search systems, a keyword list containing 5576 words or phrases is used. The performance of keyword search systems is measured by ATWV [37]. WER is used to measure the performance of the underlying LVCSR systems.

All LVCSR systems are hybrid DNN-HMM systems, which are built with the Kaldi toolkit [38]. The DNNs used for keyword search systems are trained by using the transcribed data with sMBR training [39]. The recipe of KWS in Kaldi is used to train our baseline systems. The LDA+MLLT+SAT transform is applied on the MFCC and fundamental frequency features in a context window. And the LDA+MLLT+SAT features are used to extract deep bottle-neck features. A LDA+MLLT+SAT transform is further applied on the concatenation of the LDA+MLLT+SAT features and the deep bottle-neck features. After making these transformations, the dimension of feature used in keyword search is 60. The lexicon

is provided by IARPA. The tri-gram language model used for decoding is trained by using the training transcriptions. In Algorithm 3, equal weight is used to interpolate different LMs.

4.2 Experimental results

Table 1. Performance of submodular data selection

	Data Set	FLP	LLP	FLP-10h-Random	FLP-10h-Proposed
WER(%)	*Dev10h*	64.4	75.4	79.4	73.3
	Evalpart1	66.1	77.0	80.7	74.4
ATWV	*Dev10h*	0.4349	0.2336	0.2380	0.2952
	Evalpart1	0.4222	0.2313	0.2362	0.2850

Table 1 lists the performance of different data selection algorithms. "FLP-10h-Proposed" is the system that we use our proposed algorithm to select 10 hours of data from FLP for LVCSR training. From Table 1, we can find that: (1) At the low-resource condition, WER is low for all systems; (2) The "FLP-10h-Proposed" system obtains better results than the other two systems using other sets of 10 hours of data. When comparing with the LLP system, there are 2.6% absolute WER reduction and 23.2% relative ATWV improvement on *Evalpart1*. When comparing with "FLP-10h-Random" system, there are 6.3% absolute WER reduction and 20.7% relative ATWV improvement on *Evalpart1*. (3) Although the LVCSR system trained using LLP is better than the system trained using FLP-10h-Random, the better performance in WER does not translate to that in ATWV.

Table 2. OOV statistics of keyword search systems

System	#OOV	Percentage
FLP	407	7.3%
LLP	1686	30.2%
FLP-10h-Random	1171	21%
FLP-10h-Proposed	972	17.4%

The number of OOV words or word phrases in the keyword list has great influence on the performance of the keyword search system in the low-resource condition. Table 2 lists the OOV statistics of different systems. There is 30.2% OOV in the keyword list of the LLP system. By random data selection, the OOV percentage drops from 30.2% to 21%. Although the FLP-10h-Random LVCSR system does not perform better than the LLP LVCSR system, the lower OOV percentage in FLP-10h-Random can contribute to the ATWV improvement, which can explain why the FLP-10h-Random system has better ATWV results when comparing with the LLP system.

Table 3. ATWV results of keyword searching systems

Data Set	LM	FLP	LLP	FLP-10h-Random	FLP-10h-Proposed
Dev 10h	Word-based	0.4349	0.2336	0.2380	0.2952
	Word-Morph Mixed	0.4475	0.2525	0.2478	0.2985
Evalpart1	Word-based	0.4222	0.2313	0.2362	0.2850
	Word-Morph Mixed	0.4363	0.2474	0.2386	0.2857

After using the morphological segmentation, the number of OOV words or word phrases in the keyword list is greatly reduced. There is no OOV word or word phrase in the FLP and FLP-10h-Proposed systems. There are only 7 and 6 OOV keywords or keyword phrases in the LLP and FLP-10h-Random systems respectively. Table 3 lists the keyword search results of different keyword search systems by using different types of LMs.

From Table 3, we can find that using the word-morph mixed LM in general provides performance gain to the keyword search systems. On *Evalpart1* data set, there are relative 0.2%~7% ATWV improvements. On *Dev10h* data set, there are relative 1.1%~8.1% ATWV improvements. However, the performance gain becomes insignificant when this LM is used together with our proposed data selection algorithm, and further investigation is needed.

5. CONCLUSIONS

This paper studies unsupervised data selection for the training data of an acoustic model and the vocabulary coverage of a keyword search system in a low-resource setting. We propose to use Gaussian component index based n-grams for acoustic features based submodular function optimization for data selection. As other feature based submodular functions, using this feature can avoid the computation of the similarity between all utterance pairs. Our proposed approach provides a promising improvement on the WER of the LVCSR system and the ATWV of the keyword search system. The more obvious improvement in ATWV may be attributed to the reduced number of OOV search terms in our proposed approach. The effect on the reduced number of OOV search terms has to be further investigated in the future. In zero-resource speech processing, frame-based Gaussian component posterior probabilities have been shown as an efficient data representation for detecting repeated spoken terms in two sequences of audio signals. However, we have to further verify whether Gaussian component index based n-grams can serve as a substitute for phonetic based features for the representativeness measures in terms of phonemes in data selection. And we have to study how the representativeness measures affect the words discovered.

6. REFERENCES

[1] S. Parlak and M. Saraclar, "Performance Analysis and Improvement of Turkish Broadcast News Retrieval," IEEE Trans. on Audio, Speech, and Language Processing. 2012,20(3):731-741.

[2] Y. He, B. Hutchinson, P. Baumann, M. Ostendorf, E. Fosler-Lussier, and J. Pierrehumbert, "Subword-based Modeling for Handling OOV Words in Keyword Spotting," in Proc. ICASSP 2014.

[3] K. Narasimhan, D. Karakos, R. Schwartz, S. Tsakalidis, and R. Barzilay, "Morphological Segmentation for Keyword Spotting," in Proc. EMNLP, 2014.

[4] M. Creutz, T. Hirsimaki, M. Kurimo, A. Puurula, J. Pylkkonen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke, "Morph-based Speech Recognition and Modeling of Out-of-vocabulary Words Across Languages," ACM Trans. on Speech and Language Processing, 2007,5(1):1-29.

[5] C. Ni, L. Wang, H. Liu, C.-C. Leung, L. Lu, and B. Ma. "Submodular Data Selection with Acoustic and Phonetic Features for Speech Recognition," in Proc. ICASSP 2015.

[6] P. A. Torres-Carrasquillo, E. Singer, M. K. R. J. Greene, D. Reynolds, and J. D. Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in Proc. ICSLP 2002, pp. 89 – 92.

[7] Boril, H., Zhang, Q., Angkititrakul, P., Hansen, J. H. L., Xu, D., Gilkerson, J., Richards, J. A., "A Preliminary Study of Child Vocalization on a Parallel Corpus of US and Shanghainese Toddlers," in Proc. Interspeech 2013, pp. 2405-2409.

[8] H. Lin and J. Bilmes, "How to Select a Good Training-data Subset for Transcription: Submodular Active Selection for Sequences," in Proc. Interspeech 2009, pp. 2859-2862.

[9] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Unsupervised Submodular Subset Selection for Speech Data," in Proc. ICASSP 2014, pp. 4107-4111.

[10] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels and J. Bilmes, "Submodular Subset Selection for Large-Scale Speech Training Data," in Proc. ICASSP 2014, pp. 3311- 3315.

[11] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Using Document Summarization Techniques for Speech Data Subset Selection," in Proc. NAACL/HLT-2013, pp. 721-726.

[12] Y. Shinohara, "A Submodular Optimization Approach to Sentence Set Selection," in Proc. ICASSP 2014, pp. 4140-4143.

[13] O. Siohan, and M. Bacchiani, "iVector-based Acoustic Data Selection," in Proc. Interspeech 2013, pp. 657-661.

[14] O. Siohan, "Training Data Selection Based on Context-Dependent State Matching," in Proc. ICASSP 2014, pp. 3316-3319.

[15] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active Learning for Automatic Speech Recognition," in Proc. ICASSP 2002, pp. 3904-3907.

[16] G. Riccardi and D. Hakkani-Tur, "Active and Unsupervised Learning for Automatic Speech Recognition," in Proc. Eurospeech 2003, pp. 1825-1828.

[17] G. Tur, R. E. Schapire, and D. Hakkani-Tur, "Active Learning for Spoken Language Understanding," in Proc. ICASSP 2003,pp. I-276-I-279.

[18] T. M. Kamm and G. G. L. Meyer, "Selective Sampling of Training Data for Speech Recognition," in Proc. Human Language Technology Conf., San Diego, CA, 2002.

[19] Y. Wu, R. Zhang, and A. Rudnicky, "Data Selection for Speech Recognition," in Proc. ASRU 2007, pp. 562-565.

[20] N. Itoh, T. N. Sainath, D. N. Jiang, J. Zhou, and B. Ramabhadran, "N-Best Entropy Based Data Selection for Acoustic Modeling," in Proc. ICASSP 2012, pp. 4133-4136.

[21] G. Nemhauser, L. Wolsey, and M. Fisher, "An Analysis of Approximations for Maximizing Submodular Set Function-I," Mathematical Programming,1978, 14(1):265-294.

[22] U. Feige, "A Threshold of ln n for Approximating Set Cover," Journal of the ACM (JACM), 1998,45(4):634-652.

[23] T. Kinnunen, and Haizhou Li, "An Overview of Text-independent Speaker Recognition: From Features to Supervectors," Speech Communication, 2010:52(1):12-40.

[24] Y. Zhang, anf J. R. Class, "Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams," in Proc. ASRU 2009.

[25] H. Wang, T. Lee, C.-C. Leung, B. Ma and H. Li, "Unsupervised Mining of Acoustic Subword Units With Segment-level Gaussian Posteriorgrams," in Proc. Interspeech 2013, pp.2297-2301.

[26] H. Wang, T. Lee, C.-C. Leung, B. Ma and H. Li, "Acoustic Segment Modeling with Spectral Clustering Methods," IEEE Trans. On Audio, Speech and Language Processing, 2015,23(2):264-277.

[27] L. Zhang, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Acoustic TextTiling for Story Segmentation of Spoken Documents," in Proc. ICASSP 2012, pp.5121-5124.

[28] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust Voice Activity Dection Using Long-Term Signal Variability," IEEE Trans. on Audio, Speech and Language Processing, 2011,19(3):600-613.

[29] F. Bach, "Learning with Submodular Functions: A Convex Optimization Perspective," Foundations and Trends ® in Machine Learning," 2013,6(2-3): 145-373.

[30] M. J. J. Premkumar, N. T. Vu, and T. Schultz, "Experiments towards a better LVCSR System for Tamil," in Proc. Interspeech 2013.

[31] T. Ruokolainen, O. Kohonen, S. Virpioja, and M. Kurimo, "Supervised Morphological Segmentation in a Low-Resource Learning Setting using Conditional Random Fields," in Proc. of the Seventeenth Conference on Computational Natural Language Learning, pp.29-37.

[32] M.Selvam, and A. M. Natarajan, "Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques," International Journal of Computers, 2009, 3(4):357-367.

[33] P. Smit, S. Virpioja, S.-A. Gronroos, and M. Kurimo, "Morfessor 2.0: Toolkit for statistical morphological segmentation," in Proc. ECACL 2014, pp.21-24.

[34] phonetisaurus - A WFST-driven Phoneticizer, Available On-line: https://code.google.com/p/phonetisaurus/.

[35] N. F. Chen, C. Ni, I-Fan Chen, S. Sivadas, V. T. Pham, H. Xu, X. Xiao, T. S. Lau, S. Leow, B. P. Lim, C.-C. Leung, L. Wang, C.-H. Lee, A. Goh, E. S. Chng, B. Ma, H. Li, "Low-resource Keyword Search Strategies for Tamil," in Proc. ICASSP 2015.

[36] I-F. Chen, C. Ni, B. P. Lim, N. F. Chen, C.-H. Lee, "A Keyword-Aware Grammar Framework for LVCSR-based Spoken Keyword Search," in Proc. ICASSP 2015.

[37] J. G. Fiscus, J.Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 Spoken Term Detection Evaluation," in Proc. Interspeech 2007.

[38] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

[39] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative traning of deep neural networks," in Proc. Interspeech 2013.