Construction of Spoken Language Model Including Fillers Using Filler Prediction Model

Kengo Ohta¹, Masatoshi Tsuchiya², Seiichi Nakagawa¹

¹Department of Information and Computer Sciences / ²Information and Media Center,
Toyohashi University of Technology, Japan
kohta@slp.ics.tut.ac.jp, tsuchiya@imc.tut.ac.jp, nakagawa@slp.ics.tut.ac.jp

Abstract

This paper proposes a novel method to construct a spoken language model including fillers from a corpus including no fillers using a filler prediction model. It consists of two sub-models: a filler insertion model which predicts places where fillers should be inserted, and a filler selection model which predicts appropriate fillers for given places. It converts a corpus that covers domain-relevant topics but includes no fillers into a corpus that contains fillers as well as domain-relevant topics. The experiment against the corpus of spontaneous Japanese shows that language models constructed by the proposed method achieve quite near performance of the traditional Japanese shows that language models constructed by the proposed method achieve quite near performance of the traditional

Index Terms: Spontaneous Speech, Filler, Language Model

1. Introduction

Increasing digital audio archives of lectures, presentations and meetings over various domains requests an effective indexing technology against them. Automatic speech recognition (ASR) against spontaneous speech is its essential part, because transcriptions of recorded data are necessary for indexing. Standard modern ASR systems depend on statistical language models with large vocabulary, therefore, construction of a statistical language model that covers spoken-style expressions as well as domain-relevant topics is necessary. The simplest approach to construct such a model is to train it from a large scale corpus that consists of many faithful transcriptions of spontaneous speech in the relevant domain. However, amount of available corpus is usually limited because its building task is quite expensive.

In order to cope this problem, several approaches have been proposed. A typical approach is combination of a spoken language model that covers general spoken-style expressions and a written language model that covers domain-relevant topics. For example, combination of conversational telephone speech corpus with meeting or Web corpora is adopted in meeting speech recognition[1]. As a similar approach, several methods directly manipulating N-gram probabilities such as cache models[2] were proposed. Class-based language models[3] are also used for robust estimation of N-gram probabilities with limited or unmatched data. Akita and Kawahara[4] proposed the other approach that transforms the N-gram model trained from a written-style corpus of the target domain into the N-gram model covering spoken-style expressions using a probabilistic transformation model trained from a parallel aligned corpus of the faithful transcriptions and their written-style texts. However, it is quite difficult to obtain such aligned corpus.

In this paper, we propose a new approach of this problem: using inexact transcribed corpora like shorthand notes. Inexact transcribed corpora are widely produced as shorthand notes or meeting records, and more available than faithful ones because their inexactness reduces transcription costs. And more, they can become more appropriate resource to train a spoken language model that covers domain-relevant topics than written-style corpora, because they include more spoken-style expressions in the target domain than written-style corpora. Of course, inexact transcribed corpora lack almost all disfluencies unlike faithful spoken corpora, therefore, it is necessary to restore disfluencies into them before training a spoken language model that covers disfluencies from them.

We, especially, focus on fillers among several disfluency acts, because fillers occur much more than other disfluency acts, like word fragments and incorrect or reduced pronounciations[5, 6]. This paper explains a method using a filler prediction model trained from a corpus that includes fillers but do not cover domain-relevant topics to restore fillers into inexact transcribed corpora in the target domain.

The remainder of this paper is organized as follows: Section 2 formalizes the filler prediction model. The procedure to build a language model using it will be described in Section 3. Experiments presented in Section 4 shows that the proposed method can construct language models from an inexact transcribed corpus that achieves near performance of the model trained from the real faithful corpus for the target domain. This paper concludes with concluding remarks in Section 5.

2. Filler Prediction Model

This section investigates a procedure to construct a statistical language model including fillers from an inexact transcribed corpus excluding fillers.

Two possible approaches can be supposed. The first approach consists of two steps: constructing a model which does not cover fillers from the inexact transcribed corpus and transforming it into a model including fillers using a transformation model. The second approach is transforming a corpus excluding fillers into a corpus including fillers before constructing a model. Suppose the following two sentences:

(1) This display shows that · · ·
(2) This uh display shows that · · ·

String (1) is an example sentence of the inexact transcribed corpus, and String (2) is the sentence in which a filler denoted by an underline is restored. Once a corpus in which fillers are restored like String (2) is obtained, it is quite easy to construct a model from it.

While the transformation model used in the first approach heavily depends on the structure of the target language model,
the language modeling and the corpus transformation in the second approach are clearly separated. This separation makes the second approach enable to accept advances in language modeling research area easily. Because of this advantage, the second approach will be investigated in this paper.

The second approach requires a filler prediction model that predicts both places where fillers should be inserted and filler types for the selected places. Analysis of [6] shows that many various fillers appear in real spontaneous speech. In order to avoid a data sparseness problem caused by those various fillers, we introduce an assumption that prediction of filler places and selection of filler types are independent. It means that our proposed filler prediction model consists of two sub-models: a filler insertion model which predicts places where fillers should be inserted, and a filler selection model which predicts appropriate fillers for the given places.

2.1. Filler Insertion Model

A filler insertion model predicts places where fillers should be inserted when a certain word sequence is given. In this paper, this model is formalized as a sequence labeling problem as shown in Fig. 1. The label F means that a filler should be inserted immediately after the labeled word, and the label O means contrary.

<table>
<thead>
<tr>
<th>Word sequence</th>
<th>This display shows that · · ·</th>
</tr>
</thead>
<tbody>
<tr>
<td>(BOS) pron noun verb pron</td>
<td></td>
</tr>
<tr>
<td>Label sequence</td>
<td>[F O O O O · · ·]</td>
</tr>
</tbody>
</table>

Figure 1: Example of Filler Insertion Labeling

In this paper, Conditional Random Fields (CRF)[7] is employed against this labeling problem. CRF is a discriminative probabilistic model that offers several advantages over hidden Markov models, and is employed to several statistical natural language processing tasks including language modeling[8].

When a certain word sequence \( X \) is given, the conditional probability of a label sequence \( Y \) is defined as follows:

\[
P(Y|X) = \frac{1}{Z(X)} \exp \left( \sum_{i=1}^{n} \sum_{a} \lambda_a f_a(X_i, Y_i) \right),
\]

where \( f_a \) is a feature function, \( \lambda_a \) is the weight of the feature function, and \( Z(X) \) is the normalization factor.

2.2. Filler Selection Model

A filler selection model predicts appropriate fillers for given places. We simply employ the conditional distribution of fillers over contexts as such a model.

Witten-Bell discounting[9] is employed to estimate the conditional probability \( P_s(f|h) \) of the filler \( f \) given the context \( h \) as follows:

\[
P_s(f|h) = \begin{cases} \frac{c(h,f)}{c(h)+\nu(h)} & \text{if } c(h,f) > 0 \\ \frac{r(h)}{c(h)+\nu(h)} \times P_s(f|h') & \text{otherwise} \end{cases},
\]

where \( c(h,f) \) is the frequency that the filler \( f \) occurs in the context \( h \), \( c(h) \) is the frequency of the context \( h \), \( r(h) \) is the number of different fillers which appear immediately after the context \( h \), and \( h' \) is the reduced context of \( h \).

3. Procedure to Build Language Model Using Filler Prediction Model

This section describes the procedure to build a language model using a filler prediction model. The proposed method roughly consists of two steps. The first step is training a filler prediction model based on a faithful training corpus that does not belong to the target domain. The second step is transformation from an inexact transcribed development corpus of the target domain to a corpus including fillers using a filler prediction model. Once it is obtained, building a statistical language model of the target domain from it is quite easy.

Our assumption against the filler prediction model described in Section 2 divides training of the filler prediction model into two sub-steps: training of the filler insertion model and training of the filler selection model. Each training instance of the filler insertion model is composed of a word sequence excluding fillers and its label sequence like Fig. 1, and such instances are easily obtained from the faithful training corpus. The CRF that represents the filler insertion model is trained with the toolkit CRF++, which uses LBFGS to estimate parameters and uses Gaussian prior to avoid an overfitting problem. The filler selection model is obtained from \( \lambda \)-gram statistics of the faithful training corpus as shown in Equation (2).

Because the inexact transcribed development corpus includes no fillers, it is necessary to restore fillers to it before building a statistical language model including fillers. Its procedure also consists of two steps: predicting places of fillers with the filler insertion model, and selecting fillers for the selected places with the filler selection model. Suppose the index \( i \) ranges over all words of the development corpus. The probability that a filler should be inserted immediately after the \( i \)-th word has the form

\[
P(y_i = F|X) = \sum_{(Y|y_i=F)} P(Y|X).
\]

When a uniform random variable \( Q_i \) (\( 0 \leq Q_i < 1 \)) meets the condition \( Q_i < P(y_i = F|X) \), a filler is inserted immediately after the \( i \)-th word. Once a filler insertion place is selected, the filler selection model is employed to select an appropriate filler. When a uniform random variable \( Q'_i \) (\( 0 \leq Q'_i < 1 \)) meets the equation

\[
\sum_{j=1}^{k-1} P_s(f_j|h_i) \leq Q'_i < \sum_{j=1}^{k} P_s(f_j|h_i)
\]

where \( j \) is the index of all possible fillers and \( h_i \) is the context around the \( i \)-th word, the filler \( f_k \) is selected as an appropriate filler for the given position.

The final product of the proposed method is 3-gram model trained from the corpus generated by the above procedure. It is expected that the model will match the target domain and will have a power to handle fillers because the generated corpus matches the target domain and contains fillers. Two random variables are introduced into the above procedure to simulate human’s filler arising process that is not deterministic, but is obviously stochastic. Because of these random variables, various fillers inserted corpora are generated by the proposed method even if the same training corpus and the same development corpus are given. So, the average of 10 trials will be reported as the experimental result of the proposed method in Section 4.
### Table 1: Statistics of Experimental Data

<table>
<thead>
<tr>
<th></th>
<th>Training Corpus</th>
<th>Development Corpus</th>
<th>Test Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain</strong></td>
<td>SPS</td>
<td>APS</td>
<td>APS</td>
</tr>
<tr>
<td><strong># of lectures</strong></td>
<td>900</td>
<td>400</td>
<td>50</td>
</tr>
<tr>
<td><strong>Speech length (hour)</strong></td>
<td>174.4</td>
<td>114.2</td>
<td>16.0</td>
</tr>
<tr>
<td><strong># of sentences</strong></td>
<td>263K</td>
<td>153K</td>
<td>22K</td>
</tr>
<tr>
<td><strong># of words</strong></td>
<td>1,778K</td>
<td>1,211K</td>
<td>170K</td>
</tr>
<tr>
<td><strong>Vocabulary size</strong></td>
<td>34K</td>
<td>21K</td>
<td>8K</td>
</tr>
<tr>
<td><strong># of fillers</strong></td>
<td>91K</td>
<td>79K</td>
<td>11K</td>
</tr>
<tr>
<td><strong>Filler ratio</strong></td>
<td>5.1%</td>
<td>6.5%</td>
<td>6.7%</td>
</tr>
</tbody>
</table>

SPS: simulated public speech, APS: academic presentation speech.

### 4. Evaluation

#### 4.1. Evaluation Metrics

Both the test-set perplexity and the adjusted test-set perplexity are employed to evaluate language models constructed by the proposed method. When a test corpus \( w^n \) is given, the cross entropy \( H \) of the language model \( L \) has the form

\[
H(L) = -\frac{1}{n} \log_2 P_L(w^n), \tag{4}
\]

where \( P_L(w^n) \) is the probability that the test corpus \( w^n \) occurs over the language model \( L \), and the test-set perplexity \( PP \) is defined as follows:

\[
PP = 2^{H(L)} \tag{5}
\]

The adjusted test-set perplexity described in [10] is the improved evaluation metric to consider OOVs in the test corpus, and is defined as follows:

\[
\log_2 PP^* = \log_2 PP + \frac{o}{n} \log_2 m, \tag{6}
\]

where \( o \) is the number of OOVs in the test corpus, and \( m \) is the number of the kind of OOVs in the test corpus.

#### 4.2. Experimental Data

In order to evaluate our proposed method, two corpora belonging to different domains are required. As such corpora, we use Corpus of Spontaneous Japanese (CSJ)\[6\], a large-scale database of spontaneous Japanese. It contains two main sources of spontaneous speech: academic presentation speech (APS) and simulated public speech (SPS). APS is the live recording of academic presentation in 9 different academic societies covering the fields of engineering, social science, and humanities. SPS, on the other hand, is studio recording of layman speaker’s speech of about 10–12 minutes, on everyday topics like ‘the most delightful/saddest memory of my life’.

Three subset corpora are extracted from CSJ for experiments. The training corpus is used to train a filler prediction model, and consists of 900 lectures randomly selected from SPS lectures. The development corpus is used to simulate an inexact transcribed corpus. 400 lectures are randomly selected from SPS lectures, and all fillers are removed from them before using them as the development corpus. The test corpus consists of 50 lectures randomly selected from APS lectures, and is used to evaluate the proposed method. Their statistics are described in Table 1.

Table 2 shows the difference between APS and SPS in their belonging domains. The right columns of Table 2 shows the

#### 4.3. Evaluation of Filler Insertion Model

At first, experiments that all fillers are considered as one category is conducted, in order to evaluate the filler insertion model by itself. The vocabulary size of all models is 20K. Table 3 summarizes the set of features of CRF and shows their results. The language model in the bottom line of Table 3 is trained from the faithful development corpus including fillers. It means that this model is inducted by existing fillers, so, we regard its performance as the upper bound of the proposed method. Two baseline models are also prepared: the unigram insertion model trained from the training corpus, and the trigram insertion model trained from the same corpus. Especially, the performance of the unigram insertion model is regarded as the lowest baseline because it considers a filler arising process as purely context independent stochastic process.

Table 3 reveals that the CRF which refers the preceding 2 words, the current word, and the succeeding 2 words as features is quite effective as a filler insertion model, and that a language model using it achieves quite near performance to the upper bound, and achieves better performance than the baselines’ ones. Table 3 also reveals that the CRF which ignores surface strings of domain-specific words such as nouns, verbs, and adjectives achieves the similar performance of the CRF which refers them as features. It suggests that the filler insertion model is independent from domains.

Figure 2 plots the change of \( PP \) and \( PP^* \) when training the CRF which refers preceding 2 words, the current word, and succeeding 2 words as features with different size of training corpus. We think that the currently available corpus contains enough instances to train a filler insertion model, because the decrease in \( PP \) seems to saturate with the maximum size of training corpus.
4.4. Evaluation of Filler Selection Model

The performance effect of considering contexts in filler selection is revealed at Table 4. The best model of Table 4, which considers contexts both in filler insertion and in filler selection, achieves a near performance to the upper bound.

Although $P^+$ increases 10.9% when the filler insertion model is changed from the CRF to the unigram model, $P^*$ increases only 1.3% when the filler selection model is changed from the trigram model to the unigram model. This suggests that the estimation of filler positions is more important than the selection of appropriate filler among all possible fillers.

5. Conclusions

This paper proposes a novel method to construct a spoken language model including fillers from an inexact transcribed corpus excluding fillers using a filler prediction model. The filler insertion model is changed from the CRF to the unigram model, $P^*$ increases only 1.3% when the filler selection model is changed from the trigram model to the unigram model. This suggests that the estimation of filler positions is more important than the selection of appropriate filler among all possible fillers.

6. References


