Unsupervised Approach of Data Selection for Language Model Adaptation using Generalized Word Posterior Probability

Xinhui Hu, Shigeki Matsuda and Hideki Kashioka
National Institute of Information and Communications Technology,
Hikaridai 3-5, Seikacho, Sorakugun, Kyoto, Japan
E-mail: {xinhui.hu, shigeki.matusda, hideki.kashioka}@nict.go.jp

***Abstract*— This paper reports an unsupervised approach toward data selection for language model adaptation that is used for improving spontaneous speech recognition in a speech-to-speech translation (S2ST) system. The approach is characterized by the following: 1) it obtains speech data from a real environment (sightseeing sites), in the travel domain, (2) it utilizes the recognition results of the above collected speech for the language model adaptation, (3) it applies generalized word posterior probability (GWPP) among the N-best recognition hypotheses for the base of an utterance confidence measure to select adaptation utterances, (4) it utilizes a collected proper noun lexicon to the baseline language model in the form of zeroton event, so that it has ability to recognize new proper noun words that are previously not contained in the recognition lexicon. By experiments on a Chinese speech test collected from a set of field experiments at five sightseeing areas in Japan, using the above adapted language model, average absolute reductions of 7.6% of the character error rate (CER) were obtained, which is more than the baseline language model. This reduction is over 77% of the 9.8% reduction obtained by the supervised adaptation. By manually correcting a small amount of utterances that were not selected due to their low confidences, and adding them to the above adaptation data, nearly 83% of the reduction by the supervised method can be achieved. The proposed approach effectively improves utterance selection, especially for those containing proper nouns, and is expected to reduce the cost of manual transcription.**

I. INTRODUCTION

In stochastic automatic speech recognition (ASR) systems, the performance of the language model (LM) depends heavily on the size and quality of the corpora with which it is built. Generally, larger corpora with style and a domain closer to the recognition tasks will yield better performance. However, a training corpus cannot cover everything related to a specific field and the many expressional styles that reflect the vast variations of languages. To deal with these problems, LM adaptation is typically necessary for practical ASR systems. LM adaptation is generally realized by finding external textual resources as additional training data, building a new LM as an adaptation model with these data, and using the model to adjust the baseline model. Based on whether the adaptation data are manually annotated, the adaptation approaches are categorized as supervised or unsupervised. Compared with the supervised approach whose cost is too high or impractical, the unsupervised approach is more flexible because it collects adaptation data automatically. The use of recognition results for adaptation is one kind of unsupervised approach and is becoming a promising research area. For example, in [1], based on counts from the ASR transcripts of 17 hours of adaptation speech, various adaptation strategies, including iterative adaptation and self-adaptation on the test data, were verified to obtain 51% reduction of the word error rate (7.7%) that was obtained by supervised adaptation. To cope with the sparseness of LM space, [2] uses a class-based LM trained by recognition hypotheses for adaptation. This approach has proven to be effective in improving the word accuracy of spontaneous presentation speech recognition. Improvements in reducing perplexity and the word error rate (WER) were also found by adapting the speaker's characteristics in expression using initial speech recognition results to select similar texts to build an adaptation LM [3]. Most of these researches focus on directly using the recognized results.

Compared with these studies, many other researches introduce word confidence into unsupervised LM training [4][5]. In [5], for example, adaptation is realized by multiplying the word confidences to get a weighted n-gram count when estimating a LM, and 2% WER reduction is reported. Confidence scores are also used for selecting untranscribed user utterances for improving LM [6] and detecting mis-recognized utterances [7]. In [6], according to a confidence score which is based on confidence features at different levels such as phonetic, word, and utterance, the utterances having high scores are used to enhance LM directly, while those having low score are first manually transcribed and then used to train LM. In [7], a voting approach like recognizer output voting error reduction (ROVER) is used, where sub-LMs are trained by clustered utterances and used to rescore the word lattice given by the main LM. Based on the tabulated votes of the same recognized results from all LMs, the result's correctness is determined.

To build a practical ASR system, this study focuses on evaluating the extent of the improvement with the existing training data and small amounts of adaptation data. We adopt an unsupervised approach for LM adaption using the ASR

results. The adaptation utterances are selected by a confidence score measure that is based on a generalized word posterior probability (GWPP), which has been shown to be appropriate and effective to evaluate word recognition confidence [8]. To enable obtaining utterances containing out-of-vocabulary (OOV) words, an optional proper noun lexicon is applied to expand and augment the existing recognition lexicon.

The reminder of this paper is organized as follows. Section II briefly introduces data collection from real environments in a set of field experiments. Section III will describe the collection of a proper noun lexicon, and language model construction with consideration of this lexicon. Section IV presents the GWPP algorithm and utterance confidence measures based on the GWPP within a recognition N-best. In Section V, experimental comparisons among different adapted LMs are investigated. Section VI gives conclusions on this study.

II. FIELD EXPERIMENTS AND REAL SPEECH DATA

To obtain more speech data from real environments and to evaluate the performance of an existing S2ST system, the National Institute of Information and Communications Technology (NICT) conducted a series of field experiments at five sightseeing areas in Japan. Chinese, English, Japanese, and Korean were the four target languages. The experiments were conducted between foreign tourists and Japanese staff working at the sightseeing spots. For Chinese, the native Chinese (Mandarin) speakers were tourists from different areas of China; most came from southern China where the pronunciation is often influenced by local accents. All the speech data during the experiments were collected. Many new characteristics extracted from these data are expected to be utilized to improve ASR from both the acoustical and language model aspects. For example, as described below, there is a clear characteristic that many proper nouns, like Japan locations, festival events, and souvenir names, are contained in these conversational speech data. Here, in this study, we limit to use these data as the original source of LM adaptation.

III. COLLECTED PROPER NOUN LEXICON AND LANGUAGE MODEL FOR DATA SELECTION

Because an ASR can only recognize the words that are registered in advance in its recognition lexicon, the OOV words appeared in utterances will be mapped to the closest in-vocabulary (IV) words. Most OOV words originate from such proper nouns as personal and location names. If these proper nouns cannot be recognized correctly, the utterances containing them might not be selected due to low utterance confidence scores. Even if they are selected, since the transcripts made by the ASR do not contain correct proper nouns, they will not benefit the recognitions of these kinds of words. On the other hand, however, the proper nouns play very important roles in speech communications, for examples, in the S2ST services, and demand for high recognition accuracies.

A. *Collection of optional proper noun lexicon*

To improve the recognition of proper nouns, our approach expands the recognition lexicon using a collected proper noun lexicon that supplements the baseline lexicon. Because our task is fixed to a travel domain and to several specific sightseeing spots, gathering possible proper nouns that might appear in real speech is relatively easy. In this study, the proper nouns are only limited to famous place names, Japanese personal names, festival events, and souvenir names, in five areas of Japan. These areas include Chubu (中部), Hokkaido(北海道), Kanto (関東), Kansai (関西), and Kyushu (九州). These collections are mainly from websites and existing textual database.

At present, we have built a proper noun list that covers these areas whose vocabulary is about 29.8 K words. These proper nouns are categorized into the ten groups as shown in Table 1.

Table 1 Descriptions of collected proper nouns

Category	Count	Examples
Food	1799	阿苏雪菜小鱼炒饭
Country	17	美国
Japan location	20065	濑户大桥
Sightseeing spot	2863	中富良野花公园
Souvenir	1637	奈良团扇
Given name	40	秀喜
Family name	530	前田
Quotation	10	百人一首
Organization	3131	阪急百货商店
Festival event	1508	春之熊本城节
Total	29.8 K	

B. *Language model containing zeroton event*

The language model used for data selection is trained using both the baseline corpus and the collected proper noun lexicon. The words in the proper noun lexicon are added into the training corpus. The ones that do not exist in the baseline corpus are modeled as the so-called zeroton event. The zeroton fraction is assigned to 1.0 in this study. Here, the language model is built by the MIT language modeling toolkit [10].

IV. GWPP-BASED UTTERANCE CONFIDENCE AND ADAPTATION DATA SELECTION

A. *GWPP for word reliability*

The word posterior probability is the probability of a focused word, given the acoustic observations of a sentence and a statistical speech recognizer. It offers many advantages and is interesting for a number of applications. For example, it can be used as the estimation of confidence measure. The larger the word posterior probability, the more likely the focused word is correctly recognized.

Generalized word posterior probability (GWPP) [8], a probabilistic confidence measure for optimally verifying recognized entities at the word level, is based on the posterior probability of a word in an ASR. It assesses the reliability of a focused word by looking at its reappearances in the word graph and exponentially reweighting the corresponding acoustic and language model likelihood. As its characteristics, the GWPP generalizes this word posterior probability by considering the following three practical issues:

(1) Reduced search space
(2) Relaxed time registration
(3) Reweighted acoustic and language model likelihood

The formula for GWPP(w) is as follows:

$$p([w,s,t]\,|\,x_1^T)=\sum_{\substack{\forall M,[w,s,t]_1^M \\ \exists n,1\le n\le M \\ w=w_n \\ [s,t]\cap\{s_n,t_n\}\neq\phi}}\frac{\prod_{m=1}^{M}p^{\alpha}(x_{s_m}^{t_m}\,|\,w_m)p^{\beta}(w_m\,|\,w_1^{m-1})}{p(x_1^T)} \qquad (1)$$

where [w;s,t] is the focused word *w* with starting time *s* and ending time *t* and x_1^T is the sequence of the acoustic observations. M is the number of words in the current string, and α and β are the exponential weights for the acoustic and language models, respectively.
Due to its effectiveness in evaluating word recognition confidence, we adopt the GWPP as the base of utterance confidence for selecting the recognized results in this study.

B. Utterance confidence and data selection

Due to special problem of an ASR system that recognition errors always exist, sometimes, correct recognition candidates are contained in the hypotheses other than the top one (1-best), we extend the data selection from conventional 1-best to N-best. The N-best based GWPP is defined as follows:

$$GWPP_{N-Best}(w)=(\sum_{j=1}^{N}GWPP_j(w))/N \qquad (2)$$

Here, $GWPP_j(w)$ is the GWPP of word w in the *j*th recognition hypothesis. N is the number of N-best. In this study, N=10. The $GWPP_{N-Best}(w)$ is the average GWPP of word w in the N-best it belongs to.
Based on the $GWPP_{N-Best}(w)$, an utterance confidence is therefore proposed as:

$$CF_{utterance}=(\sum_{i=1}^{M}GWPP_{N-Best}(w_i))/M \qquad (3)$$

Here, the w_i is *i*th word in a recognition hypothesis and the M is the word number within this hypothesis.
Adaptation utterances are chosen from all N-best results. Assuming that $CF_{threshold}$ is a threshold, an utterance is selected if its utterance confidence exceeds the $CF_{threshold}$.

C. Adapted Language Model

The linear interpolated LM is adopted for building an adapted LM, it is formulated as [11]:

$$LM_{adapted}=\lambda\cdot LM_{baseLM}+(1-\lambda)\cdot LM_{adaptation} \qquad (4)$$

Here, LM_{baseLM} is the LM trained by the baseline corpus, and $LM_{adaptation}$ is the LM trained by the adaptation utterances. λ is the interpolation coefficient, which is determined by evaluating the perplexity of the development set to the adapted LM.

V. EXPERIMENTS AND RESULTS

A. Data Setting for Experiments

Two kinds of data were used in this study: a Chinese textual corpus [9] for the training baseline LMs and a set of Chinese speech data collected from the field experiments mentioned in the previous sections. Each kind of data contained training data and test sets. The second one also contained a development set. The detailed information of these data are shown in Table 2. The word counts and the OOV rates of the collected speech data (preceded with "Adp") are computed using their manual transcripts. For computing the OOV rates, BaseTrain, a baseline training set, was used.
From this table, we know that the OOV rate of the collected data is clearly higher than the baseline one. Since the AdaTestP is specially used for evaluating proper noun recognition, we intentionally selected it so that at least one proper noun was contained in each utterance.

Table 2 Descriptions of data sets

Set	Purpose	Utter-ances	Words	OOV rate
BaseTrain	Training	510 K	464.5 K	
BaseTest	Test	510	2600	0.26%
AdpTrain	Adaptation	14.7 K	71.2 K	7.33%
AdpTest	Test	524	2545	6.56%
AdpDev	Development	506	2598	6.41%
AdpTestP	Test OOV	528	2935	15.57%

B. The best threshold and selected utterances

Figure 1 shows the selected utterances from the recognition results (N-best) of data AdpTrain when using the language model for utterance selection, and the perplexities of

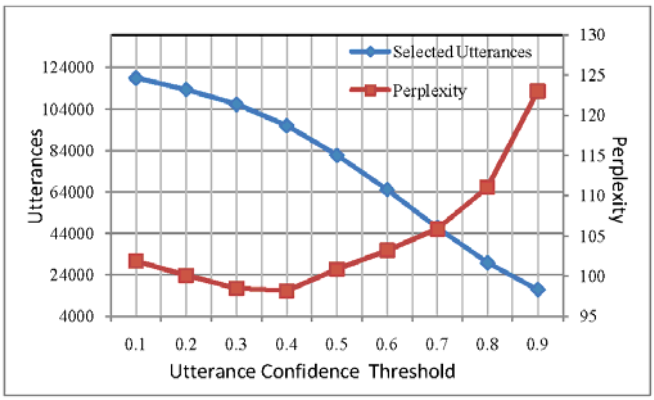

Fig. 1 Selected utterances from 10-best and perplexities of development set to the adapted LMs in different utterance confidences threshold

development data AdpDev to the adapted LMs, in different utterance confidence threshold. These adapted LMs are trained by their corresponding selected utterances.

The optimized threshold is found at $CF_{threshold}$ =0.4 and about 96 K utterances were selected from all 10-best results. These utterances correspond to 11K original utterances shown in table 2.

C. *Evaluation models and experiment types*

To evaluate the performance of using the selected utterances, we compared five types of LM:

(1) BaseLM – trained by data BaseTrain.
(2) SimpUnsup– adapted LM using all recognized results of data AdpTrain.
(3) Supervised – adapted LM using all manually checked transcripts of data AdpTrain.
(4) GWPP – adapted LM using the selected utterances from data AdpTrain at the best threshold.
(5) GWPPMC – adapted LM using two utterance sets; one is the data used in GWPP, another is the un-selected utterances (2.3K) due to low confidence. The later is manually corrected before for adaptation.

For all the above models, using the proper noun lexicon (PnLex) and without it (NoPnLex) to train the baseline LM are compared.

The recognition results of data AdpTest using the above LMs are shown in Fig. 2.

D. *Improvement of proper noun recognition*

Table 3 shows the recognition results of AdpTestP using the GWPP model. PnCER is the CER on only proper nouns. The performance of the proper noun recognition is largely improved by using the proper noun lexicon.

Table 3 Recognition results of set AdpTestP

	NoPnLex	PnLex
CER[%]	43.50	32.36
PnCER[%]	65.37	34.44

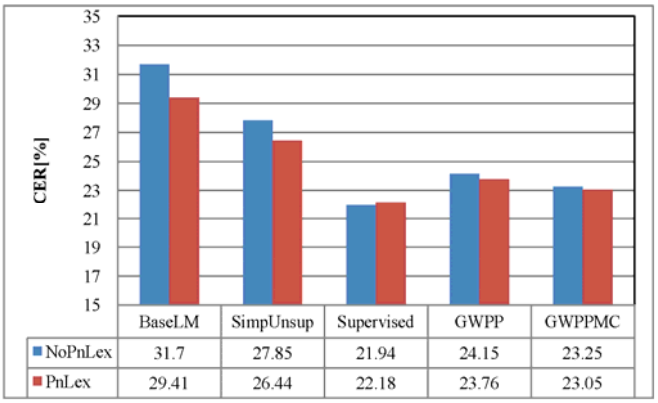

Fig. 2 Recognition results of data AdpTest using different adapted LMs.

VI. CONCLUSIONS

We proposed an unsupervised approach to select utterances from ASR output for LM adaptation. To obtain more utterances containing proper noun words, we chose a collected proper noun lexicon, and added its words into LM training in the form of zeroton event. This is verified effective to recognize proper nouns that do not exist in the original training corpus. For examples of this study, word "加藤清正公像, 黒川温泉, 旧細川邢部邸, 熊本机场，天草," are not collected in the baseline training corpus, however, they are recognized correctly with the help of using the collected proper nouns. Based on the facts that correct recognition candidates are sometimes contained in the hypothesis other than 1-best, an N-best based GWPP utterance confidence measure is proposed to select effective utterance.

By experiments on a set of test speech data collected from real environmental field experiments, the selected utterances by the GWPP approach were proven to be more effective for LM adaptation than the SimpUnsup which uses simply all recognized results as adaptation data. Compared with the SimpUnsup, CER decreased 2.6% from 26.4 to 23.8%. Of the improvements (to the BaseLM) obtained by the Supervised (9.8% in NoPnLex, 7.2% in PnLex), 77% and 78% were respectively achieved by the GWPP approach (7.6% and 5.7%). We also investigated the semi-supervised approach based on the GWPP by using a small amount of manually corrected speech transcripts. We found that by correcting only those unselected utterances (2.3K), and adding them to the GWPP adaptation data, nearly 83% of the reduction by the supervised method can be achieved (In PnLex case). This shows that the manual annotation of speech can be efficiently improved.

Although the collected proper noun lexicon used in this study has ability to improve the recognition performance, its improvement will become small with the increase of valid additional lexicon to the LM. As shown in Fig.2, for examples, the improvement of using the collected lexicon is the biggest in the case of BaseLM, then, the SimplUnsup is the second in which some new proper nouns are contained in the selected utterances and these words are no longer used as

zeroton event in the LM. In the case of Supervised where all correct words of adaptation data are added to the training corpus, the recognition becomes worse in PnLex than in NoPnLex. The differences between NoPnLex and PnLex in the cases of GWPP and GWPPMC are very small. This fact also suggest that these LMs have tendencies to be close to the Supervised one.

Future work will investigate the relationship between the order of N-best and the adaptation performance, and compare more utterance confidence measures based on the GWPP. Different variations of zeroton fraction for using additional lexicon in LM training will be studied. Furthermore, more detailed works will be on selecting utterances for semi-supervised approach, which is important to improve the efficiency of manual speech transcription.

ACKNOWLEDGMENT

The database used in this study was provided by a project supported by the Ministry of Internal Affairs and communications, Japan in 2009.

REFERENCES

[1] M. Bacchiani and B. Roark, "Unsupervised Language Model Adaptation," Proc. ICASSP 2003, pp. I224-227, 2003.

[2] T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui, "Unsupervised Class-based Language Model Adaptation For Spontaneous Speech Recognition," Proc. ICASSP 2003, pp. I-236-239, 2003.

[3] H. Nanjo and T. Kawahara, "Unsupervised Language Model Adaptation for Lecture Speech Recognition," Proc. SSPR, pp. 75-78, 2003.

[4] R. Gretter and G. Riccardi, "On-line Learning of Language Models with Word error Probability Distributions," Proc. ICASSP 2001, pp. 557-560, 2001.

[5] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised Acoustic and Language Model Training with Small Amounts of Labelled Data," Proc. 2009, pp. 4297-4300, 2009.

[6] M. Nakano and T. J. Hazen, "Using untranscribed user utterances for improving language models based on confidence scoring," Proc. Interspeech 2003, 417-420, 2003.

[7] K. Fujinaga, H. Kokubo, H. Yamamoto, G. Kikui, and H. Shimodaira, "Mis-recognized utterance detection using multiple language models generated by clustered sentences," Proc. Interspeech 2003, pp. 2793-2796, 2003.

[8] W.K. Lo and F.K. Soong, "Generalized Posterior Probability for Minimum Error Verification of Recognized Sentences," Proc. ICASSP 2005, pp. I85-88, 2005.

[9] X. Hu, R. Isotani, H. Kawai, and S. Nakamura, "Construction and Evaluation of an Annotated Chinese Conversational Corpus in Travel Domain for the Language Model of Speech Recognition," Proc. Interspeech 2010, pp. 1910-1913, 2010.

[10] https://code.google.com/p/mitlm/

[11] R. Kneser and V. Steinbiss, "On the Dynamic Adaptation of Stochastic Language Models," Proc. ICASSP 1993, pp586-589, 1993.