EFFECTIVE KEYWORD SEARCH FOR LOW-RESOURCED CONVERSATIONAL SPEECH

Rasa Lileikytė[⋆], *Thiago Fraga-Silva*[†], *Lori Lamel*[⋆], *Jean-Luc Gauvain*[⋆]
Antoine Laurent[†], *Guangpu Huang*[⋆]

[⋆]LIMSI, CNRS, Université Paris–Saclay, 508 Campus Universitaire F–91405 Orsay, France
[†]Vocapia Research, 28 rue Jean Rostand, 91400 Orsay, France
[⋆]{lileikyte, lamel, gauvain, huang}@limsi.fr, [†]{thfraga, laurent}vocapia.fr

ABSTRACT

In this paper we aim to enhance keyword search for conversational telephone speech under low-resourced conditions. Two techniques to improve the detection of out-of-vocabulary keywords are assessed in this study: using extra text resources to augment the lexicon and language model, and via subword units for keyword search. Two approaches for data augmentation are explored to extend the limited amount of transcribed conversational speech: using conversational-like Web data and texts generated by recurrent neural networks. Contrastive comparisons of subword-based systems are performed to evaluate the benefits of multiple subword decodings and single decoding. Keyword search results are reported for all the techniques, but only some improve performance. Results are reported for the Mongolian and Igbo languages using data from the 2016 Babel program.

Index Terms— Speech recognition, keyword search, text augmentation, language modeling, low-resourced languages

1. INTRODUCTION

Today's speech recognition systems make use of statistical acoustic and language models (LMs) which are trained on large data sets. System performance generally improves with increasing training data. Low-resourced languages are considered those with a low availability on the Internet, and usually have limited text resources, with little or no available transcribed audio or pronunciation dictionaries.

Training language models under low-resourced conditions is a challenge. Web data are frequently used to improve language model for broadcast news, as in [1, 2]. Conversational speech has specific syntactic and semantic nature that is significantly different from written language. There are little conversational-like Web texts for low-resourced languages. Web data usage for low-resourced languages was investigated in [3, 4, 5]. Alternatively, texts generated with recurrent neural networks (RNNs) demonstrated gains in [6, 7].

In the keyword search (KWS) task the out-of-vocabulary (OOV) keywords usually are poorly detected and degrade keyword search performance. Various methods have been proposed to address this problem. One approach is converting word lattices to phoneme lattices and performing phoneme based search [8, 9]. Some studies [10, 11] propose using lattices of subword units. The proxy approach is used in [12], where keyword search allows matches to vocabulary words which are phonetically similar to the specified keyword. KWS performance improvement using joint decoding is investigated in [13], using multiple system combination in [14], and multilingual acoustic models in [15, 16].

This paper explores two techniques to improve a keyword search system for low-resourced conversational speech, with the aim of increasing the detection of OOV keywords. *1) Extra text resources are assessed to augment language model and lexicon*. Documents collected from the Web are used for language model training. These texts were gathered by submitting conversational-like queries to a search engine in order to reach conversational-like data [4]. Additionally, texts generated by RNNs are explored. *2) Different approaches for the use of subwords are explored to determine the impact on keyword search*. First, two ways of subword decoding are investigated: multiple decoding where each character n-gram subword set is decoded separately and then keyword hits are combined; and single decoding when different n-gram subword texts are concatenated. We also investigate the impact of n-gram subwords size, various sets of concatenated subword texts, and concatenated subword texts with the word texts.

2. DATA

All the experiments reported in this paper use data provided by the IARPA-Babel program [17] for Mongolian and Igbo.

Mongolian (IARPA-babel401b-v2.0b), more specifically Halh Mongolian is a Mongolic language spoken in Mongolia by approximately 3 million speakers. The official standard spelling uses Mongolian Cyrillic. Igbo (IARPA-babel306b-v2.0c) is a Niger-Congo language (Volta-Niger) spoken in south-eastern Nigeria by about 25 million people. It is based on Latin alphabet with additional dotted characters.

The data are comprised of spontaneous telephone conversations, with about 40 hours of manually transcribed training data. About a 85 million and 120 million word text corpus was collected from the Web for Mongolian and Igbo respec-

Table 1. Mongolian results using various texts for LM training: manual transcriptions (*trs*); Web data (*web*); RNN generated text (*rnn*). For KWS word units are used.

Vocab	LM	OOV	WER	MTWV			
		%	%	All	INV-INV	OOV-INV	OOV-OOV
23k	trs (baseline)	4.3	48.1	0.460	0.516	-	0.138
23k	trs+rnn	4.3	47.9	0.461	0.516	-	0.142
100k	trs+web	1.9	47.0	0.505	0.529	**0.470**	0.266
100k	trs+web+rnn	1.9	46.8	0.504	0.529	0.456	0.267
700k	trs+web	0.9	47.1	0.503	0.522	0.435	0.309
700k	trs+web+rnn	0.9	46.7	0.500	0.523	0.386	**0.325**

Table 2. Igbo results using various texts for LM training: manual transcriptions (*trs*); Web data (*web*); RNN generated text (*rnn*). For KWS word units are used.

Vocab	LM	OOV	WER	MTWV			
		%	%	All	INV-INV	OOV-INV	OOV-OOV
17k	trs (baseline)	2.4	54.7	0.326	0.343	-	0.288
17k	trs+rnn	2.4	54.4	0.330	0.350	-	0.277
100k	trs+web	1.8	54.9	0.330	0.340	0.240	0.308
100k	trs+web+rnn	1.8	54.9	0.329	0.340	0.242	0.306
700k	trs+web	1.3	55.2	0.330	0.341	0.233	**0.323**
700k	trs+web+rnn	1.3	55.0	0.333	0.346	**0.245**	0.321

tively. The Web data was collected (Wikipedia, subtitles and other sources), filtered by BBN and shared with the Babelon participants [4]. Additional 120 million word text corpus for Mongolian and 90 million for Igbo was generated using RNNs [18] by LIMSI and provided to the Babelon team.

All results are reported on the official Babel 10 hour development data set. For the keyword search experiments, the official 2016 year list of development keywords provided by NIST was used. The Mongolian development keyword list contains 2404 keywords, and Igbo contains 2364 keywords. Based on the vocabulary of their respective transcriptions, for Mongolian there are 358 and for Igbo 601 OOV keywords. A keyword may be a single word or a sequence of words. If any word in the keyword list is out-of-vocabulary then the keyword is considered OOV with respect to the system's vocabulary. The remaining keywords are in-vocabulary (INV).

3. SYSTEM OVERVIEW

3.1. Speech-to-text system

In our experiments the speech-to-text (STT) systems are built via a flat start training, where the initial segmentation is performed without any a priori information. It uses left-to-right 3-state hidden Markov models (HMMs) with Gaussian mixture observation densities, in total about 10k tied states with about 15 components per state [19]. Next, deep-neural network (DNN) is used to estimate the HMM state likelihoods replacing the GMMs [20]. The 6-layer DNN models have about 10M parameters, and the softmax output layer targets HMM states. Word position dependent and word position independent acoustic models are used in the word- and subword-based systems respectively. They are trained on multilingual stacked bottleneck features provided to the Babelon team by BUT [21].

Back-off trigram LMs with Kneser-Ney smoothing were trained using the LIMSI STK toolkit. The vocabularies consist of all words from the training transcriptions and the most likely words from Web texts. The experiments use phonemic pronunciation lexicons, where the grapheme-to-phoneme mappings were provided by NWU to the Babelon members (similar as in [22]). Mongolian is represented with 29 units and Igbo with 32, along with 4 units for silence and fillers.

For each speech segment a word lattice is generated, the final hypotheses are then obtained using consensus decoding [23]. The speech-to-text system performance is measured with the commonly used word error rate (WER) metric.

3.2. Keyword search system

For the keyword search we use the methods proposed in [10], with a focus on OOV keywords performance improvement. A word and a subword consensus networks are generated from decoding lattices. Both consensus networks are searched to locate all sequences of words and subwords that correspond to each keyword. Keyword search is carried out with cross-word search, ignoring word boundaries, and splitting words in keyword term. Resulting word and subword based keyword hits are combined. The keyword scores are normalized using keyword-specific thresholding and exponential normalization [24]. From 3 to 7 character n-grams (or letters) cross-word subword units are used.

Keyword search results are reported in terms of the maximum term-weighted value (MTWV) [25]. To observe the

Table 3. KWS performance reported when different character n-gram subword sets (from 3-gram to 7-gram) are decoded separately, then resulting keyword hits are combined. Various texts for subword LM are used: manual transcriptions (*trs*), RNN generated text (*rnn*). MTWV results reported using subword units.

LM		MTWV Mongolian			MTWV Igbo		
subw	source	All	INV	OOV	All	INV	OOV
3-gram	trs	0.380	0.394	0.302	0.248	0.244	0.271
4-gram	trs	0.393	0.405	0.329	0.258	0.251	**0.280**
5-gram	trs	0.406	0.420	0.336	0.260	0.255	0.279
6-gram	trs	0.418	0.434	0.334	0.253	0.246	0.275
7-gram	trs	0.425	0.439	**0.350**	0.263	0.259	0.277
3-gram	rnn	0.363	0.375	0.295	0.248	0.237	**0.285**
4-gram	rnn	0.372	0.382	0.320	0.251	0.243	0.283
5-gram	rnn	0.378	0.386	**0.334**	0.245	0.234	0.278
6-gram	rnn	0.401	0.413	0.330	0.249	0.244	0.268
7-gram	rnn	0.405	0.418	0.329	0.245	0.241	0.261
5-way combine	trs	0.441	0.448	**0.408**	0.257	0.250	0.285
5-way combine	rnn	0.417	0.424	0.384	0.252	0.241	0.287
10-way combine	trs+rnn	0.418	0.424	0.396	0.238	0.227	**0.292**

Table 4. KWS performance reported when texts with different character n-gram subwords (from 3-gram to 7-gram, denoted as (*3to7*)) and/or (*word*) units are concatenated and then single decoding is performed. Subword units are based on: manual transcriptions (*subw-trs*), RNN generated text (*subw-rnn*). MTWV results reported using subword units.

LM		MTWV Mongolian			MTWV Igbo		
subw-trs	subw-rnn	All	INV	OOV	All	INV	OOV
3to7	-	0.416	0.434	0.312	0.275	0.273	0.282
3to7+word	-	0.419	0.446	0.263	0.297	0.300	0.291
-	3to7	0.407	0.424	0.316	0.259	0.249	0.295
-	3to7+word	0.428	0.443	**0.341**	0.144	0.152	0.120
3to7	3to7	0.422	0.440	0.322	0.273	0.269	0.286
3to7+word	3to7	0.442	0.464	0.322	0.293	0.297	0.288
3to7	3to7+word	0.423	0.444	0.304	0.278	0.273	**0.296**
3to7+word	3to7+word	0.323	0.353	0.149	0.278	0.284	0.265

impact of augmented text on keyword search, performance is reported for different keywords: INV-INV, OOV-INV, OOV-OOV. When words from augmented texts are added to the lexicon, some originally OOV words become INV (OOV-INV), while others remain OOV (OOV-OOV). The INV keywords are considered with respect to the original lexicon (INV-INV).

4. DATA AUGMENTATION FOR STT & KWS

Data augmentation using the Web texts provided by BBN was assessed to augment the lexicon and language model. Some low-resourced languages may have little or no text resources on the Web. Rather than only using Web data, we also introduced additional texts generated with RNNs [18] based on training transcripts. RNN has 2 hidden layers and 512 neurons per layer. Training transcripts were randomly shuffled and split into five non-overlapping subsets. For each split, an RNN was trained using four sets and reserving the fifth set for validation. The RNN keeps the same vocabulary and does not address the OOV detection problem significantly.

Results obtained with data augmentation are shown in Tables 1 and 2 for Mongolian and Igbo, respectively. For Mongolian a 100k lexicon was selected using both the Web data and RNN generated data. With this lexicon the OOV rate is reduced in half and a 1.3% absolute WER improvement is obtained over the baseline (48.1% vs 46.8%). RNN texts helped to improve system by 0.2% on top of Web texts. The WER remains almost the same even if the lexicon is increased to 700k. For Igbo, using the RNN texts with 100k lexicon leads to 0.3% WER absolute reduction compared to the baseline (54.7% vs 54.4%). Web data did not bring WER improvement. For Igbo, LMs trained on transcriptions and RNN texts are more accurate than LMs including Web data. This may be in part due to the large amount of English in the Web texts even after filtering.

Full-word based KWS results are also given for Mongolian and Igbo. Adding texts improves the overall KWS performance, with the largest gains from the better lexical coverage (OOV-INV). Without ignoring word boundaries and splitting words in KWS, OOV-OOV drops by 0.1 to 0.2 absolute.

Table 5. KWS performance combining the best full-word and subword systems.

Lang	MTWV			
	All	INV-INV	OOV-INV	OOV-OOV
Mongolian	0.515	0.529	0.470	0.486
Igbo	0.332	0.346	0.245	0.323

5. IMPROVING KEYWORD SEARCH

In this section we apply subword search technique with the aim of improving the detection of OOV keywords. First we explore the impact of n-gram size, then we compare the results of multiple decodings and single decoding.

Keyword search results using multiple decoding when each n-gram subword set is decoded separately are given in Table 3. For Mongolian the 7-gram transcript-based system improves the MTWV for OOVs by 0.21 compared to the baseline. For Igbo 3-gram RNN-based system shows the highest OOV result among the separately decoded subword sets, but with no improvement compared to the baseline. Subword units reduce the OOV rate but generate false combinations. Since words are longer in the Mongolian language, character 7-gram subwords are more beneficial, however, Igbo results are better with 3-grams due to the shorter words of this language. The differences across the RNN and training transcript-based subwords are not significant for both languages. For Mongolian the combination of 5 decoding transcript-based subword systems leads to OOV improvement of 0.27 MTWV absolute over the baseline (0.138 vs 0.408). As Igbo took advantage from RNN usage, combining all 10 subword systems including both RNN and transcript-based subword outputs gives the best OOV result, but with only a tiny improvement over the baseline (0.288 vs 0.292).

Multiple decoding with different n-gram subword sets and then combining keyword hits is an expensive process. Table 4 presents the results of a single decoding when texts from 3 to 7-gram subword sets and/or full-words are concatenated. LMs are interpolated with 0.8 coefficients for transcript-based subwords, and 0.2 for RNN subwords. Concatenated subword texts lead to a high OOV detection for both languages (entries with *3to7*). When full-word texts are concatenated along with the subword texts, INV detection is higher, but OOV performance degrades in the some cases. For Mongolian RNN subwords plus full-words improve OOV by 0.2 MTWV absolute compared to the baseline (0.138 vs 0.341). For Igbo interpolating LMs of transcript-based subwords and RNN-based subwords plus full-words, shows the best OOV result with a tiny gain of 0.01 absolute (0.288 vs 0.296).

Comparing OOV best results of multiple and single decodings (Table 3, Table 4), the performance with the latter is less good for Mongolian, and slightly better for Igbo. Table 5 presents the results when keyword hits of the best full-word and subword systems are combined. The final combination leads OOV-OOV to 0.35 absolute gain over the baseline for Mongolian, and to 0.04 absolute for Igbo.

6. CONCLUSIONS

In this paper we explored two techniques aiming to improve keyword search performance for low-resourced conversational speech, with a focus on OOV keywords. The experiments were performed for Mongolian and Igbo.

The first technique improves lexical coverage and language model by augmenting training texts: using Web data and via texts generated by RNNs. For Mongolian, extra Web resources obtain WER absolute gain of 1.3%, but no gains are obtained for Igbo, which may be due to the large number of English words in the Web texts. RNN generated texts lead to WER improvements for both languages: 0.2% absolute gain for Mongolian, and 0.3% absolute for Igbo. Using word-based units the KWS performance is improved with a large gain for OOV-INV from the better lexical coverage. Ignoring word boundaries and splitting words in keyword search improves the OOV-OOV detection. Word-based RNNs keep the same vocabulary and do not affect OOV-OOV significantly.

A single decoding with a language model estimated on all subword n-gram texts concatenated, results in a modest performance loss as carrying out multiple decodings with different n-gram subword sets followed by merging keyword hits and is much less costly in terms of computation. This was observed with the RNN-based and transcription-based subwords. For Mongolian, subwords improve OOV-OOV detection by 0.27 MTWV absolute over the baseline, and for Igbo almost no improvement is observed.

The largest gains are obtained when outputs of the best full-word system and subword system are combined. The proposed techniques lead to significant OOV-OOV improvement by 0.35 MTWV absolute comparing to the baseline for Mongolian, and by tiny 0.04 absolute for Igbo.

7. ACKNOWLEDGMENTS

We would like to thank our IARPA-Babel partners for sharing resources (BUT for the bottleneck features, BBN for the Web data, and NWU for the grapheme to phoneme mappings).

This research was in part supported by the French National Agency for Research as part of the SALSA (Speech And Language technologies for Security Applications) project under grant ANR-14-CE28-0021, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

8. REFERENCES

[1] T. Schlippe, L. Gren, N. T. Vu, and T. Schultz, "Unsupervised language model adaptation for automatic speech recognition of broadcast news using web 2.0," in *Interspeech*, 2013, pp. 2698–2702.

[2] R. Lileikytė, A. Gorin, L. Lamel, J. L. Gauvain, and T. Fraga-Silva, "Lithuanian broadcast speech transcription using semi-supervised acoustic model training," *SLTU*, vol. 81, pp. 107–113, 2016.

[3] G. Mendels, E. Cooper, V. Soto, J. Hirschberg, M. Gales, K. Knill, A. Ragni, and H. Wang, "Improving speech recognition and keyword search for low resource languages using web data," in *Interspeech*, 2015, pp. 829–833.

[4] L. Zhang, D. Karakos, W. Hartmann, R. Hsiao, R. Schwartz, and S. Tsakalidis, "Enhancing low resource keyword spotting with automatically retrieved web documents," in *Interspeech*, 2015, pp. 839–843.

[5] R. Lileikyte, L. Lamel, and J. L. Gauvain, "Conversational telephone speech recognition for Lithuanian," in *SLSP*, 2015, pp. 164–172.

[6] G. Huang, A. Gorin, J. L. Gauvain, and L. Lamel, "Machine translation based data augmentation for Cantonese keyword spotting," in *ICASSP*, 2016, pp. 6020–6024.

[7] A. Gorin, R. Lileikyte, G. Huang, L. Lamel, J. L. Gauvain, and A. Laurent, "Language model data augmentation for keyword spotting in low-resourced training conditions," in *Interspeech*, 2016, pp. 775–780.

[8] O. Siohan and M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing," in *Interspeech*, 2005, pp. 52–55.

[9] D. Karakos, I. Bulyko, R. Schwartz, S. Tsakalidis, L. Nguyen, and J. Makhoul, "Normalization of phonetic keyword search scores," in *ICASSP*, 2014, pp. 7834–7838.

[10] W. Hartmann, V. B. Le, A. Messaoudi, L. Lamel, and J. L. Gauvain, "Comparing decoding strategies for subword-based keyword spotting in low-resourced languages," in *Interspeech*, 2014, pp. 2764–2768.

[11] 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 *ICASSP*, 2014, pp. 7864–7868.

[12] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," in *ASRU*, 2013, pp. 416–421.

[13] H. Wang, A. Ragni, M. J. F. Gales, K. M. Knill, P.C. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *Interspeech*, 2015, pp. 3660–3664.

[14] W. Hartmann, L. Zhang, K. Barnes, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Comparison of multiple system combination techniques for keyword spotting," in *Interspeech*, 2016, pp. 1913–1917.

[15] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, et al., "Multilingual representations for low resource speech recognition and keyword search," in *ASRU*, 2015, pp. 259–266.

[16] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Multilingual features based keyword search for very low-resource languages," in *Interspeech*, 2015, pp. 1260–1264.

[17] M. Harper, "The BABEL program and low resource speech technology," in *ASRU*, 2013.

[18] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *SLT*, 2012, pp. 234–239.

[19] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech communication*, vol. 37, no. 1, pp. 89–108, 2002.

[20] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[21] F. Grézl and M. Karafiát, "Bottle-neck feature extraction structures for multilingual training and porting," in *SLTU*, 2016, vol. 81, pp. 144–151.

[22] M. Davel, D. Karakos, E. Barnard, C. V. Heerden, R. Schwartz, and S. Tsakalidis, "Exploring minimal pronunciation modeling for low resource languages," *Interspeech*, pp. 538–542, 2015.

[23] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[24] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, R. Hsiao, G. Saikumar, I. Bulyko, L. Nguyen, et al., "Score normalization and system combination for improved keyword spotting," in *ASRU*, 2013, pp. 210–215.

[25] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *SIGIR*, 2007, vol. 7, pp. 51–57.