WESPAC IX 2006
The 9th Western Pacific Acoustics Conference
Seoul, Korea, June 26-28, 2006

AUTOMATIC TRANSCRIPTION OF MEETINGS USING TOPIC-ORIENTED LANGUAGE MODEL ADAPTATION

Yuya AKITA Carlos TRONCOSO Tatsuya KAWAHARA

Academic Center for Computing and Media Studies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

This paper presents an automatic speech recognition (ASR) system dedicated for meetings of the National Congress of Japan. The distinctive features of the congressional meeting speech are wide distribution and frequent change of topics. For more accurate transcription, such topics should be emphasized in a language model one after another. Therefore, we introduce two approaches for topic adaptation: PLSA-based approach and trigger-based approach. The PLSA-based adaptation is performed turn by turn to emphasize topics in individual pair of a question and answer. Since topics are treated in a probabilistic manner, robust adaptation is realized. On the other hand, the trigger-based adaptation stresses relevant words to the word history, thus long-distance context can be reflected into a language model. These two approaches were evaluated on real meetings of the Congress, and significant improvement of perplexity was obtained by both approaches. We also compared their effects on reduction of word error rates.

KEYWORDS: Speech recognition, Language model adaptation, PLSA, Trigger model

INTRODUCTION

Recently, research targets of automatic speech recognition (ASR) have shifted to spontaneous speech such as lectures and meetings. Automatic transcription and summarization of such speech materials are promising applications of ASR. We are investigating an ASR framework for meetings of the National Congress (Diet) of Japan.

Meeting in the Congress is always transcribed as a record. Furthermore, some sessions are digitally archived, so closed captions will be needed in the near future. At present, utterances are taken down in shorthand, and then edited afterwards for documentation by professional stenographers who are specially trained in the national institute. This scheme is obviously costly, so its termination is decided by the Japanese government. On the other hand, it is difficult in Japanese to transcribe in real time by typing, since a large number of homonyms appear in Japanese sentences and selection of correct words (in *kanji* notation) takes much time. Therefore, ASR will be useful for creation of records and closed captions. However, ASR for this kind of speech has rarely been studied.

As a similar task, the NIST "Rich Transcription (RT)" project has dealt with meetings[1]. The target is relatively informal meetings on business or research, where participants make utterances more spontaneously. Thus, ASR was difficult, and word error rate (WER) was around 30–40%. ASR for courtroom speech has also been investigated[2], and comparable WER was obtained, due to the adverse acoustic condition, emotional speech and spontaneity of speech. The speech of the National Congress is relatively more formal and acoustic condition is better than these kinds of speech. However, highly accurate or almost perfect transcription is needed for the National Congress since all utterances must be recorded completely, while accurate transcription is not necessarily required for general meetings.

For accurate transcription, one of distinctive problems to be solved is wide distribution of topics. The topics of the meetings include budget, economy, security, education and foreign affairs, and the topic changes frequently and abruptly by (speaker) turns, even if the same speaker talks. A uniform language model cannot properly represent these topics, hence it should be adapted to current topics. In this paper, we incorporate a PLSA-based approach[3] and a trigger-based approach[4] to adapt a language model. The former uses probabilistic mapping to automatically defined topic clusters, while the latter directly utilizes correlation between words. Both approaches can represent relatively longer context dependency, however, their methodologies are quite different. Therefore, we evaluate and compare these two approaches using real speech of the congressional meeting.

TASK AND BASELINE SYSTEM

Task Description. We have been preparing a speech corpus of meetings in the National Congress (the House of Representatives). Participants of meetings are mainly members of the Cabinet, members of the Congress and government officials. The majority of the corpus originates from the committee of budget, in which a variety of domestic and international issues are discussed and the topic changes frequently even in a same person's talk. The utterances were transcribed manually and faithfully, and used as a reference text. Audio data is also manually segmented into speaker turns by detection of speaker changes.

Specifically, audio data of the committee held on February 14, 2003 was used as a test-set in this work. There are 23 participants in this time and the duration of speech is about 5.5 hours. The total number of words is 62,512. Interpellators, who are members of the Congress, state their opinions and ask questions to ministers. Ministers and government officials answer the questions. Apparently, interpellators speak longer than others, and officials speak very few times. The total number of turns is 296. Within each turn, utterances are not segmented.

Baseline System. The ASR system for the Congress meetings is based on our system originally developed for panel discussions[3]. Figure 1 shows an overview of the system. As the decoder, our Julius[5] is used.

In this work, the Corpus of Spontaneous Japanese (CSJ)[6] is used to train the speaker-independent acoustic model[7]. The CSJ consists of many oral presentations. We use 2,496 talks, and total amount of speech is 486 hours. The acoustic model contains approximately 8,000 triphones. The numbers of shared states and mixture components are 3,000 and 16, respectively. Speaking styles in the CSJ is spontaneous compared to those in other large-scale speech databases such as ASJ Japanese Newspaper Article Sentences (JNAS) and ATR phonetically balanced sentences (BLA). Actually, our preliminary experiment showed significant difference of WER (8% absolute) between the CSJ-based model and ATR-BLA model. For the acoustic model, MLLR-based speaker adaptation is performed using initial transcripts.

The language model is a mixture of topic-oriented *Minutes* model and speaking-style-oriented *Lecture* model, since there is no single corpus sufficiently covering these two linguistic aspects. Specifications of respective models are shown in Table 1. For the *Minutes* model, we collected the four-year minutes of the National Congress from the 145th ordinary session in 1999 to the 155th extraordinary session in 2002. All meetings including plenary sessions, committees and public hearing are used. Text of the minutes is

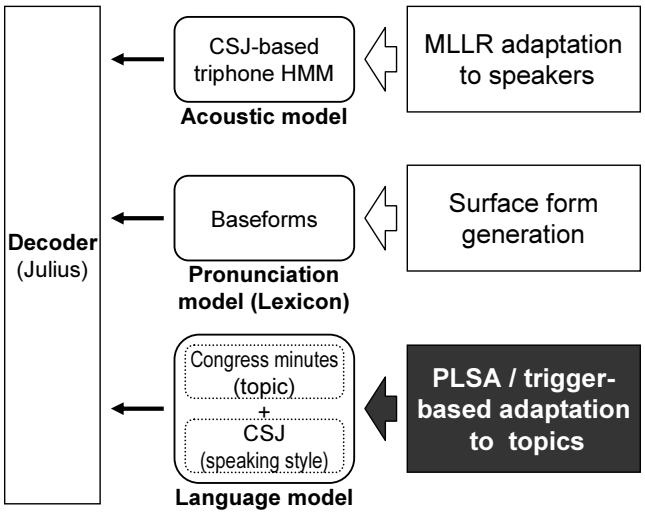

Figure 1. *Overview of the ASR system*

Table 1. *Specifications of language models*

Model	Minutes	Lecture	Baseline
Training corpus	Minutes of the National Congress (1999-2002)	Corpus of Spontaneous Japanese	—
#Words	70M	2.9M	—
#Uniq. words	72K	37K	—
#Documents	2,866	359	—
Vocab. size	29K	5.8K	30K
Perplexity	69.48	95.70	62.34
OOV rate	3.56%	9.43%	0.47%

basically faithful record of utterances except that fillers and disfluency are removed and some colloquial expressions are modified. For the *Lecture* model, the CSJ is used. We use part of the corpus by excluding presentations at academic meetings. Interjections, repairs and colloquial expressions are faithfully transcribed, while they are rarely contained in the minutes of the Congress.

The pronunciation lexicon contains standard Japanese pronunciations (baseforms) and spoken-style variants (surface forms) with their own pronunciation probabilities. These surface forms and probabilities are derived from baseforms using a statistical model of pronunciation variations[8].

LANGUAGE MODEL ADAPTATION BASED ON PLSA

The baseline language model is adapted to every speaker turn, as the topic may change every time. However, definition of the topic is indistinct, and it makes modeling of such characteristics difficult. We have proposed an adaptation scheme of language model taking speaking styles as well as topics into account[3]. The approach is based on probabilistic latent semantic analysis (PLSA)[9]. PLSA is a characterization of documents using a sub-space, where a document d is represented as a set of word occurrence probabilities

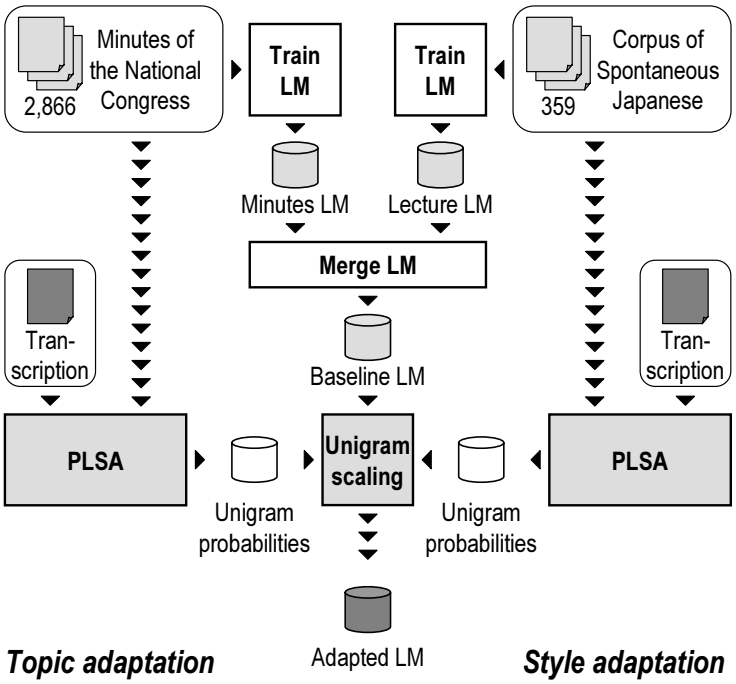

***Figure 2**. Overview of PLSA-based adaptation method*

$\{P(w|d)\}$ defined by (1);

$$P(w|d) = \sum_{j=1}^{N} P(w|t_j)P(t_j|d), \tag{1}$$

where t_j is an unseen variable known as a latent variable, and N is the total number of latent variables (i.e., dimensions of the sub-space). Probabilities $\{P(w|t_j)\}$ and $\{P(t_j|d)\}$ correspond to the base of the sub-space and to the coordinates of document d in the sub-space, respectively. These probabilities are estimated by EM algorithm using word occurrence statistics.

Figure 2 shows proposed adaptation method based on the PLSA framework. Since topic and speaker characteristics are covered by different corpora as shown in Table 1, PLSA is performed using each corpus, and respective sub-spaces are constructed. As for the minutes of the National Congress, documents are separated by the kind and date of meetings, and the total number of documents is 2,866. Meanwhile, the total number of talks in the CSJ is 1,245, and 359 documents are generated by concatenating all talks made by the same speaker. Consequently, speaking styles peculiar to individual speakers are expected to be mainly extracted by PLSA. Based on preliminary experiments, the numbers of latent variables were determined as 250 and 200 for topic-oriented and speaking-style-oriented PLSA, respectively.

Language model adaptation is done by projecting the initial transcription into the sub-spaces. The projection is performed only for unigrams; bigram and trigram probabilities are approximately calculated using the unigram scaling technique[10]. The estimation of probability for trigram $w_{i-2}w_{i-1}w_i$ is formulated in (2).

$$P'(w_i|w_{i-2}w_{i-1}) \propto \frac{P(w_i|d)}{P(w_i)} P(w_i|w_{i-2}w_{i-1}), \tag{2}$$

where $P(w_i)$ is a unigram probability in the baseline language model, and $P(w_i|d)$ is that obtained from PLSA. Finally, an adapted language model is generated by interpolating the topic-projected and speaking-style-projected models.

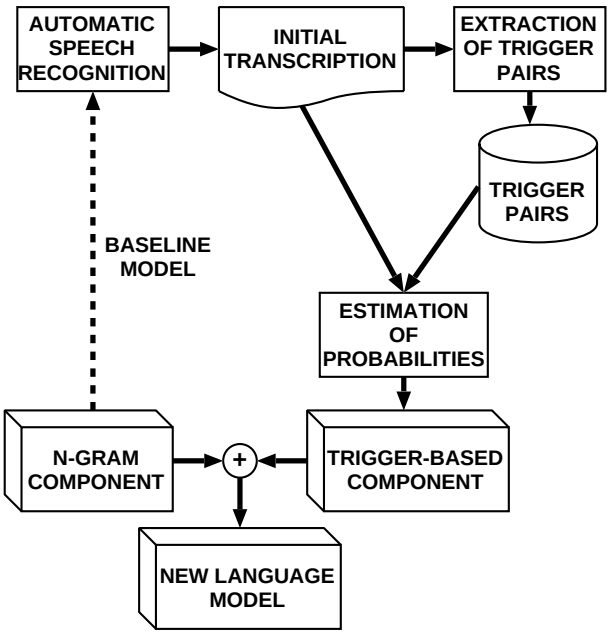

***Figure 3**. Overview of trigger-based adaptation method*

LANGUAGE MODEL ADAPTATION BASED ON TRIGGER MODEL

Since each turn of the meeting focuses on a particular topic, we expect to find topic-related words around the turn. In order to capture these long-distance dependencies, we have proposed to use the trigger-based language model adaptation[4]. We also incorporate this adaptation technique to the ASR system for the meetings.

Figure 3 illustrates the outline of the proposed approach. First, ASR is performed with a standard n-gram as the baseline language model, yielding the initial speech recognition results. The trigger pairs are then extracted and their probabilities are estimated from the initial transcription. Finally, the resulting trigger-based component is combined with the n-gram component to produce a new language model.

Extraction of trigger pairs from initial transcription. A trigger pair is a pair of content words that are semantically related to each other. Trigger pairs can be represented as $A \rightarrow B$, which means that the occurrence of A "triggers" the appearance of B, that is, if A appears in a document, it is likely that B will come up afterwards.

Task-dependent trigger pairs are extracted from the initial transcription, namely the K-best ASR hypotheses. For the selection of pairs, instead of the average mutual information (AMI) used in [11], we use the term frequency/inverse document frequency (TF/IDF) measure.

The TF/IDF value of a term t_k in a document D_i is computed as follows:

$$v_{ik} = \frac{tf_{ik} \log\left(N/df_k\right)}{\sqrt{\sum_{j=1}^{T}(tf_{ij})^2[\log(N/df_j)]^2}} \tag{3}$$

where tf_{ik} is the frequency of occurrence of t_k in D_i, N is the total number of documents, df_k is the number of documents that contain t_k, and T is the number of terms in D_i. In this work, we use a text window as the document unit.

We create all possible word pairs, including pairs of the same words (*self-triggers*), with the base forms of all content words with a TF/IDF value above a threshold. Part-of-speech (POS)-based filtering is introduced to discard function words.

Since the initial transcription contains errors, in order to minimize the adverse effect of erroneous trigger pairs, we introduce two methods to get rid of as many incorrect trigger pairs as possible. First, we use the confidence score that is provided by the ASR system to eliminate the trigger pairs whose component words have a confidence score lower than a threshold. Then, we compare the trigger pairs extracted from the initial transcription with pairs extracted from a large corpus, and we discard the trigger pairs that do not belong to the intersection of the two sets.

Probability estimation from initial transcription. The probabilities of the trigger pairs are estimated from the K-best ASR hypotheses by using a text window to calculate the co-occurrence frequency of the pairs inside it. Given a trigger pair $w_1 \rightarrow w_2$, this text window consists of the L words preceding w_2.

The probability of each trigger pair is computed as follows:

$$P_{TP}^{IT}(w_2|w_1) = \frac{N(w_1, w_2)}{\sum_j N(w_1, w_j)} \tag{4}$$

where $N(w_1, w_2)$ denotes the number of times the words w_1 and w_2 co-occur within the text window, and j runs throughout all words triggered by w_1.

Proposed trigger-based language model. The proposed trigger-based language model is then constructed by linearly interpolating the probabilities of the trigger pairs with those of the baseline n-gram model, so that both long and short-distance dependencies can be captured at the same time.

The probability of the proposed language model for a word w_i given the word history $H = w_{i-L}, \cdots w_{i-1} \triangleq w_{i-L}^{i-1}$ is computed in the following way:

$$P_{LM}(w_i|H) = \frac{1}{L} \sum_{j=i-L}^{i-1} P_{LM}(w_i|w_j) \tag{5}$$

$$P_{LM}(w_i|w_j) = \begin{cases} P_{NG}(w_i|w_{i-n+1}^{i-1}), \text{ if } P_{TP}^{IT}(w_k|w_j) = 0, \forall k \\ \lambda P_{NG}(w_i|w_{i-n+1}^{i-1}) + (1-\lambda) P_{TP}^{IT}(w_i|w_j), \text{ otherwise} \end{cases} \tag{6}$$

Here L is the number of words in the history H; P_{NG} is the probability of the n-gram component, which uses only the last $n-1$ words of H (i.e. $n \ll L$); P_{TP}^{IT} is the probability of the trigger-based component, computed by equation (2); and λ is the language model interpolation weight. When there are no words triggered by w_j, the n-gram model alone is used. Otherwise, the n-gram probabilities are linearly interpolated with the probabilities from the trigger pairs.

EXPERIMENTAL EVALUATION

We evaluated the two adaptation approaches incorporated to the baseline system. First, the effect of language model adaptation on perplexity was investigated. Figure 4 shows perplexity by the baseline model and adapted models. For comparison, we tested adaptation using manual transcription as well as the initial ASR result obtained by the baseline model. In case of manual transcription, average reduction of 11.3%, 7.4% and 16.8% were obtained by PLSA-based topic, style and both adaptation, respectively. Using the ASR result, perplexity was reduced by 8.7%, 5.7% and 13.4%, respectively. The figures indicate that both topic-oriented and style-oriented adaptation effectively reduce perplexity. As for trigger-based adaptation, reduction of 39.5% and 21.2% were obtained with manual and automatic transcription, respectively. Trigger-based adaptation realized higher performance than PLSA-based adaptation. However, the degradations with automatic transcription are relatively 20.2% and 46.3% in PLSA-based and trigger-based adaptation, respectively.

Adapted language models were also tested on ASR. In this ASR experiment, only automatic transcription was used for adaptation, and PLSA was performed for both topic- and speaking-style adaptation. Note

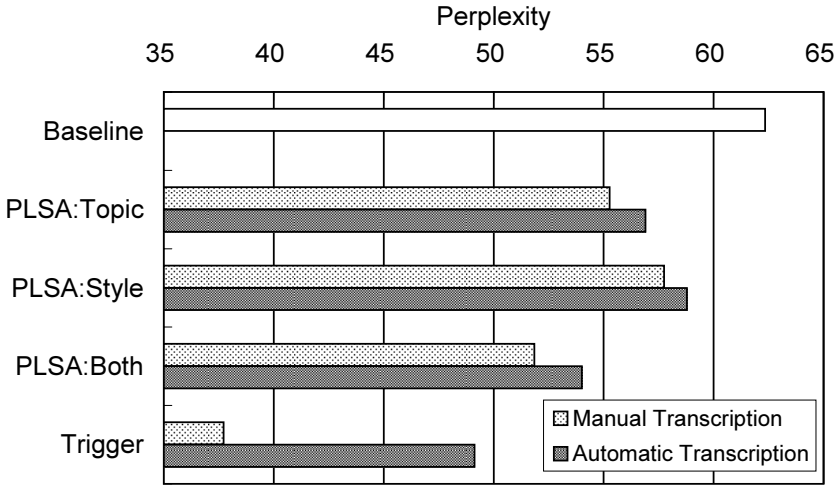

Figure 4*. Reduction of perplexity by language model adaptation*

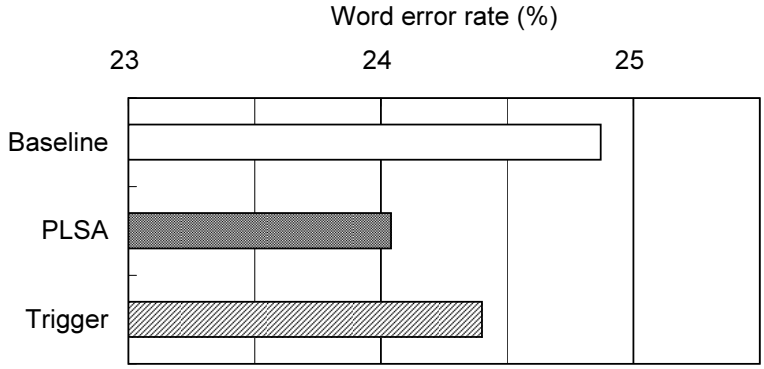

Figure 5*. Reduction of WER by language model adaptation*

that the trigger-based language model was applied by rescoring initial transcription, while the PLSA-based model was used in redecoding. Resulting word error rates (WER) are shown in Figure 5. WER by the baseline system was 24.9%. By PLSA-based and trigger-based adaptation, relative improvements of 3.3% and 1.9% were obtained, respectively. In this case PLSA-based adaptation outperformed trigger-based adaptation which yielded larger reduction of perplexity. The major reason of these results is that adverse trigger pairs derived from errors in automatic transcription cancelled improvement by correct trigger pairs. PLSA-based approach is more robust than trigger-based approach, since it is based on probabilistic mapping of transcription. Similar results were reported by Tam and Shultz[12], where adaptation based on cache model and LDA (Latent Dirichlet Allocation) were compared.

CONCLUSIONS

This paper presented two approaches of topic-oriented language model adaptation for transcription of meetings of the National Congress. In the Congress, multiple speakers talk in turn, and different topics are observed in each turn. Thus, we adopted PLSA-based and trigger-based adaptation to reflect contextual information to a language model. The former provides adapted probabilities of language model by mapping text to the probabilistic topic space. Speaking-style adaptation is also available in the same framework. The

latter emphasizes probability of a word based on occurrence of correlated words. These two approaches were evaluated and compared using real congressional speech, and trigger-based adaptation realized larger reduction of perplexity than PLSA-based adaptation. On the contrary, smaller WER was obtained by PLSA-based adaptation, because of its robustness against ASR errors.

ACKNOWLEDGMENTS

The authors are grateful to the House of Representatives of Japan for providing us with speech data of meetings and partially supporting this research.

REFERENCES

1. J.S. Garofolo, C.D. Laprun, and J.G. Fiscus, "The Rich Transcription 2004 Spring Meeting Recognition Evaluation," in *Proc. ICASSP Meeting Recognition Workshop*, 2004.

2. R. Prasad, L. Nguyen, R. Schwartz, and J. Makhoul, "Automatic Transcription of Courtroom Speech," in *Proc. ICSLP*, 2002.

3. Y. Akita and T. Kawahara, "Language Model Adaptation based on PLSA of Topics and Speakers for Automatic Transcription of Panel Discussions," *IEICE Transactions*, vol. E88-D, no. 3, pp. 439–445, 2005.

4. C. Troncoso and T. Kawahara, "Trigger-Based Language Model Adaptation for Automatic Transcription of Panel Discussions," *IEICE Transactions*, vol. E89-D, no. 3, pp. 1024–1031, 2006.

5. T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository," in *Proc. ICSLP*, 2004.

6. S. Furui, K. Maekawa, and H. Isahara, "Toward the Realization of Spontaneous Speech Recognition —Introduction of a Japanese Priority Program and Preliminary Results—," in *Proc. ICSLP*, 2000.

7. T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark Test for Speech Recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR*, 2003.

8. Y. Akita and T. Kawahara, "Generalized Statistical Modeling of Pronunciation Variations using Variable-length Phone Context," in *Proc. ICASSP*, 2005.

9. T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. SIG-IR*, 1999.

10. D. Gildea and T. Hofmann, "Topic-based Language Models using EM," in *Proc. Eurospeech*, 1999.

11. R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer, Speech and Language*, vol. 10, pp. 187–228, 1996.

12. Y.-C. Tam and T. Schultz, "Dynamic Language Model Adaptation using Variational Bayes Inference," in *Proc. Eurospeech*, 2005.