AUTOMATIC TRANSCRIPTION OF LECTURE SPEECH USING TOPIC-INDEPENDENT LANGUAGE MODELING

Kazuomi Kato *Hiroaki Nanjo* *Tatsuya Kawahara*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

We approach lecture speech recognition with a topic-independent language model and its adaptation. As lecture speech has its characteristic style that is different from newspapers and conversations, dedicated language modeling is needed. The problem is that, although lectures have many keywords specific to the topic and fields, available corpus of each domain is limited in size. Thus, we introduce topic-independent modeling with a vocabulary selection mechanism based on a mutual information criterion. It realizes better coverage and accuracy with small complexity than the conventional word frequency-based method. This baseline model is adapted to specific lectures using preprint texts. We have tried automatic transcription of oral presentations and achieved a word error rate of 23.6% on the average.

1. INTRODUCTION

Automatic transcription of lecture speech is significant both in research and applications. As a numbers of lectures and speeches in public are manually transcribed as a document, there are large demands for semi-automating the process. Lecture speech is regarded as an intermediate between read speech of newspaper corpora and conversational speech. Speakers of lectures use not only formal expressions but also colloquial ones, but not so casually as private conversations. Utterances contain disfluency especially filled pauses, but at least speakers try to be fluent. In some portions, it is similar to broadcast news, but the lecturers are not professional in speaking.

The other prominent feature is that lectures have specific topics, and the topics are often so technical that the vocabulary of one lecture is different from those of newspapers, daily conversations and even other lectures. For example, the vocabulary in ICSLP presentations is specific to the spoken language processing and different from other technical fields. This feature causes a serious difficulty in language modeling since it is not easy to collect lecture transcription data on specific topics large enough for training statistical models. In fact, the total size of available lecture corpora is much smaller than newspapers and broadcast news even if we ignore the difference in topics.

Therefore, we adopt an approach which first constructs a topic-independent language model and then adapts it to specific lectures to be transcribed. The effect of the topic-independent model and initial results of automatic transcription of oral presentations are demonstrated in this paper.

2. SYSTEM OVERVIEW

Automatic transcription of lecture speech is realized by the following steps as illustrated in Figure 1.

1. Train topic-independent language model

   This general model is to cover expressions dependent on lecture-style speech that is different from newspapers and conversations. It is also designed to be independent of specific fields and topics in order to make effective use of as many transcriptions of various lectures as possible.

2. Adapt language model to target lectures

   The language model is adapted to specific lectures to be transcribed. Specifically, it must predict keywords of relevant fields and topics as well as their linguistic statistics. In this work, keyword extraction and language model adaptation are done by using preprint papers of the lectures.

We adopted a similar approach in language model adaptation using MAP estimation for dialogue speech recognition of various topics[1]. However, the domain of lecture speech is much wider and the joint vocabulary size is much larger although keywords are so specific to topics. Thus, we introduce a vocabulary selection mechanism. Eliminating non-topic keywords will reduce recognition errors and improve efficiency.

3. TOPIC-INDEPENDENT LANGUAGE MODEL

3.1. Vocabulary Selection

We have proposed topic-independent modeling of filler words and demonstrated its effectiveness in key-phrase

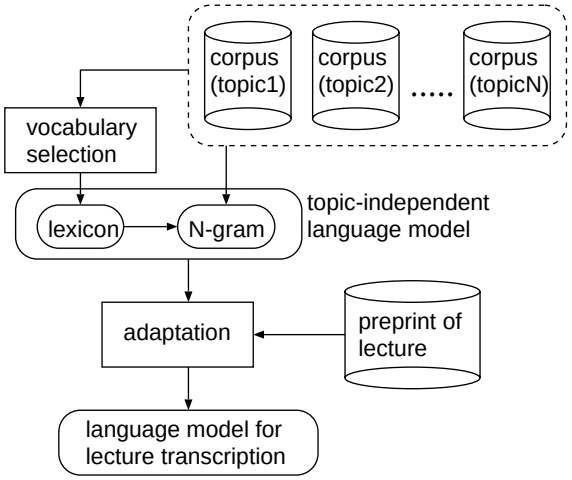

Figure 1: Overview of system

command verification for a voice-operated projector[2]. The model extracts filler words that are unique to lecture-style utterances and independent of specific topics. In this rather simple task, a vocabulary of hundreds of entries was proven to be sufficient.

The method is extended to the baseline language model for lecture speech recognition. It defines a vocabulary used in lectures by excluding keywords specific to topics. For the purpose, we adopt an information-theoretic criterion that is widely used for topic identification. Specifically, mutual information between a word w and topics T is computed. Suppose there are a set of topics $T = \{t_1, \ldots, t_n\}$, the mutual information $I(T; w)$ for a word w indicates non-uniformity of the frequency of the word w in various topics, or how much the word correlates with specific topics.

$$\begin{aligned} I(T;w) &= H(T) - H(T/w) \\ &= \sum_T P(t_i) \log \frac{1}{P(t_i)} - \sum_T P(t_i/w) \log \frac{1}{P(t_i/w)} \end{aligned}$$

Unlike topic identification, we pick up the words that appear in various topics universally, or whose $I(T; w)$ values are small. The resultant word set will give reasonable coverage to inputs of any domains and be robust against the change of topics. A similar approach to classify vocabulary words was taken on a broadcast news database[3]. The effect of topic-independent modeling is also seen in [4] as the general English model occupies about 90% in the topic-mixture model on broadcast news.

3.2. Training and Evaluation

For the training procedure, we simply use multiple corpora of different topics. Topic labels are not necessary since topic identification is not the purpose. As the training material, we have collected transcription of lectures and panel discussions available via World Wide Web. In total, 55 corpora covers various topics from information technology, medicine to politics. The total text size is 837K words and the number of different lexical entries amounts to around 32K.

Based on the selected lexicon, a word 3-gram model is trained with the original set of corpora.

For comparison, we also performed the conventional vocabulary selection based on the word frequency in the overall texts. When we pick up top 5000 words, 23% of the entries are not included by the other method. The proposed method based on the mutual information incorporates more entries of verbs and adjectives generally used in lectures, while the frequency-based lexicon has more nouns that can be regarded as topic words.

As a preliminary evaluation, recognition of lecture-style sentences from a portion of a television program is conducted. We used 115 utterances by three speakers in total. The language model is integrated with our recognition program Julius and gender-dependent triphone HMM of 2000 states and 16 mixture components trained with 132 speakers, which are available as the IPA Japanese dictation toolkit[5]. Lexical coverage, perplexity and word error rates are listed in Table 1 for the two models: one is based on the mutual information and the other by the word frequency. The vocabulary size of each lexicon is 5000 in this experiment. The proposed language model based on the mutual information criterion realizes better coverage, smaller perplexity and higher accuracy than the conventional method using the word frequency criterion. The result shows the effectiveness of our topic-independent modeling of lecture speech.

Other models of various sizes of vocabulary trained with newspaper corpus is also compared. When we use a language model trained with a newspaper corpus of 7 years, the coverage with 20K vocabulary is only 93.8% and the word error rate is 28.5% which is 1.5 times higher than our model. This suggests that a different model is needed for the lecture-style.

4. AUTOMATIC TRANSCRIPTION OF LECTURES

4.1. Model Adaptation

The baseline language model is adapted to lectures to be automatically transcribed. We make use of preprint papers of the lectures which we assume are available in an electronic form beforehand. The purpose of the process is to incorporate technical keywords into the lexicon and to adapt the word 3-gram model to the current topic. A specific language model is generated with the preprint papers and then merged into the topic-independent model.

Adaptation of N-gram statistics is realized by linear combination of the two models [6][3][7]. A probability of word w given a history h is defined as a sum of the prob-

Table 1: Coverage, perplexity and word error rate for lecture-style sentences

language model	lecture corpus		newspaper corpus		
	mutual information	word frequency	word frequency		
(size of vocabulary)	5K	5K	5K	20K	60K
coverage	95.5%	94.4%	85.4%	93.8%	99.5%
perplexity	64.9	66.6	140.3	130.2	144.6
word error (speaker1)	19.0%	21.5%	44.6%	29.6%	34.8%
(speaker2)	18.9%	20.8%	43.8%	27.7%	32.3%
(speaker3)	20.1%	22.1%	43.8%	28.2%	32.5%

Table 2: Lexical coverage by topic adaptation (%)

	lecture corpus	
	mutual information	word frequency
speaker	8K	8K
A	98.9 ← 93.3	98.6 ← 89.1
B	98.4 ← 90.8	97.8 ← 89.8
C	99.0 ← 90.9	98.7 ← 87.9

(adapted ← topic-independent)

Table 3: Comparison of word error rate (%)

	lecture corpus	
speaker	mutual information	word frequency
A	21.5	21.9
B	30.8	31.6
C	34.4	36.4

(vocabulary 8K)

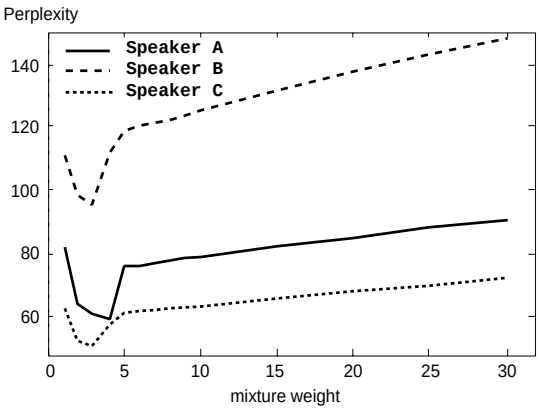

Figure 2: Test-set perplexity at various mixture weight (vocabulary 8K)

ability given by the baseline model $P_0(w|h)$ and that estimated with the preprint text $P_1(w|h)$, by normalizing with respective text sizes N_0, N_1 and weighting the adaptation text with a constant λ.

$$P(w|h) = \frac{N_0}{N_0 + \lambda N_1} P_0(w|h) + \frac{\lambda N_1}{N_0 + \lambda N_1} P_1(w|h)$$

Although the parameter λ can be estimated with the leaving-one-out method[7], in this work we perform experiments by changing the value of the only parameter.

4.2. Experiments

We have tried automatic transcription of oral presentations in our department. We pick up three speakers (A,B,C) who presented around 10 minutes using view-graphs. All presentations are somehow related with speech processing. Their preprint papers of 25 pages (=10K words) were available for language model adaptation. The same decoder and acoustic model are used as in the previous section. Through preliminary tests, it was found that a vocabulary size of 8K gives best performance.

Lexical coverage on oral presentations by the speakers (A,B,C) is shown in Table 2. New keywords in the preprint that are not covered by the baseline lexicon are added. Such examples include 'bigram' and 'morph'. Without adaptation, the lexicon based on the mutual information criterion achieves wider coverage than that of the conventional method. The adaptation process significantly improves the coverage to 98-99%. Final coverage by the both models is almost comparable through the adaptation process. Although the lexicon using the mutual information criterion has a lot of lecture-style words that are used in any lectures, many of such words are covered by the preprint paper.

Test-set word perplexity is shown in Figure 2 at various values of the weight λ on the preprint text. It is observed that the perplexity is reduced by 25% through the adaptation and gets minimum around $\lambda = 3$.

The comparison of word error rates by the mutual information criterion and the conventional method is shown in Table 3. The language model with the mutual information criterion achieves higher accuracy in all presentations. However, the difference is not significant as in the coverage after adaptation.

There is a problem in language modeling in relation to filled pauses. Usual archives of lectures that are publicly available are manually modified so that such disflu-

Table 4: Recognition results by incorporating filler model (%)

speaker	ratio of fillers	word error rate	
		adapted (λ=3)	filler added
A	4.9	21.5	18.0
B	7.6	30.8	26.4
C	13.7	34.4	26.4

(vocabulary 8K)

ency events are removed. Thus, the generated language model cannot cope with them. Therefore, we incorporate estimates of probabilities of typical fillers by referring a dialogue corpus. As shown in Table 4, filler words occupy 5-14% of the presentations. Thus, addition of them to the model brings out improvement of the accuracy by 3.5-8.0%.

The final recognition results for the three speakers (A,B,C) are given in Table 4. On the average, the error rate is 23.6%. The figure is in-between of that for broadcast news and that for conversational speech such as the Switchboard corpus. It is reasonable when we consider the acoustic and linguistic characteristics of lecture speech.

5. CONCLUSIONS

We have presented an initial trial of automatic transcription of lecture speech by focusing on its language modeling. The topic-independent model that is trained with various lecture corpora is demonstrated to realize better coverage and accuracy as a baseline model. The model is adapted with preprint papers to the topics and applied to recognition of oral presentations. Coverage of more than 98% and word error rate of 18-27% are achieved with the 8K lexicon after the adaptation. This fact suggests the feasibility of automatic transcription with the proposed framework.

Extensive data collection of lecture speech is being carried out by the project starting last year, so the model will be improved. With the function to attach confidence measures to recognition results[8], the system will be useful for semi-automatic transcription with post-processing by human.

Acknowledgment: The authors are grateful to Prof. Sadaoki Furui and other members of Science and Technology Agency Priority Program on “Spontaneous Speech: Corpus and Processing Technology”.

References

[1] H.Masataki, Y.Sagisaka, K.Hisaki, and T.Kawahara. Task adaptation using MAP estimation in n-gram language modeling. In *Proc. IEEE-ICASSP*, pages 783–786, 1997.

[2] T.Kawahara and S.Doshita. Topic independent language model for key-phrase detection and verification. In *Proc. IEEE-ICASSP*, pages 685–688, 1999.

[3] K.Seymore, S.Chen, and R.Rosenfeld. Nonlinear interpolation of topic models for language model adaptation. In *Proc. ICSLP*, pages 2503–2506, 1998.

[4] T.Imai, R.Schwartz, F.Kubala, and L.Nguyen. Improved topic discrimination of broadcast news using a model of multiple simultaneous topics. In *Proc. IEEE-ICASSP*, pages 727–730, 1997.

[5] T.Kawahara et al. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proc. ICSLP*, 2000.

[6] R.Kneser and V.Steinbiss. On the dynamic adaptation of stochastic language models. In *Proc. IEEE-ICASSP*, volume 2, pages 586–589, 1993.

[7] F.Wessel and A.Baader. Robust dialogue-state dependent language modeling using leaving-one-out. In *Proc. IEEE-ICASSP*, pages 741–744, 1999.

[8] T.Kawahara, C.-H.Lee, and B.-H.Juang. Flexible speech understanding based on combined key-phrase detection and verification. *IEEE Trans. Speech & Audio Process.*, 6(6):558–568, 1998.