AUTOMATIC LECTURE TRANSCRIPTION BY EXPLOITING PRESENTATION SLIDE INFORMATION FOR LANGUAGE MODEL ADAPTATION

Tatsuya Kawahara   Yusuke Nemoto   Yuya Akita

Kyoto University, Academic Center for Computing and Media Studies
Sakyo-ku, Kyoto 606-8501, Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

The paper addresses language model adaptation for automatic lecture transcription by fully exploiting presentation slide information used in the lecture. As the text in the presentation slides is small in its size and fragmentary in its content, a robust adaptation scheme is addressed by focusing on the keyword and topic information. Several methods are investigated and combined; first, global topic adaptation is conducted based on PLSA (Probabilistic Latent Semantic Analysis) using keywords appearing in all slides. Web text is also retrieved to enhance the relevant text. Then, local preference of the keywords are reflected with a cache model by referring to the slide used during each utterance. Experimental evaluations on real lectures show that the proposed method combining the global and local slide information achieves a significant improvement of recognition accuracy, especially in the detection rate of content keywords.

Index Terms— speech recognition, language model, lectures, PLSA, cache model

1. INTRODUCTION

Targets of automatic speech recognition (ASR) have become broader to include a variety of human-to-human communications such as lectures and meetings. These kinds of audio recordings are increasingly archived and distributed via networked digital media. ASR will provide an index and metadata to enable efficient search and access to the speech materials.

Along with this direction, not a few projects have been conducted. In early 2000s, the Corpus of Spontaneous Japanese (CSJ)[1] was compiled. It contains a thousand of academic presentations at technical conferences, recorded using a close-talking microphone. We achieved a word error rate (WER) of around 20% with this corpus[2]. Recordings of oral presentations and seminars were also conducted in European projects such as TED corpus[3] and CHIL projects.

Classroom lectures at universities are also being digitally archived. Their automatic transcription and indexing are also studied at MIT[4] and Microsoft[5] under the iCampus project. Similar efforts are being conducted at Kyoto University and Tokyo Institute of Technology in Japan[6].

ASR technology is also expected to be useful for real-time captioning for deaf people. Currently, note-taking for deaf students in universities (at least in Japan) are being conducted by student volunteers, but these volunteers are not professional stenographers and cannot write down or type in every utterance of the lecture. Moreover, many lectures at universities are so technical that "out-of-field" volunteers cannot catch the content or technical words, for example, engineering students cannot help medical students. Therefore, an ASR system, combined with rapid adaptation of language model, is highly desirable. In the previous works[7][4], texts of relevant textbooks and technical papers were used to enhance N-gram language models.

With the pervasiveness of PC projectors and presentation software, more and more lecturers are getting to use presentation slides in classroom lectures in these years. Apparently, the presentation slides provide useful information for language model adaptation. First, the slide file, as a whole, provide a list technical keywords used in the lecture and topics covered by the lecture, which we refer to as global information. Moreover, we can obtain more precise prediction of content words in utterances by referring to the currently used slide, which we call local information.

On the other hand, slide information is limited in the text size and they contain only keywords and key phrases, but no carrier phrases. It is difficult to train a reliable N-gram language model with the text in the slides. Yamazaki et al. proposed language model adaptation by interpolating N-gram extracted from the slide text with a baseline model, but demonstrated only a little improvement[6]. Actually, our preliminary experiments showed the degradation in word accuracy with the simple N-gram interpolation scheme. Therefore, it is necessary to explore a more robust adaptation scheme by focusing on topic and keyword information.

In this paper, we investigate several methods to exploit slide information, both globally and locally, for effective and efficient language model adaptation for lecture transcription. Since we are targeting real-time captioning, we are more interested in efficient adaptation methods, which do not require much computation during the lectures nor re-decoding with the adapted model (re-scoring of an N-best list or word graph is feasible). Specifically, we introduce global topic adaptation based on PLSA and Web retrieval, and then reflect local word preference based on a cache model. Experimental evaluations were conducted with classroom lectures and tutorials given in Kyoto University.

2. GLOBAL TOPIC ADAPTATION BASED ON PLSA

PLSA (Probabilistic Latent Semantic Analysis)[8][9] is used to characterize documents in a corpus using word occurrence statistics. A topic sub-space, where each dimension represents some topic, is constructed by the expectation-maximization (EM) algorithm with a corpus of topic-annotated documents. The dimensions are optimally determined to distinguish documents in the training corpus. By projecting a document d (=slide text in this work) into this sub-space, a set of word occurrence probabilities $\{P(w|d)\}$ are obtained:

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

where $\{t_j\}$ are latent variables corresponding to dimensions, i.e., topics. N is a number of latent variables, i.e., dimensions of the sub-space.

Since this projection is based on a “bag-of-keyword” model, it is expected to robustly work with the slide text consisting of short key-phrases. In the PLSA framework, moreover, the probability $P(w|d)$ is estimated via latent topic variables t_j for words w which do not appear in the slide text d. This property will also realize robust adaptation of the language model with a small size of the slide text. The above equation provides only unigram estimates, so we apply them to a trigram model by the scaling technique[9][10]:

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

In this work, all text data of the slides used in the target lecture is used for adaptation, but only content words (nouns and verbs excluding numbers and pronouns) are extracted and their occurrence probabilities are adapted.

The adaptation is conducted offline given a whole slide text, before the ASR of lecture speech is conducted.

3. ADAPTATION WITH WEB TEXT

As an alternative method of offline language model adaptation, we have also investigated the incorporation of related Web text[11][12][13][14].

Characteristic keywords are extracted from the slide text to constitute a search query to Web. In this work, we selected three keywords based on the tf-idf measure from each slide, and made a query with them to collect Web text related with the slide. This process was iterated through all slides. In order to extract sentences, from the collected texts, which are matched in the style used in lectures, we proposed a sentence filtering method using the baseline language model[13].

Then, another language model is trained with the selected text, and interpolated with the baseline model to generate an adapted model.

4. LOCAL ADAPTATION WITH CACHE MODEL

In a cache model[15], preceding words in a history C, much longer than the N-gram model, are stored and their occurrence probabilities are heightened, assuming that they are more likely to be re-used. The cache model probability is obtained by

$$P(w_n|C) = \frac{1}{|C|} \sum_{w_h \in C} \delta(w_n, w_h) \tag{3}$$

where Kronecker delta δ becomes 1 when w_n matches w_h.

This probability is linearly interpolated with the baseline trigram model with some weight, and used for re-scoring an N-best list output by the baseline ASR system. In the orthodox cache model, the contextual history C is given by the initial ASR hypotheses. In this work, we extend this scheme so that the slide text used during the utterance is incorporated to the context C. Here, we assume that temporal information of the slide usage, i.e. which slide is used in a specific time period, is available.[1] In this manner, dynamic adaptation of the language model using the slide information is realized by predicting the words which appear in the corresponding slide.

5. EXPERIMENTAL EVALUATIONS

5.1. Database and Setup

The proposed adaptation methods have been evaluated with real lectures given at Kyoto University. We have recorded two kinds of lectures: one is ordinary classroom lectures in several courses of the Computer Science (CS) department, and the other is lectures of summer tutorial seminars on automatic speech recognition (ASR). As the test-set for this work, we used three lectures of different courses from the former category, and twelve lectures from the latter category. The speaking style of the latter category is more formal, since many of the lecturers and students came from outside of Kyoto University and the course is very intensive. Each lecture was given by different professors and lasted about 90 minutes, containing 14K to 22K words.

For the baseline ASR system, we used our Julius system (rev.3.5)[2]. The acoustic model was a tied-state triphone HMM

[1] This function is realized by several kinds of software.

[2] http://julius.sourceforge.jp/

Table 1. Evaluation with perplexity

	CS course	ASR tutorial
baseline	152.8	115.3
PLSA (slide; all words)	237.0	139.0
PLSA (**slide**; content words)	165.5	110.2
PLSA (utterance; ASR)	133.3	101.6
PLSA (utterance; oracle)	126.4	97.7
Web text	124.3	105.6

Table 2. Evaluation with word accuracy

	CS course	ASR tutorial
baseline	58.80	71.83
PLSA (slide)	59.41	72.40
Web text	60.50	72.37
cache (slide)	60.11	72.44
cache (ASR)	59.63	72.24
cache (slide+ASR)	60.30	72.66
PLSA + cache (slide)	60.68	72.98
PLSA + cache (ASR)	60.42	72.80
PLSA + cache (slide+ASR)	60.97	73.11

with 192K Gaussian components and trained with the academic presentation speech set (257 hours) of the CSJ. We conducted SAT and unsupervised MLLR adaptation. The baseline trigram language model was also trained with the CSJ (7M words) and the lexicon size was 50K words. Note that a number of talks on ASR in several technical conferences were included in the CSJ, which match the tutorial lectures in the test-set in terms of the topic.

Prior to ASR, all out-of-vocabulary (OOV) words which appeared in the slide text of the respective lectures were added to the lexicon, but this addition brought only a small WER improvement (0.2% absolute). This is referred to as the baseline in the following sub-sections.

5.2. Evaluation of Global Adaptation

First, we evaluated the global adaptation based on PLSA using the whole slide text. The topic sub-space of PLSA was generated using the lecture transcripts of the CSJ. The results of test-set perplexity are listed in Table 1. Initially, we used all observed words for PLSA, but the test-set perplexity was degraded by more than 20%. By limiting to content words for PLSA, a perplexity reduction is achieved for the ASR tutorials. In our previous study on PLSA-based language model adaptation using the initial transcription result[10], we observed the effect of adapting non-content words. However, in the slide text, key-phrases are mostly used and information on non-content words are hardly extracted. This property makes it necessary to select content words for applying PLSA to the slide text.

A similar adaptation method using the initial transcript output by the baseline ASR system realizes a greater reduction of perplexity, because it adapts not only content words but also carrier words. However, this scheme cannot be applied to real-time captioning since it essentially requires decoding whole lectures twice. The perplexity reduction using the correct transcript (oracle case) is also given for reference.

Next, we also conducted the adaptation based on Web text retrieval. The parameters such as interpolation weights were determined in a preliminary experiment. The method achieved a significant perplexity reduction, especially for the CS course lectures. Compared with the ASR tutorials, the topics of the CS courses are not well covered by the CSJ, thus effect of incorporating external texts becomes larger.

Then, we conducted an ASR evaluation. The word accuracy is summarized in Table 2. The results are consistent with the perplexity evaluation. The PLSA-based adaptation using the slide text realizes an improvement of 0.6% absolute for the ASR tutorial lectures, while the Web-based adaptation improves by 1.7% absolute for the CS course lectures.

5.3. Evaluation of Local Adaptation

Next, we implemented and compared several adaptation methods based on the cache model. Several parameters such as the history length $|C|$ in the initial ASR hypotheses and the linear interpolation weight with the baseline language model were determined with cross-validation by dividing the test-set of the ASR tutorials into two, and the same values were used in the CS course lectures. Specifically, the history length $|C|$ of 60 and the cache model weight of 0.1 were used.

The evaluation in word accuracy is included in the middle portion of Table 2. The method using the slide text as a cache realizes better accuracy than the conventional cache model using the contextual words in the ASR hypotheses. We also observed that the combination of these two methods gives a further improvement. These results confirm the effect of the local and dynamic adaptation of the language model.

5.4. Combination of Global and Local Adaptation

Then, we combined the global adaptation based on PLSA and the local adaptation based on the cache model. Here, we used the PLSA-based method rather than the Web text collection, because it is much faster and more convenient in many classrooms, where Internet access is not always available.

The results are summarized in the lowermost portion of Table 2. We confirmed a combined effect of the two methods, and achieved an improvement of the word accuracy by 2.2% and 1.3% absolute over the baseline, for the CS course lectures and the ASR tutorials, respectively.

Table 3. Evaluation with keyword detection rate (F-measure)

	CS course	ASR tutorial
baseline	70.87	85.28
PLSA (slide)	74.78	86.64
Web text	75.09	85.78
cache (slide)	75.92	86.01
cache (ASR)	73.07	86.40
cache (slide+ASR)	75.96	87.59
PLSA + cache (slide)	78.37	88.18
PLSA + cache (ASR)	76.76	87.39
PLSA + cache (slide+ASR)	78.60	88.24

5.5. Evaluation with Keyword Detection Rate

Detection of keywords is a more important measure than the simple word error rate when we consider the applications of ASR such as indexing used for document retrieval and speech summarization. It is also presumably important for human comprehension of the transcript.

Therefore, we made another evaluation using the keyword detection rate. Here, we define keywords by the content words (nouns and verbs excluding numbers and pronouns) that appear in the slide text. Then, we compute F-measure, which is a mean of the recall rate of keywords included in utterances and the precision of keywords detected in ASR results. The results are listed in Table 3. Compared with the baseline, the proposed method combining the global and local adaptation improves the detection rate by 7.5% and 3.0% absolute for the CS course lectures and the ASR tutorials, respectively. The improvement is realized primarily through the recall rate, and it is much larger than that of the simple word accuracy. Thus, the proposed methods are more effective in detection of keywords. Notice that, in this evaluation measure, the proposed methods based on PLSA and/or the cache model using the slide information gives comparable or even better accuracy than the Web text collection method, because these methods are more oriented to topic words.

6. CONCLUSIONS

We have investigated several language model adaptation methods which exploit presentation slide information for automatic lecture transcription. First, N-gram probabilities are re-scaled with lecture-dependent unigram probabilities estimated by PLSA using all slides of the lecture. Then, N-best hypotheses of the initial speech recognition results are rescored using word probabilities enhanced with a cache model using the slide corresponding to each utterance. Experimental evaluations on real lectures show that the proposed methods using the global and local slide information achieve significant improvements of recognition accuracy, especially in the detection rate of content keywords. Thus, the approach based on PLSA and the cache model is robust and effective.

We are designing and implementing a captioning system for lectures, and plan to conduct trials in our university.

Acknowledgment: The work is partially supported by Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communications, Japan.

7. REFERENCES

[1] S.Furui. Recent advances in spontaneous speech recognition and understanding. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 1–6, 2003.

[2] H.Nanjo and T.Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. Speech & Audio Process.*, 12(4):391–400, 2004.

[3] E.Leeuwis, M.Federico, and M.Cettolo. Language modeling and transcription of the TED corpus lectures. In *Proc. IEEE-ICASSP*, volume 1, pages 232–235, 2003.

[4] A.Park, T.Hazen, and J.Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *Proc. IEEE-ICASSP*, volume 1, pages 497–500, 2005.

[5] C.Chelba and A.Acero. Indexing uncertainty for spoken document search. In *Proc. INTERSPEECH*, pages 61–64, 2005.

[6] H.Yamazaki, K.Iwano, K.Shinoda, S.Furui, and H.Yokota. Dynamic language model adaptation using presentation slides for lecture speech recognition. In *Proc. INTERSPEECH*, pages 2349–2352, 2007.

[7] K.Kato, H.Nanjo, and T.Kawahara. Automatic transcription of lecture speech using topic-independent language modeling. In *Proc. ICSLP*, volume 1, pages 162–165, 2000.

[8] T.Hoffman. Probabilistic latent semantic indexing. In *Proc. SIG-IR*, 1999.

[9] D.Gildea and T.Hofmann. Topic-based language models using EM. In *Proc. EUROSPEECH*, pages 2167–2170, 1999.

[10] Y.Akita and T.Kawahara. Language model adaptation based on PLSA of topics and speakers. In *Proc. ICSLP*, pages 1045–1048, 2004.

[11] R.Sarikaya, A.Gravano, and Y.Gao. Rapid language model development using external resources for new spoken dialog domains. In *Proc. IEEE-ICASSP*, volume 1, pages 573–576, 2005.

[12] T.Ng, M.Ostendorf, M.-Y.Hwang, M.Siu, I.Bulyko, and X.Lei. Web-data augmented language models for mandarin conversational speech recognition. In *Proc. IEEE-ICASSP*, volume 1, pages 589–592, 2005.

[13] T.Misu and T.Kawahara. A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts. In *Proc. INTERSPEECH*, pages 9–12, 2006.

[14] Y.Akita, Y.Nemoto, and T.Kawahara. PLSA-based topic detection in meetings for adaptation of lexicon and language model. In *Proc. INTERSPEECH*, pages 602–605, 2007.

[15] R.Khun and R.De Mori. A cache-based natural language model for speech recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 12(6):570–583, 1990.