SMT-Based Lexicon Expansion for Broadcast Transcription

Manon Ichiki*, Aiko Hagiwara*, Hitoshi Ito*, Kazuo Onoe†, Shoei Sato* and Akio Kobayashi‡
* NHK Science and Technology Research Laboratories, Tokyo, Japan
E-mail:{ichiki.m-fq,hagiwara.a-iy,itou.h-ce,satou.s-gu}@nhk.or.jp
† NHK Muroran Broadcasting Station, Hokkaido, Japan
E-mail:onoe.k-ec@nhk.or.jp
‡ NHK Engineering System, Tokyo, Japan
E-mail:kobayashi.a-fs@nhk.or.jp

***Abstract*—We describe a method of lexicon expansion to tackle variations of spontaneous speech. The variations of utterances are found widely in the programs such as conversations talk shows and are typically observed as unintelligible utterances with a high speech-rate. Unlike read speech in news programs, these variations often severely degrade automatic speech recognition (ASR) performance. Then, these variations are considered as new versions of original entries in the ASR lexicon. The new entries are generated based on the SMT approach, in which translation models are trained from corpus translating phoneme sequence in a lexicon into the sequence obtained by phoneme recognition. We introduce a new method in which unreliable entries are removed from the lexicon. Our SMT-based approach achieved a 0.1 % WER reduction for a variety of broadcasting programs.**

I. Introduction

NHK (Japan Broadcasting Corp.) has studied closed-captioning based on automatic speech recognition (ASR) to resolve accessibility issues for hard-of-hearing people and launched a closed-captioning system for live news shows in 2000 [1], [2]. While the system achieves high ASR performance when decoding read and clean speech in news programs, its accuracy is drastically degraded when decoding spontaneous speech in infotainment programs and talk shows. To overcome such drawbacks, we use a "re-speak" method. The system decodes "rephrased" clean speech uttered by other speakers. Since this way is not economical compared with captioning through decoding speech "directly", the ASR system that covers whole program hours is demanded.

One of the major challenges in decoding spontaneous speech is robustness against variability in utterances. Typically, in talk shows, speech is less articulate than read speech due to its high rate of speech and the existence of filled pauses. In particular, a high rate of speech can easily cause deletions of phonemes that should appear in the original pronunciations and fusion of phonemes with its adjacent phonemes. Even if not these phonemes are not deleted, it lead to performance degradation when the duration of the phonemes is less than the number of states in a hidden Markov model (HMM). Therefore, we try to improved phoneme sequence in a lexicon by expanding with the variations based on a statistical method.

For estimation of word pronunciations or a lexicon, grapheme-to-phoneme (G2P) methods have been studied [3]. Typically, G2P is utilized for estimation of unknown-word pronunciations [4]. On the other hand, the statistical machine translation (SMT) tool Moses is also used for pronunciation estimation [5].

Because, the Japanese language is free from a phonographic writing system, the G2P methods have been studied less so far. The proposed method estimates variations of pronunciation from orthographic phoneme sequences in the lexicon instead of their character sequence. The estimation of spoken phoneme sequences are obtained by SMT trained by corpus paring orthographic phonemes consisting of phoneme defined in the lexicon and spontaneous spoken phonemes of speech corpus. Although the G2P-based phoneme sequence conversion/generation has been reported [6], there remains the issue of over-generation of variations that adversely affect the ASR performance. To address the issue, we propose a three-pass approach that suppresses the generation of such *outsider* entries in the expanded lexicon. In the first pass of our proposed lexicon expansion, the SMT generates pronunciation variations so that they cover a sufficient number of phonological phenomena that the original lexicon cannot capture. In the second pass, the training speech data are re-aligned by using the expanded lexicon. After re-aligning the training data, the unnecessary variations in the expanded lexicon according to their relative frequencies in the training data. While it is an empirical and simply way, it was never tried in the above mentioned literature [4], [5]. In [4], the lexicon was designed for excluding the pronunciations unobservable in the alignments. In addition to discarding pronunciations, in [5], the pronunciations that caused word errors were removed from the lexicon. In this paper, we will show our approach is an efficient solution to over-generation of pronunciation variations. In the third pass, another SMT phrase table was trained by aligned word corpus paired orthographic phonemes with spoken phonemes. This additional procedure will reduce the over-generation of pronunciation. This procedure reduces variations estimated by applying translation rules across the word boundaries. Because the reduced variations are not able to specify corresponding word, these are not adequate to our lexicon expansion. As the three-pass approach is expected to generate only promising pronunciation variations, the ASR performance should be improved.

II. Pronunciation Conversion Model

This section gives a brief review of statistical machine translation (SMT) followed by our approach to lexicon expansion.

A. Statistical Machine Translation

In SMT, given a sequence in the source language, $\boldsymbol{f}$, the optimal translated sequence in the target language, $\hat{\mathbf{e}}$, is obtained by

$$\begin{aligned} \hat{\mathbf{e}} &= \arg\max_{\boldsymbol{e}} P(\boldsymbol{e}|\boldsymbol{f}) \\ &= \arg\max_{\boldsymbol{e}} P(\boldsymbol{f}|\boldsymbol{e})P(\boldsymbol{e}), \end{aligned} \tag{1}$$

where $P(\boldsymbol{e})$ is the target-language model and $P(\boldsymbol{f}|\boldsymbol{e})$ is the source-to-target translation model. While the conventional SMT solves sequence-to-sequence conversion problems for two different languages, in G2P, the problem is considered as the conversion from graphemes to phonemes.

B. Lexicon Expansion

From a phonological viewpoint, the variability in spontaneous speech caused by, for example, the high rate of speech can be viewed as substitutions, deletions, and insertions of phonemes against the original pronunciation entries. Therefore, the variability in speech can be measured as the Levenshtein distances or differences between the original sequences and ones that mimic the utterance whenever the originals are available as reference labels. When a set of conversion rules are trained from the differences between sequences based on SMT, a lexicon that includes promising pronunciation variations can be build for robust ASR. In the SMT-based approach, source and target phoneme sequences are required for conversion modeling. The source phoneme sequences can be derived from forced-alignment between input acoustic features and HMM states based on orthographic pronunciations in the original lexicon. On the other hand, as the target sequences are indeterminate due to the variability or fluctuations in speech, we can never prepare the reference phoneme labels in a determinative way. Then, instead, the phoneme recognition results are used as targets. The SMT-based approach leads to a stochastic model including a set of conversion rules trained from these two phoneme sequences as training data. The key issue inherent in conversion modeling is over-generation of entries. In particular, in spontaneous speech, for example, short words such as Japanese particles are too ambiguous to be assigned to certain sequences by phoneme conversion. It is obvious that the ASR performance will degrade when such kind of pronunciation variations are registered into the lexicon. Even with words that have more phonemes as their pronunciations when they are observed with high frequency, the language model tends to assign high scores to the variations as well as the original pronunciation. Therefore, over-generation of pronunciations will probably degrade the ASR performance. In this paper, we propose a three-pass approach that removes such *outsider* entries from the expanded lexicon.

C. Utterance-Based Lexicon Expansion

Figure 1 shows an overview of the proposed lexicon expansion methods. In the utterance-based method, in the first pass, speech inputs as training data are aligned according to the original lexicon (baseline) while they are decoded by using a phoneme-based language model. From the resulting two phoneme sequences, we obtain a phrase table, which is a set of phoneme conversion rules for converting orthographic sequences to their variations. The application of the phrase table to the baseline lexicon leads to an expanded lexicon that reflects the variability in speech.

As described above, the expanded lexicon includes outsider entries that would be hardly ever used in real-world conversations. Then, we attempt to re-align the speech inputs by using the obtained expanded lexicon and remove outsiders according to the relative frequencies. The relative frequency, $f(p_i^k)$, is given by

$$f(p_i^k) = \frac{c(p_i^k)}{\sum_j c(p_j^k)}, \tag{2}$$

where p_i^k is the entry whose orthographic form corresponds to w^k, and $c(p_i^k)$ is the frequency of p_i^k. The new entry is dropped from the expanded lexicon if $f(p_i^k) < \alpha$, where α is the cutoff value, which is empirically determined. The entries are expected to survive through aligning the speech inputs if they are more promising and robust than the original one.

D. Word-Based Lexicon Expansion

By aligning phonemes with the expanded lexicon, we can perform word-by-word translation because exact word boundaries are obtained, unlike when using the former method. In this word-based lexicon expansion method, a word-by-word phrase table is obtained from two sets of alignments, and new entries are similarly generated by using the baseline lexicon. In the second pass, the outsiders are also removed from the expanded lexicon according to their relative frequencies to obtain a final lexicon. Since the obtained conversion rules are applied to the limited range within words, we can avoid the problem of ownerships of phonemes in contrast to the utterance-based expansion method. Therefore, the word-based expansion method is expected to capture the variations observed in words more effectively than the utterance-based method.

III. Experiments

A. Lexicon Expansion Setup

In the first pass of lexicon expansion, at the beginning, a translation model was trained from phoneme alignments and phoneme recognition results. The training data for translation modeling, or the parallel corpus, is shown in Table II. The phoneme recognition results corresponding to the target language in STM context were obtained as outputs from the DNN-HMM hybrid decoder, whose setup is described

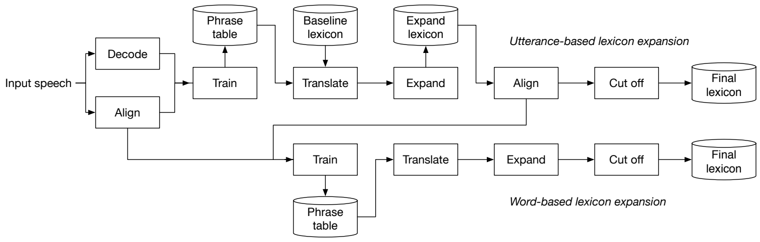

Fig. 1. Overview of Lexicon Expansion

TABLE I
EVALUATION DATA

PID	shows	#hours	#utterances	#words	PP	OOV(%)
001	5	8.3	11.6k	101.1k	211.7	1.0
002	5	2.1	1.9k	18.8k	221.2	0.6
003	2	2.5	3.0k	23.2k	236.2	0.8
004	5	2.1	3.3k	23.6k	277.3	1.3
005	5	4.1	3.7k	38.5k	154.8	0.5
006	5	2.5	1.9k	19.2k	164.5	0.5
007	5	2.7	2.8k	27.5k	160.4	0.6
008	5	6.7	5.8k	70.2k	116.7	0.5
009	5	3.8	5.5k	45.8k	244.2	1.2
010	5	4.2	3.8k	38.8k	183.7	0.7
011	5	3.6	5.0k	33.6k	286.8	0.7
012	5	2.1	1.7k	19.1k	128.2	0.5
013	5	4.5	4.3k	34.8k	310.9	0.8
all	62	49.0	54.3k	494.1k	194.7	0.8

TABLE II
TRAINING DATA FOR LEXICON EXPANSION

	#hours	#utterances	#phonemes
Training data	1817	107M	29.5M

in the following section. On the other hand, the alignments corresponding to the source language were generated by using the baseline lexicon. By regarding the alignments as reference labels, the phoneme error rate (PER) for the recognition results was 1.28 %.

From the phoneme recognition results, we trained a 4-gram language model, whose vocabulary consists of 40 Japanese phonemes using the KenLM toolkit [7]. For training the translation model, we performed the alignment by using GIZA++ [8]. To avoid the explosive increase in the number of unique phrases, we limited the phoneme sequence length to nine.

In the second pass of lexicon expansion, the training data were re-aligned by using the expanded lexicon. After alignment, unlikely entries were removed from the lexicon according to the cutoff value, α, in Eq. (2), and we obtained a final lexicon. Table III shows the entry sizes of the expanded dictionaries in the utterance-based lexicon expansion method.

B. Speech Recognition Setup

As listed in Table I, we used a variety of programs for testing the ASR performance. The programs are characterized by their topics, such as current affairs, sports events, and infotainment, as well as by the existence of background noise/music like opening/closing themes, jingles, and cheers in the stadium. Some programs consist of conversations among some casts. For the reference labels, the perplexity (PP) and out of vocabulary (OOV) rate were measured using a baseline trigram LM. The LM was trained from Japanese closed-caption texts (239M) with a 200k vocabulary. For the evaluation, we used a DNN-HMM hybrid decoder based on the Kaldi toolkit [9].

For acoustic modeling, we prepared 1400 hours of speech materials from broadcast programs such as news shows. According to the Kaldi's Wall Street Journal setup, we aligned the training data based on the Gaussian mixture models with 6841 senones. We utilized a 40-dimensional log-mel-filterbank with log energy plus their deltas and delta-deltas as an acoustic input feature and configured 11 frame splicing features as inputs to the DNN. The DNN consisted of eight hidden layers with 2048 sigmoid units each and was trained in the conventional way of pre-training and fine-tuning.

C. Experimental Results

For the utterance-based method, we compared the results from the baseline lexicon and the expanded ones while changing the number of entries according to the cutoff values in Eq.(2). Table V lists the overall word error rates (WERs) for the evaluation data. As shown in the table, the result of the lexicon (α=0.05) gave a poor WER performance compared with the baseline result. As is evident from Table III, the existence of many unnecessary variations caused a severe performance degradation. Apparently, the generated entries competed against correct entries that have similar pronunciations. On the other hand, among the cutoff conditions,

TABLE III
ENTRY SIZE OF EXPANDED LEXICON (UTTERANCE-BASED)

α (cutoff)	#entries	#added
baseline	368.0k	–
α=0.0	661.3k	293.3k
α=0.05	382.9k	15.0k
α=0.1	378.6k	10.7k
α=0.2	374.9k	7.0k
α=0.3	373.4k	5.5k
α=0.4	372.1k	4.2k
α=0.5	370.8k	2.9k

TABLE IV
PHRASE TABLES

	#entries
utterance-based	444.8k
word-based	41.9k

the lexicon (α=0.3) achieved the best WER of 40.6 % and reduced the absolute error reduction by 0.1 %. Although the cutoff based on the relative frequencies improved the performance, the WER stayed unchanged when the lexicon had few additional entries.

Table VI shows detailed results for each program and a comparison between the baseline, utterance-based lexicon expansion method, and word-based lexicon expansion method. The results show that the word-based expansion method achieved higher WERs than the utterance-based method. The similar results were reported in [4]. This is probably caused by the small difference between alignments generated by the original lexicon and the expanded one. Actually, as shown in Table IV, the phrase table size for generation of utterance-based lexicon is ten times larger than the table for the word-based lexicon. A phrase table including a sufficient number of phoneme variations could not be generated in the word-based lexicon expansion. A matched-pair testing [10] showed that the program (PID=006) was significantly improved when the utterance-based lexicon (α=0.3) was used. We did not achieve significant improvements for other broadcast programs.

In our investigation, even with high cutoff values, some of the high-relative-frequency outsiders still survived. All of the outsiders were not always removed from the lexicon when using relative frequencies as measures because they are estimated from small amounts of training data. One of the solutions to reducing the side effects caused by the outsiders is to reflect the probabilities of the pronunciation variants in the language model. Although the statistics or frequencies of the variants in the training data can be used in language modeling, the data size is typically far smaller than that of the text data, and it is not enough to estimate a robust language model. Thus, we need to use the data without reference labels to obtain the robust model.

IV. CONCLUSION

We described a new way of word pronunciation estimation based on statistical machine translation. The experimental results showed that our three-pass approach achieved promising results compared with the results from the original word lexicon. In future work, we will examine another way of entry selection and investigate how to deal with variations across word boundaries.

TABLE V
OVERALL RESULTS (UTTERANCE-BASED EXPANSION)

α (cutoff)	WER (%)
baseline	40.7
α=0.0	45.0
α=0.05	40.9
α=0.1	40.7
α=0.3	**40.6**
α=0.4	**40.6**
α=0.5	40.7

TABLE VI
DETAILED RESULTS

PID	baseline	utterance-based		word-based
		α=0.0	α=0.3	
001	52.4	56.7	**52.3**	52.4
002	40.9	45.6	40.9	41.0
003	42.0	46.2	**41.9**	42.0
004	57.3	62.7	57.3	57.3
005	29.1	33.0	29.1	29.2
006	28.2	31.1	**28.0**	**28.1**
007	32.9	37.1	32.9	32.9
008	23.9	27.2	**23.8**	**23.8**
009	59.4	64.3	59.4	**59.3**
010	38.4	42.8	38.4	38.4
011	45.4	51.1	**45.3**	45.4
012	19.1	22.8	**19.0**	19.1
013	38.2	42.7	**38.0**	38.2
all	40.7	45.0	40.6	40.7

REFERENCES

[1] T. Imai, S. Homma, A. Kobayashi, T. Oku, and S. Sato, "Speech recognition with a seamlessly updated language model for real-time closed-captioning," in *Proc. Interspeech*, 2010, pp. 262–265.

[2] A. Kobayashi, Y. Fujita, T. Oku, S. Sato, S. Homma, T. Arai, and T. Imai, "Live closed-captioning system using hybrid automatic speech recognition for broadcast news," in *Proc. NAB Broadcast Engineering Conference*, 2013, pp. 277–283.

[3] B. Maximilian, and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," in *Speech Communication*, 2008, pp. 434–451.

[4] A. Laurent, P. Deléglise, and S. Meignier, "Grapheme to phoneme conversion using an SMT system," in *Proc. Interspeech*, 2009, pp. 708–711

[5] A. Laurent, S. Meignier, and P. Deléglise, "Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions," in *Computer, Speech and Language*, 2014, vol.28, no. 4, pp. 979–996.

[6] L. Lu, and A. Ghoshal, and S. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *Proc. ASRU*, 2013, pp. 374–379.

[7] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proc. EMNLP*, 2011, pp. 187–197.

[8] F. Och, and H. Ney, "A systematic comparison of various statistical alignment models," in *Computational Linguistics*, 2003, vol.29, no.1, pp. 19–52.

[9] 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 *Proc. ASRU*, 2011.

[10] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.