Expansion of Training Texts to Generate a Topic-Dependent Language Model for Meeting Speech Recognition

Kazushige Egashira, Kazuya Kojima, Masaru Yamashita, Katsuya Yamauchi and Shoichi Matsunaga
Nagasaki University, Nagasaki, Japan
E-mail: {b608216, b608244, masaru, yamauchi, mat}@cis.nagasaki-u.ac.jp Tel: +81-95-819-2700

Abstract— **This paper proposes expansion methods for training texts (baseline) to generate a topic-dependent language model for more accurate recognition of meeting speech. To prepare a universal language model that can cope with the variety of topics discussed in meetings is very difficult. Our strategy is to generate topic-dependent training texts based on two methods. The first is text collection from web pages using queries that consist of topic-dependent confident terms; these terms were selected from preparatory recognition results based on the TF-IDF (TF; Term Frequency, IDF; Inversed Document Frequency) values of each term. The second technique is text generation using participants' names. Our topic-dependent language model was generated using these new texts and the baseline corpus. The language model generated by the proposed strategy reduced the perplexity by 16.4% and out-of-vocabulary rate by 37.5%, respectively, compared with the language model that used only the baseline corpus. This improvement was confirmed through meeting speech recognition as well.**

I. INTRODUCTION

Minutes of meeting are usually taken manually, and this involves enormous time and effort. Automatic minute taking (AMT) systems are expected to reduce this. Recognizing meeting speech is an important function of AMTs; however, in certain cases, such as when the topics discussed and vocabulary used are diverse, this can be difficult. Developing a topic-dependent language model will be useful in addressing the above difficulty and in recognizing the meeting speech accurately.

Some studies have involved using topic-dependent training texts to develop the aforementioned language model. Additionally, previous studies have used a large corpus of texts. Tur et al. [1, 2] used a corpus comprising meeting speech transcriptions that were collected manually. Hain et al. [3, 4] used a collection of texts sourced from telephone conversations and the Internet and preparatory speech recognition results of the target meeting. To develop a system to automatically transcribe minutes of the meeting of the Japanese National Congress, Akita et al. [5] used a large corpus of transcriptions from previous Congress meetings.

As it is difficult to prepare topic-dependent training texts for individual meetings, we posited methods more efficient and practical than collecting large amounts of text. In this paper, we proposed two methods to develop a corpus of topic-dependent training texts: text collection from web pages and text generation using participants' names.

Words that appear in the transcription of a meeting are indicative of the meeting's topic. Certain words used in a topic might not be presented in the common corpus. We proposed a method that used web texts which were collected by the queries determined from preparatory speech recognition results of a meeting. The queries, consisting of topic-dependent confident terms, were selected on the basis of the TF-IDF (TF; Term Frequency, IDF; Inversed Document Frequency) values of each term in the preparatory result.

Additionally, prior information on meetings can be useful for topic-dependent expansion. Participants' names constitute prior information and usually appear repeatedly in meeting speech. The out-of-vocabulary (OOV) rate of the language model generated from the common corpus is high and frequently contains participants' names. We proposed a method to add texts containing participants' names to the common corpus.

Our topic-dependent language model was generated using these new texts and the common corpus. We used newspaper text and lecture transcription as the common (baseline) corpus, because the former contains a variety of topics and the latter, spontaneous speech similar to meeting speech. Although, the newspaper text used written language style, meeting speech used spoken language style. To generate a language model for spontaneous speech, paraphrasing predicates from written to spoken language was found to be effective [6]. We converted the language style in the newspaper text from written to spoken language style.

The improvement in the language models generated using the expanded texts was confirmed through the OOV rate, perplexity (PP), and performance of meeting speech recognition.

II. FLOW AND DATABASE OF RECOGNITION

A. Flow of Meeting Recognition Procedure

The flow of our meeting speech recognition system was shown in Fig.1. There are multiple participants in our target meeting. It was necessary to detect voice activity sections accurately for each participant, because noises and utterances from other speakers can lower recognition accuracy. This

detection process is indicated as VAD (Voice Activity Detection) in the figure.

Preparatory recognition was performed using an initial language model and a lexicon; those were made from baseline corpus. Then, search queries were determined using the results of the preparatory recognition. Using the queries, additional topic-related training texts for the target meeting were obtained from web pages. New texts containing participants' names were also generated. Finally, speech recognition was carried out using the topic-dependent language model and lexicon that were generated using the baseline corpus, obtained topic-related texts, and newly generated texts containing participants' names.

B. *Database of Meeting Speech*

We used 12 mock meeting speeches in which prepared agendas were discussed. Six types of agendas (e.g., catalog editor meeting in mail-order house, teachers' meeting in an elementary school) were used. The participants in the meetings spoke on the basis of a prepared rough scenario for each agenda. Two types of scenarios for each agenda were prepared: one was for five participants and the other for six. One of the participants played the role of a moderator.

The average length of the meetings was approximately 50 minutes. The meetings were recorded using lapel microphones attached to each speaker. Participants' speeches were recorded separately on five or six synchronized channels. The sampling frequency, which was 44.1 kHz during recording, was down-sampled to 16 kHz. We prepared labels for the utterance periods and transcriptions of each participant to evaluate VAD and speech recognition performance.

A Japanese newspaper database (50.5 MB) and a transcription of lectures (24.4 MB) from the Corpus of Spontaneous Japanese (CSJ [7]) were used as the baseline corpus.

III. EXPANSION OF TRAINING TEXTS

We examined expansion methods for training texts to generate a topic-dependent language model and lexicon for each target meeting. The texts were prepared using the following three methods:

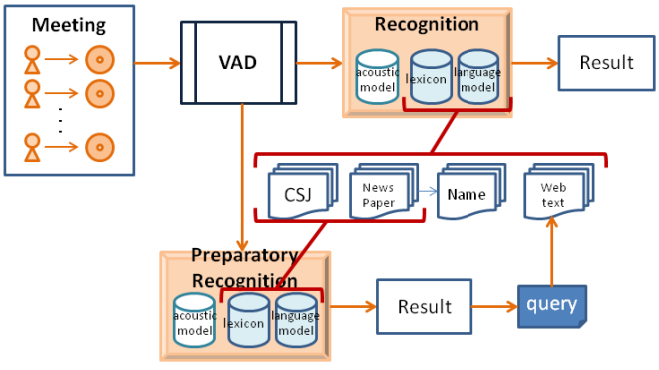

Fig. 1 Flow of meeting recognition

- style conversions from written to spoken language (conventional method)
- text generation using participants' names (method 1)
- text collection from web pages (method 2)

We proposed a way to combine the baseline corpus with the text developed using these methods.

A. *Style conversions from written to spoken language (conventional method)*

In order to recognize meeting speech accurately, it is ideal to generate a language model using transcriptions that consist of a spoken language style. While the lecture transcription consisted of a spoken language style, a large part of the text in the newspapers consisted of a written language style. Therefore, the newspaper texts were converted so that they consisted of the spoken language style. The method for doing so has been studied in previous research (e.g., rule-based text transformation using lexical and syntactic information, and statistical transformation of N-gram entries in [5]). In this study, the difference between the Japanese expressions *"keitai"* and *"joutai"* was examined. The former is an honorific expression and the latter, a normal one. The sentences in the newspaper were changed using 14-style conversion rules decided in advance.

B. *Text generation using participants' names (method 1)*

Prior information on meetings, such as the agenda and participants' names, is usually acquired before the meeting begins. Usually, participants' names frequently appear in meeting speech. We examined OOV using a baseline language model. The results showed that the majority of the 19.4% OOV was participants' names. It was difficult to source all participants' names from the newspaper text and lecture transcriptions. One way to resolve this was to add participants' names directly into the lexicon. However, this method does not take into account the links between words. Therefore, we generated a language model and lexicon from training texts that included participants' names. We randomly chose seven sentences from the newspaper text that contains names of people and replaced people's names with participants' names in each meeting. Then, we added the replaced sentences (40 KB) to the training text.

C. *Text collection from web pages (method 2)*

We collected documents related to the target meeting from web pages using the preparatory recognition results. Then, we added the documents to the training text for each meeting.

The query was determined using the results of preparatory recognition based on TF-IDF values. We used the TF-IDF as the one of the method to determine the particular terms for each meeting. Eleven kinds of nouns that were used in a morpheme analysis [8] constituted the items in the query. We searched the Internet using the queries and collected the top 100 documents of search result list.

TF and IDF are defined as

$$TF(i,j) = \frac{\text{term } j \text{ frequency}}{\text{number of morphemics in document } i} \quad (1)$$

$$IDF(j) = \log \frac{\text{number of all texts in database}}{\text{number of texts containing term } j} \quad (2)$$

$$TF\text{-}IDF(i,j) = TF(i,j) \times IDF(i) \quad (3)$$

Newspapers were used as the database because they provide information on many topics. Words whose TF-IDF values were higher than the threshold were used for the keywords that constituted the queries. If 100 documents were not collected by the first query, more documents were collected through the next query. The next query was determined by a higher threshold than the initial query was. Additionally, the next query had less keywords than the first. The documents collected in the first search were excepted. This procedure was repeated until 100 documents were collected. The initial threshold for TF-IDF value was 0.02. As an example, the queries made for the agenda "catalog editor meeting in mail-order house" are shown in Table I.

IV. Experiment

A. Language Models

We generated the language model from seven kinds of combination of training texts. Table II shows the combinations. Model "c" is regarded as the baseline. In order to simulate the ideal documents collected from the web, we used transcription of meeting spoke on the same scenario as "similar text" (30 KB).

B. Evaluation by PP and OOV

Figure 2 shows the averaged PP and OOV for the 12 meetings. In comparison to model "a," the PP of model "b" was lower. The result showed that the style conversion of the newspaper texts was effective for reducing OOV. On the other hand, the differences of PP and OOV rate between model "c" and "d" were small. When the lecture transcriptions that contained spoken language were included, the effect was lost. Therefore, it is not necessary to convert the style if the training texts contain a large amount of spoken language style texts.

The OOV rate of model "e" (0.63%) was lower than that of model "d" (0.80%). This indicates that method 1 was effective for decreasing the OOV rate. The occurrence percentage of participants' names in OOV words was 19% in model "d," which is higher than the percentage for other words. The percentage was approximately 1% in model "e," because OOV words such as participants' names that frequently occurred in other models did not occur in this one. On the other hand, when we added participants' names to the training corpus, PP increased by a small amount. In our proposed method, we prepared a set of seven sentences, where each set contained each participant's name. Then, it is considered that the occurrence probability of the word sequences in every set was increased. However, this increase did not generate a large negative effect on PP.

The PP and OOV rates for model "f" (112, 0.50%) were lower than those for model "e" (135, 0.80%), including web texts which were collected by proposing procedure using the TF-IDF value performed well to make the language model and the lexicon adapt to target meeting. Therefore, method 2 was effective for generating topic-dependent language model.

When we used model "g," which was generated using a training corpus that included the transcription of meeting spoke on the same scenario, it was observed that the PP and OOV rate using "g" were lower bounds (37, 0.42%). Then, in comparison to the results of our proposed method and the method using model "g," the former could decrease the OOV rate sufficiently. However, the PP derived using our approach was higher. Thus, it is necessary to develop a better method to decrease the PP in future.

C. Evaluation of the meeting speech recognition experiment

1) VAD: In this experiment, we fit Gaussian mixture models to the acoustic features of utterance and non-utterance sections. The acoustic features were 12 mfcc, Δmfcc, power, and Δpower. The difference between the likelihood of utterance and non-utterance sections in each analytical frame was calculated. If the difference was larger than the threshold value, the frame was classified as an utterance sections; otherwise, the frame was classified as a non-utterance section. The speaker who had the highest power value on the frame was classified as the target speaker. The utterance and non-utterance sections that were smoothed through the following procedure: utterance sections that were too short were treated as non-utterance sections, and non-utterance sections that were too short were treated as utterance sections.

We detected the utterance sections of nine meetings, for which utterance periods were prepared. As a result, the

TABLE I
EXAMPLE OF QUERIES FOR A MEETING AGENDA
(CATALOG EDITOR MEETING IN MAIL-ORDER HOUSE)

Query	Keywords	Number of texts
1	item catalog page operation project special-topic logistics eco	57 (2.1MB)
2	item catalog page operation project special-topic	43 (1.0MB)

The keywords were in Japanese.

TABLE II
COMBINATION OF TRAINING TEXT

Language model	Newspaper		CSJ	Participants' names	Web	Similar text
	Original	Style conversion				
a	✓					
b		✓				
c (baseline)	✓		✓			
d		✓	✓			
e		✓	✓	✓		
f (proposed)		✓	✓	✓	✓	
g		✓	✓	✓		✓

average for f-measure was 0.971 and standard deviation was 0.008. The performance to detect utterance sections was considered to be high enough for meeting speech recognition.

2) Meeting speech recognition experiment: We conducted the meeting speech recognition experiment using language models "d," "e," "f," and "g." In this experiment, we regarded model "d" as the baseline. The acoustic model used in this experiment was part of CSJ, which comprises 819 academic presentation speeches with a combined length of 274.4 hours. The conventional 3000 states 16-mixture hidden Markov models were constructed with the set of 27 Japanese phonemes, while the feature parameters were 12 mfcc, Δmfcc, power, and Δpower.

Table III shows the word correct (Corr) and the word accuracy (Acc) for each language model. In comparison to the baseline model "d," the Corr and Acc of proposed models "e" and "f" were higher. The result showed that the proposed methods were effective. According to the small amount of texts added to the baseline corpus, the recognition performance was increased a little by the addition of participants' names (method 1) .

The Corr and the Acc of model "g" were not high enough. The result suggested that the acoustic model that was generated using CSJ was not suitable for recognizing meeting speech. Therefore, our future challenge is to determine the methods to construct a suitable acoustic model for recognizing meeting speech.

V. CONCLUSIONS

This paper proposed methods to expand training text from a common (baseline) corpus to a topic-dependent one for accurate meeting speech recognition. The participants' names were included in the training text by adding seven sentences that contained the participants' names in each meeting (method 1). The web text was collected using queries that consisted of topic-dependent confident terms selected from preparatory recognition results based on the TF-IDF value of each term (method 2). The topic-dependent language model was generated using these new texts and the baseline corpus.

The language model generated using the training texts

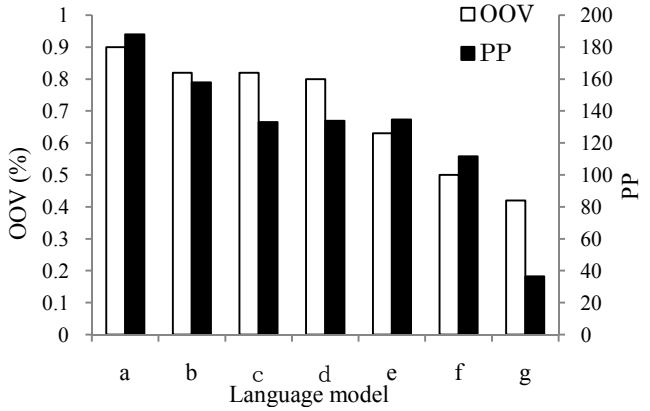

Fig. 2 Result of OOV and PP.

TABLE III
RESULTS OF MEETING RECOGNITION EXPERIMENT [%]

Language model	Corr	Acc
d (baseline)	62.19	50.66
e	62.38	50.99
f	63.66	52.87
g	67.39	58.16

containing participants' names reduced the OOV rate by 21.2% in comparison with the baseline model. The method to include prior information, such as the participants' names, on meetings performed well to reduce OOVs of the language model. Furthermore, the PP and OOV were reduced by adding the web text in method 2. The language model generated by the proposed strategies reduced the perplexity by 12% and OOV rate by 40%.

This paper also examined the advantage of style conversion from written language to spoken language in the conventional method. Our result showed that it was not necessary to convert the style if the training texts contained a large amount of spoken language.

The improvement caused by our expansion methods was also confirmed through the meeting speech recognition. The proposed procedure achieved higher performance than the baseline; however, the recognition accuracy was not high. It was considered that the acoustic model used in the experiment was not suitable for recognizing meeting speech. Our future challenge is to develop a method to construct a suitable acoustic model for recognizing meeting speech.

REFERENCES

[1] G. Tur and A. Stolcke, "Unsupervised language model adaptation for meeting recognition," *Proc. ICASSP*, Vol. 4, pp. 173-176, 2007.

[2] D. Vergyri, A. Stolcke and G. Tur, "Exploiting user feedback for language model adaptation in meeting recognition," *Proc. ICASSP*, pp. 4737-4740, 2009.

[3] T. Hain, L. Burget, J. Dines, P. N. Garner, A. E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln and V. Wan, "The AMIDA 2009 meeting transcription system," *Proc. INTERSPEECH*, pp. 358-361, 2010.

[4] S. Kombrink, T. Mikolov, M. Karafiat and L. Burget, "Recurrent neural network based language modeling in meeting recognition," *Proc. INTERSPEECH*, pp. 2877-2880, 2011.

[5] Y. Akita, M. Mimura and T. Kawahara, "Automatic transcription system for meeting of the Japanese national congress," *Proc. INTERSPEECH*, pp.84-87, 2009.

[6] N. Kaji, M. Okamoto and S. Kurohashi, "Paraphrasing predicates from written language to spoken language using the web," *Proc. Human Language Technology Conference*, pp.241-248, 2004.

[7] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," *SSPR-2003*, MMO2, 2003.

[8] http://chasen-legacy.sourceforge.jp/