A Method for Style Adaptation to Spontaneous Speech by Using a Semi-linear Interpolation Technique

Nobuyasu Itoh, Masafumi Nishimura, and Shinsuke Mori
IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.

Abstract

This paper deals with a method for adapting a language model created from written-text corpora to spontaneous speech by using a semi-linear interpolation technique. Sizes and topic coverages of spoken language corpora are usually far smaller those of written-text corpora. We propose an approach to adapt a base language model to the styles of spontaneous speech on the basis of the following assumptions. The words that are topic-independent, that is to say, common in spontaneous speech should be predicted mainly by a model created from spontaneous speech corpora (style model), while the base model is more reliable for predicting topic-related words, because they are difficult to predict from a model based on a small corpus. We classified all words into disfluencies and normal words. The normal words are classified into two more categories; common words and topic words according to mutual information. For each category, the qualified models (base or style) with the optimal weights for linear interpolation are selected. In other words, a different linear combination of the models is used for each category of a predicted word. We conducted experiments by using a spoken-language corpus of Japanese for creating the style model. We achieved 159.1 in test-set perplexity compared with the baseline of 189.3 (simple linear interpolation) and the perplexity of the style specific model, which was 230.7.

1 Introduction

An N-gram based language model is a typical statistical model which is widely used, particularly for speech recognition. In order to build a good language model, we need a set of learning data (a corpus) that is large enough to give accurate parameters. Large text corpora that cover broad areas are newspapers or broadcast news transcriptions. The former are available in many languages and are dominant in creating language models for current dictation systems. However, they are strictly written texts. Broadcast news transcriptions are obviously different in style from newspapers, and those of interviews and discussions have the typical characteristics of the spoken language domain[1]. But except for a few corpora (i.e. Broadcast News in English), the sizes are unsatisfactory. Since transcribing spontaneous speech accurately is requires so much time[1], the corpus sizes are far smaller than those of written texts even in English. For example, Switchboard [2], which is famous for its fully spontaneous speech corpus, has only approximately 2 million words. In particular, Japanese is far more time-consuming and difficult to transcribe than English, which is one of the reasons why the sizes and topic coverages of Japanese spontaneous speech corpora are strictly limited.

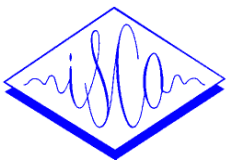
We believe that a general language model for spontaneous speech is a key to developing a transcription system for broader coverage such as lectures, interviews, and presentations. In this paper, we report an efforts towards creating a topic-independent language model for lecture transcription by adapting a base language model to the style of spontaneous speech in lectures. In Section 2, we describe an overview of our Japanese spontaneous-speech corpus, comparing it with others. In Section 3, we propose a method for adapting a base language model to the styles of spontaneous speech on the basis of a semi-linear interpolation technique. In Section 4, we present some experimental results. In the last section we discuss the efficiency of our method.

2 A spontaneous speech corpus

In Japan, some organizations (ATR, ASJ, and JEIDA[2]) provide spontaneous speech corpora[3]. However, since their focuses are dialogues on limited topics such as hotel reservations and queries on travel information, they are not sufficient even for the learning of styles. We therefore created a corpus of spontaneous speech from broadcast lectures.

We selected a total of 148 lectures broadcast by the University of the Air in 1998, transcribed them, and tokenized each sentence into words. Fig. 1 and Table 1 present a sample and some statistics about our corpus respectively, where the number of unique words does not include disfluencies[3]. Disfluencies are written down in phonetic spellings and with the tag of '<>'. A

[1] In BN, filled pauses were not transcribed.
[2] Japan Electronic Industry Development Association
[3] In Fig. 1, we modified some of the content words in order to avoid violating intellectual property laws



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Table 1: A broadcast-lecture corpus of Japanese

No. of subjects	78
No. of speakers	97
No. of words	1,098,888
No. of uniq. words	23,929
Disfluency	99,419 (9.1%± 4.0)
Filler	8.75%
Fragment	0.35%

d tag is added to mark word fragments. In addition, we also tagged prolonged sounds at word endings (i.e. wo<o:>), which are sometimes heard in Japanese. The reference [4] describes word units used in the corpus.

Shriberg[5] classified disfluencies on the basis of how the actual utterance must be modified to obtain the original (fluent) utterance. The three main types, filled pause, repetition, and deletion account for 85% of all disfluencies. According to this classification, "<e:>" and "<a:>", which play very similar role to "uh" or "um" in English, are filled pauses. Repetition means unintended addition of the same word, and deletion means a lack of a word which should be present in a fluent utterance. They also reported that in about 25% of repetitions and deletions word fragments, which do not compose a normal word, were observed in the Switchboard[2] corpus. In this work, we use the term *disfluency* in reference to interjectional words and phrases (*Filler* in Table 1) including filled pause, and word fragments. Other types of disfluency are not tagged. According to the statistics on our corpus, the rate of disfluencies is 9.1%, which supports the belief that our corpus mostly consists of spontaneous speech.

We made a preliminary examination on the coverage of disfluencies and other words (hereafter called *normal* words). Table 2 shows the word coverage of our base vocabulary for broadcast lectures except for disfluencies. The base vocabulary list was created mostly from newspaper articles and archives of on-line texts on net forums, and its size is approximately 75K. To compare the results to those of other sources of test data,

I *also* *this* *ceremony* *[case]*
<eto> 、 watashi mo 、 kono shiki wo<o:> 、

[case] *attend* *do*
<e:> ni sanka shite 、 <ano> *d*<kan> sanka

do *[tense]* *but* *people* *[case]*
itashi mashita 、 keredomo hitobito no 、 <sono:>

zeal *and* *Japan* *in* *be seen* *not*
、 netsui 、 sorekara nihon dewa mirare naku

[tense] *like* *school* *[case]* *at*
natta youna 、 <sono> 、 gakko ni oita ...

Figure 1: A sample of transcription

Table 2: Coverage of the base vocabulary (75K)

The University of the Air	98.6%
Nikkei newspaper	99.6%
Mainichi newspaper	98.7%

the coverages for some articles of two major Japanese newspapers are also listed. The table shows that the vocabulary list covers well the words that appeared in the broadcast lecture corpus and that the differences is slight. On the other hand, Fig. 2 shows the coverages of n most frequent disfluencies. Only a few of them cover the majority. For example, 24 disfluencies, which can be merged into 12 by ignoring ambiguities of transcription, account for 90% of all disfluencies.

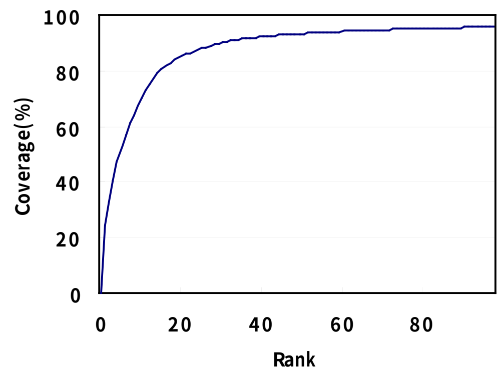


Figure 2: Coverage of disfluencies

3 Semi-linear interpolation techniques

Our broadcast lecture corpus includes many topics on various areas from literature to information science. However its size is only about 1.1M words, which is obviously insufficient for building a general language model for spontaneous speech. We therefore propose to apply semi-linear interpolation techniques to style adaptation. In other words, we aim at adapting the style of a base language model, which is created from large corpora of written texts, to spontaneous speech by using the broadcast lecture corpus. The basic idea is that all words are classified into the following categories, and that qualified models with the optimal weights for linear interpolation are selected. The categories are

- disfluencies
- words that frequently appear in broadcast lecture corpus independently of the topics (hereafter referred as *common* words)

- topic-related words (hereafter referred as *topic* words)

This category-based adaptation is therefore based on linear interpolation. But the total adaptation is nonlinear, which is the reason why we refer it as *semi-linear*.

3.1 Disfluency model

Stolcke[6] proposed a *clean-up* model for predicting speech disfluencies. In this model, disfluencies themselves are assumed to be normal-word events, though they do not help the prediction of the next word. Siu[7] also reported a similar model called *skip* N-gram. Siu treated each disfluency differently (extending the N-gram order *or* skipping the disfluency) depending on the function, and obtained a small improvement in perplexity.

On the other hand, in Japanese, the probabilities of disfluencies are usually estimated by heuristic methods instead of learning from a corpus. For example, Ohtsuki [8] treated disfluencies as commas or the beginning of a sentence in their language model.

In this work, we combine the base language model with that created from our broadcast-lecture corpus. The learning data of the base model are newspaper articles, where disfluencies almost never appear. Simple combination degrades the accuracy of the prediction of disfluencies rather than improves it. Thus, we tested the following approaches, where different models are used for predicting disfluencies and normal words, and compared them.

[Model 1] (Dis1)
Disfluencies are predicted by uni-gram probabilities that are learned from the broadcast lecture corpus, and normal words are predicted by the linear-interpolated model of the base tri-gram and broadcast lecture tri-gram, where all disfluencies are skipped in the word history. The following is the formula of this model. V_D denotes the set of disfluencies. P_{ua} and P_{base} designate that they are estimated from the broadcast lecture (the University of the Air) corpus and the newspaper corpora respectively. h and h_s means a word history and that with skipped disfluencies respectively. λ denotes an interpolation coefficient.

$$P(w_n \mid h) = \begin{cases} P_{ua}(w_n) & (\text{if } w_n \in V_D) \\ \lambda P_{ua}(w_n \mid h_s) + & \\ \quad (1-\lambda)\delta_{\overline{D}} P_{base}(w_n \mid h_s) & (\text{if } w_n \notin V_D) \end{cases}$$
$$\delta_{\overline{D}} = 1 - \sum_{w_n \in V_D} P_{ua}(w_n \mid h_s)$$

[Model 2] (Dis2)
Disfluencies are predicted by tri-gram probabilities that are learned from the broadcast lecture corpus. Normal words are predicted by the same formula as Model 1.

$$P(w_n \mid h) = P_{ua}(w_n \mid h_s) \quad (\text{if } w_n \in V_D)$$

[Model 3] (Dis3)
Normal words are predicted by the interpolated model of normal N-gram and skip N-gram. That is to say, the probabilities of normal words are calculated by the following formula, where γ is an interpolation coefficient. Disfluencies are predicted by the same formula as Model 2.

$$P(w_n \mid h) =$$
$$\lambda P_{ua}(w_n \mid h) + (1-\lambda)\delta_{\overline{D}} P_{base}(w_n \mid h_s) \ (\text{if } w_n \notin V_D)$$
$$where$$
$$P_{ua}(w_n \mid h) = \gamma P_{ua}(w_n \mid h_s) + (1-\gamma) P_{ua}(w_n \mid w_{n-1})$$

3.2 Category-based adaptation of normal words

In this section, we describe a method for adapting the probabilities of normal words to spontaneous speech. Seymore[9] proposed classifying a vocabulary list into three categories: General, On-topic, and Off-topic, and to use a different model for predicting the word of each category. However, their reported result using their proposed model is inferior to the results of simple linear interpolation. We therefore use the best-interpolated model for each of the word categories. The basic assumptions in our approach are that the words that are common in the broadcast lecture corpus should be predicted mainly by a model created from it, while the base model is more reliable for predicting topic words. The following is the formulated expression, where V_C and V_T designate *common* words and *topic* words respectively.

$$P(w_n \mid h) = \begin{cases} \lambda_C P_{ua}(w_n \mid h) + (1-\lambda_C)\delta_{\overline{D}} P_{base}(w_n \mid h) & \\ \quad w_n \in V_C & \\ \delta_T(\lambda_T P_{ua}(w_n \mid h) + (1-\lambda_T)\delta_{\overline{D}} P_{base}(w_n \mid h)) & \\ \quad w_n \in V_T \end{cases}$$

δ_T is the normalized coefficient and calculated by the following formula:

$$\delta_T(h) = \frac{1 - \sum_{w_n \in V_C} \lambda_C P_{ua}(w_n|h) + (1-\lambda_C)\delta_{\overline{D}} P_{base}(w_n|h)}{\sum_{w_n \in V_T} (\lambda_T P_{ua}(w_n|h) + (1-\lambda_T)\delta_{\overline{D}} P_{base}(w_n|h))}$$

Many research projects have been conducted on methods for measuring how much a word is used depending on various topics. Relative entropy, χ^2, and mutual

Table 3: Experimental results

Model	Perplexity
Lecture LM	230.7
Baseline LM	189.3
Dis1	185.4
Dis2	182.4
Dis3	176.5
Dis3 + CBA(293)	161.5
Dis3 + CBA(1296)	159.1
Dis3 + CBA(6018)	165.6

information are well known approaches. We adopted mutual information, as used by Kawahara[10].

$$(1) \qquad I(T;w) = -\sum_t P(t) log P(t) + \sum_t P(t \mid w) log P(t \mid w)$$

4 Experiments

Our base language model is created from a total of 193M words, most of which are newspaper articles. On the other hand, we divided the lecture speech corpus into 10 subsets, 9 of which are used as learning data. The remaining subset data is divided again into 2 subsets. One is used for estimation of interpolation coefficients, and the other one is for test data. The texts of a single lecture were not divided into learning and test data.

On the basis of formula (1), we selected common words from the learning data of the 9 subsets. The 30 disfluencies with the highest frequencies were added to the base vocabulary list (75K). Any normal words that were out-of-vocabulary were not added to the list.

Table 3 shows our experimental results. Lecture LM, CBA, and the parenthetic numbers refer to the lecture specific model created only from the broadcast lecture corpus, the category-based adaptation of normal words, and the number of words selected as *common*. Baseline LM is that created by simple linear interpolation of two models, the base and the lecture specific model. The difference of the model Dis1 and Dis2 is whether disfluencies are predicted by uni-gram *or* tri-gram estimation. The result suggests that disfluencies are also context dependent, but the improvement is small. Comparing the results of Baseline LM and Dis2, we found the skip N-gram model is slightly better than the baseline model. Dis3, which is an extension of the model Dis2 (skip N-gram), improved the perplexity further. This result supports the view that disfluencies are also useful for predicting the next normal word, which is a conclusion obtained from some experiments in English. The category-based adaptation of common word probabilities improved the best case of the disfluency models by approximately 15%.

5 Conclusion

In this work, we have described a method for adapting a base language model to the style of spontaneous speech in lectures by using a lecture-speech corpus. In our method, words are classified into categories, disfluencies, common words and topic words, and the best combination of the two models is used for each of them. The perplexity can be reduced to 159.1 from 189.3 (baseline). We also compared several models for disfluencies. Our combined model (Dis3), which is one of the extended skip N-gram models, achieved the best results of all of them.

Acknowledgment

The authors are grateful to The Mainichi Newspaper Co. (CD Mainichi 91-95), Nippon Keizai Shimbun, Inc., and the University of the Air.

References

[1] Iyer, R. and Ostendorf, M., "Relevance weighting for combining multi-domain data for n-gram language modeling," *Comp. Speech and Lang.*, Vol. 13, pp. 267-282, (1999).

[2] Godfrey, J.J. et al., "Switchboard: Telephone speech corpus for research and development," *Proc. of ICASSP*, pp. 517-520, (1992).

[3] Yamamoto, M., "Current status of construction of spoken dialog database," *Journal of the Acoustical Society of Japan*, Vol. 54, No. 11, pp. 797-802, (1998) in Japanese.

[4] Itoh, N. et al., "A word-based Japanese language model," *Journal of Natural Language Processing*, Vol. 6, No. 2, pp. 9-27, (1999) in Japanese.

[5] Shriberg, E., "Preliminaries to a theory of speech disfluencies," Ph. D. thesis, University of California at Berkeley, (1994).

[6] Stolcke, A. and Shriberg, E., "Statistical language modeling for speech disfluencies," *Proc. of ICASSP*, pp. 405-408, (1996).

[7] Sui, M. and Ostendorf, M., "Modeling disfluencies in conversational speech," *Proc. of ICSLP*, Vol. 1, pp. 386-389, (1996).

[8] Ohtsuki et al., "Improvements in Japanese broadcast news transcription," *Proc. of DARPA Broadcast News Workshop*, pp. 231-236, (1999).

[9] Seymore, K. et al., "Nonlinear interpolation of topic for language model adaptation," *Proc. of ICSLP*, pp. 2503-2506, (1998).

[10] Kawahara, T. and Doshita, S., "Topic independent language model for key-phrase detection and verification," *Proc. of ICASSP*, pp. 685-688, (1999).