EFFICIENT ESTIMATION OF LANGUAGE MODEL STATISTICS OF SPONTANEOUS SPEECH VIA STATISTICAL TRANSFORMATION MODEL

Yuya Akita *Tatsuya Kawahara*

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

ABSTRACT

One of the most significant problems in language modeling of spontaneous speech such as meetings and lectures is that only limited amount of matched training data, i.e. faithful transcript for the relevant task domain, is available. In this paper, we propose a novel transformation approach to estimate language model statistics of spontaneous speech from a document-style text database, which is often available with a large scale. The proposed statistical transformation model is designed for modeling characteristic linguistic phenomena in spontaneous speech and estimating their occurrence probabilities. These contextual patterns and probabilities are derived from a small amount of parallel aligned corpus of the faithful transcripts and their document-style texts. To realize wide coverage and reliable estimation, a model based on part-of-speech (POS) is also prepared to provide a back-off scheme from a word-based model. The approach has been successfully applied to estimation of the language model for National Congress meetings from their minute archives, and significant reduction of test-set perplexity is achieved.

1. INTRODUCTION

Recently, main targets of automatic speech recognition (ASR) research have shifted to spontaneous speech such as meetings and conversation. One of major issues in spontaneous speech recognition is building an appropriate language model covering spoken-style expressions as well as domain-relevant topics. For better modeling, large amount of well-matched data is needed. However, amount of available spoken-style text, especially faithful transcript, is usually limited because of transcription costs. Therefore, efficient modeling utilizing such limited data is required. On the other hand, there are often large relevant text databases available, which are not faithful transcript but made for documentation. These include proceedings of lectures, minutes of meetings or closed captions for broadcasts.

The conventional approach of language modeling for spontaneous speech recognition is combination of these document-style text databases with some spontaneous speech corpus. For example, written text from textbooks or proceedings of lectures is complemented by transcripts of conversational telephone speech (CTS) of Switchboard and Fisher corpora for automatic transcription of lecture speech[1, 2]. Also, combination of CTS corpus with meeting or Web corpora is adopted in meeting speech recognition[3, 4]. Respective corpus represents specific linguistic characteristics such as topics and styles. Though these synthesis-based approaches are simple and effective, resulting model also contains irrelevant linguistic expressions which may degrade ASR performance. Another approach, proposed by Schramm et al.[5], generates simulated spoken-style text by randomly inserting fillers into written-style text. However, this approach handles only fillers, and contextual statistics are not reliably estimated.

In this paper, we propose a transformation approach which estimates language model statistics (N-gram counts) of spontaneous speech from a document-style large corpus using transformation model. The transformation model, which is widely used in statistical machine translation (SMT), is trained with small amount of parallel aligned corpus of these two styles. The proposed approach is applied to efficiently construct a language model of meetings, by estimating statistics of "virtual" faithful transcripts from a large amount of documented minutes.

2. TRANSFORMATION BASED ON STATISTICAL MACHINE TRANSLATION FRAMEWORK

Statistical machine translation[6] generates sentence Y of target language from sentence X of source language, which maximizes posterior probability $P(Y|X)$ based on Bayes' rule.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

Here $P(X)$ is not actually used because it does not affect choice of Y for given X. $P(X|Y)$ is usually computed by translation model.

In this work, we consider document-style and spoken language as different ones, denoted by X and Y, respectively, and try to estimate spoken language model $P(Y)$, which is represented as Equation (2) by modifying Equation (1).

$$P(Y) = P(X)\frac{P(Y|X)}{P(X|Y)} \tag{2}$$

Table 1. Major differences between spontaneous speech and document-style text

Type	$P(Y\|X)$	$P(X\|Y)$	Document-style text X	→	Spoken language Y
Insertion of fillers	Estimate	1	I think ...	→	(pause) **ah** I think ...
Deletion of postpositional particles	Estimate	Estimate	*watashi* ***wa*** *omoimasu* (I / my / me) (think) "*wa*" indicates the nominative case (= I)	→	*watashi omoimasu*
Substitution of colloquial expressions	Estimate	1	***iroiro na*** (various)	→	***ironna***

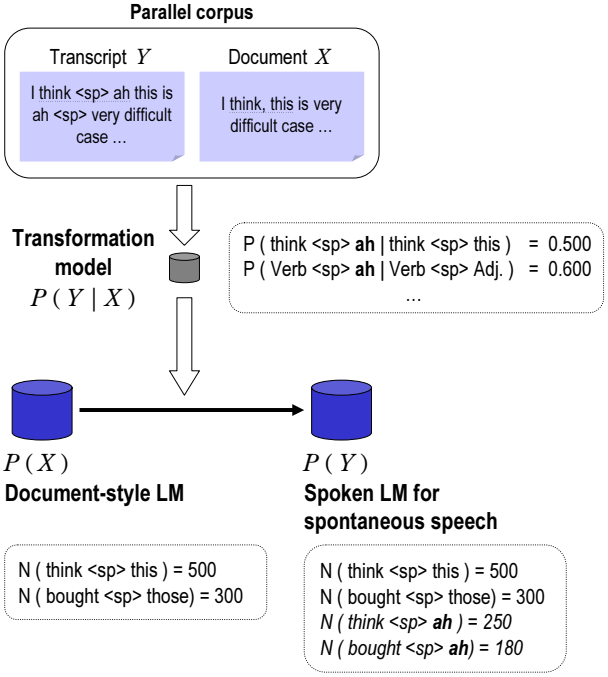

Fig. 1. Conceptual image of SMT-based transformation

When we assume n-gram language model for both X and Y, estimation of $P(Y)$ and $P(X)$ is replaced by estimation of occurrence counts of n-word sequence $(x_1^n,\ y_1^n)$, as below:

$$N(y_1^n) = N(x_1^n)\frac{P(Y|X)}{P(X|Y)} \tag{3}$$

where N denotes occurrence count of N-gram entries. If we further limit contextual information for $P(Y|X)$ and $P(X|Y)$ to n-words, Equation (3) is revised as follows:

$$N(y_1^n) = N(x_1^n)\frac{P(y_1^n|x_1^n)}{P(x_1^n|y_1^n)} \tag{4}$$

Equation (3) and (4) indicates that N-gram count $N(y_1^n)$ of spoken language model can be estimated using that of document-style model $N(x_1^n)$. Figure 1 shows conceptual image of this transformation. The conditional probabilities $P(Y|X)$ and $P(X|Y)$, namely, transformation model can be estimated using a parallel corpus of faithful transcripts and their document-style texts.

3. ESTIMATION OF LANGUAGE MODEL STATISTICS OF SPONTANEOUS SPEECH

3.1. Model of spontaneous speech

In this paper, language model of spontaneous Japanese is addressed within the above framework. Differences between Japanese spontaneous and document-style expressions are classified into three categories shown in Table 1. For these cases, we examine conditional probabilities $P(Y|X)$ and $P(X|Y)$ that correspond to probabilities of spontaneous and document-style transformation, respectively.

One of major characteristics is insertion of fillers. Fillers are often observed at the beginning or the end of utterances or pauses. However, they do not necessarily appear at these points, and frequency depends on actual filler words and contexts. Hence, insertion probability $P(Y|X)$ must be estimated. On the contrary, deletion probability $P(X|Y)$ is always 1, since fillers must be removed from transcripts for documentation.

The second characteristic is deletion of postpositional particles that indicate relationship between words or phrases, for example, linguistic case structure. Note that not all postpositional particles are deleted, for example, those indicating the nominative case are often omitted while those indicating the possessive case are rarely dropped. Also, postpositional particles cannot be inserted at all possible points when transforming a transcript into document-style. Thus, both $P(Y|X)$ and $P(X|Y)$ must be estimated.

The third one is substitution to colloquial words and phrases for polite expressions. Such substitution is often used for smooth utterance. Similar to the first case, not all possible words are actually substituted, however, appeared colloquial expressions must be always corrected in document-style text. So $P(Y|X)$ should be estimated, while $P(X|Y)$ is set to 1.

3.2. Estimation of transformation model probabilities

The key issue of the proposed method is estimation of conditional probabilities $P(Y|X)$. Training text data for transformation model is faithful transcripts and minutes of meetings

that were made by editing the former one by focusing on the phenomena described in the previous sub-section. These corrected expressions are also annotated for making alignments.

First, statistics of word-based patterns $N(y_1^n)$ and $N(x_1^n)$ are counted for different portions. Since we assume that the proposed method is applied to trigram language model, length of pattern (i.e., number of words in pattern) is three. Then, using these statistics, word-based transformation probabilities $P_{\mathrm{word}}(y_1^n|x_1^n)$ are estimated as following:

$$P_{\mathrm{word}}(y_1^n|x_1^n) = \frac{N(y_1^n)}{N(x_1^n)} \tag{5}$$

We also introduce a model based on part-of-speech (POS) since the word-based counts encounters data sparseness problem with small amount of the parallel corpus. POS tag is given to every word using a morphological analyzer. Aggregating statistics of same POS patterns, wider coverage and more reliable estimation is expected. Transformation probabilities $P_{\mathrm{POS}}(x_1^n|y_1^n)$ for POS-based patterns are estimated in the same way as Equation (5). Conditional probabilities $P_{\mathrm{word}}(x_1^n|y_1^n)$ and $P_{\mathrm{word}}(x_1^n|y_1^n)$ are also estimated accordingly.

3.3. Application of transformation model

Using Equation (4), the transformation model is applied to N-gram entries of a document-style language model that was trained with a large corpus. At first, each word-based pattern is applied to N-gram entries in turn, and transformation is performed with $P_{\mathrm{word}}(y_1^n|x_1^n)$ when the pattern is matched. If not matched, then POS-based pattern is compared, and applied with $P_{\mathrm{POS}}(y_1^n|x_1^n)$ if matched. In this way, a back-off scheme is realized. As a result, N-gram entries for spontaneous speech are generated with their estimated occurrence counts that are calculated by multiplying the transformation probabilities of the applied pattern to the count of the original model, as shown in Equation (4).

4. EXPERIMENTAL EVALUATION

4.1. Task description

We have evaluated the proposed method on the meeting transcription task. We collected four-year archive of the minutes of the National Congress of Japan for training of the baseline language model. The text size is 71M words. They were transcribed and edited by professional stenographers in the Congress. They are not faithful transcripts since typical spontaneous expressions such as fillers and end-of-sentence expressions are deleted or modified for documentation. Therefore, we made accurate transcripts for another set of meetings to train the transformation model. Total amount of training transcripts is 666K, which is too small to directly train a language model.

Table 2. Specifications of language models

Model	Minutes ("Baseline")	CSJ	Mixture ("+CSJ")	Transformed ("Proposed")
Training corpus	Minutes of the National Congress	Corpus of Spontaneous Japanese	Minutes + CSJ	Minutes + parallel
Style	Document	Spontaneous	—	Spontaneous
#Words	71M	2.9M	71M+2.9M	71M+0.7M
Vocabulary size	30,386	10,110	31,019	30,431
#Trigrams	3.63M	0.25M	3.77M	4.14M

For the test-set, we prepared transcript of yet another meeting. It contains variety of topics such as economy, foreign affairs and national security. The total number of distinct speakers is 23, so various spontaneous expressions are also observed. The size of the test-set is 63K words.

For comparison, we prepared another spoken-style language model using the Corpus of Spontaneous Japanese[7] which consists of a large number of extemporaneous public speeches, thus not matched to the task domain. The text size is 2.9M words. Interjections, repairs and colloquial expressions are faithfully transcribed, while they are rarely contained in the minutes of the Congress. Since our previous works[8] used a mixture model of the two models, we set it up for comparison. Mixture weights of the Minutes model and the CSJ model was optimally determined using the test-set.

Specifications of these language models are shown in Table 2.

4.2. Experimental results

The transformation model was constructed from the parallel corpus of 666K words. To suppress overgeneration of sparse patterns, those appeared more than twice in the training text were used. The total number of patterns is 5,719, which includes 2,187 (38%) POS-based patterns. Examples of obtained patterns are shown in Table 3.

Then, the transformation model was applied to the baseline model, and a transformed language model was obtained. The vocabulary size of the new model was 30,431 and the number of trigram entries was 4.14 million. The vocabulary size is almost same with that of the baseline model, and smaller than that of the mixture model, because only words characteristic to spontaneous speech were added by the proposed method while irrelevant topic words were also included in the mixture model. Moreover, the number of trigram entries in the new model is larger than the mixture model. The proposed method successfully predicted more trigrams of spontaneously spoken expressions than the synthesis-based approach.

Figure 2 shows perplexity and out-of-vocabulary (OOV) rates by the three models. These two measures were calculated using the test-set transcript. OOV words were included in computation of perplexity. Perplexity of the base-

Table 3. Examples of transformation patterns

Pattern					Prob.
word-based:					
<pause> *sore wa*			→	<pause> ***ah*** *sore wa*	0.011
<pause> *sore wa*			→	<pause> ***ano*** *sore wa*	0.007
(that is)				(inserted fillers)	
POS-based:					
PN N PP			→	w_1 w_2 ***ah***	0.095
PN N PP			→	w_2 ***ah*** w_3	0.048
(ex.)					
nippon	*keizai*	*wa*	→	*nihon keizai* ***ah***	0.095
nippon	*keizai*	*wa*	→	*keizai* ***ah*** *wa*	0.048
(Japan	economy	is)		(inserted fillers)	

PN : proper noun, N : noun, PP : postpositional particle
w_i : i-th word in the POS-tag sequence.

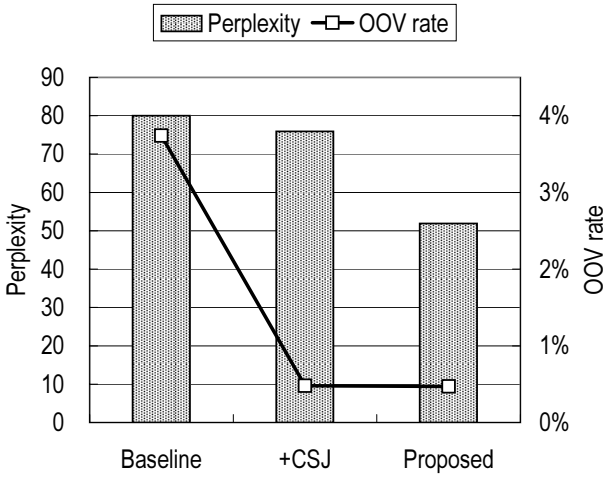

Fig. 2. Reduction of perplexity and OOV rate

line model, the mixture model ("+CSJ") and the transformed model ("Proposed") is 80.0, 75.9 and 51.9, respectively. OOV rates are 3.74%, 0.47% and 0.47%, respectively. Compared with the baseline model, the mixture model did not reduce perplexity while OOV rate was drastically reduced. Increase of trigram entries did not necessarily lead to improvement of prediction performance in the mixture model. Meanwhile, the proposed method reduced both perplexity and OOV rate, because it generated effective trigram entries with appropriate probabilities.

For reference, we also constructed a language model using the parallel corpus only. Perplexity and OOV rate by the model are 65.5 and 1.43%, respectively. The model contains too few trigram entries to realize reliable prediction performance. We also trained a mixture model of the Minutes and the parallel corpora. Perplexity by the resulting model is 56.2 and larger than that by the proposed method. We further tested a mixture model of the proposed model and the parallel-corpus-based model, by which perplexity of 43.4 was obtained. Note that the vocabulary of the latter two models is same as "Proposed" model. These results demonstrate that the proposed method could generate more various and effective trigram entries than a simple interpolation method.

5. CONCLUSIONS

We have presented a transformation approach to estimate a language model of spontaneous speech. The transformation model contains context-dependent probabilistic patterns of transformation from document-style to spontaneous speech. These patterns and their probabilities are determined based on occurrence statistics in a parallel corpus of faithful transcripts and their document-style texts. Since this training data is small, contexts are backed-off to POS-based ones, which provide more robust prediction. The transformation model is applied to N-gram counts of document-style language model, and N-gram entries of spontaneous speech are generated with estimated occurrence counts. In experimental evaluation, the proposed method efficiently and effectively generated spoken language model and reduced both perplexity and OOV rate. We are currently evaluating speech recognition performance with this model and hopefully report the results in near future.

6. REFERENCES

[1] A. Park, T. Hazen, and J. Glass, "Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling," in *Proc. ICASSP*, 2005, vol. 1, pp. 497–500.

[2] L. Lamel, G. Adda, E. Bilinski, and J.-L. Gauvain, "Transcribing Lectures and Seminars," in *Proc. Eurospeech*, 2005, pp. 1657–1660.

[3] F. Metze, C. Fügen, Y. Pan, and A. Waibel, "Automatically Transcribing Meetings using Distant Microphones," in *Proc. ICASSP*, 2005, vol. 1, pp. 989–992.

[4] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals, "Transcription of Conference Room Meetings: An Investigation," in *Proc. Eurospeech*, 2005, pp. 1661–1664.

[5] H. Schramm, X.L. Aubert, C. Meyer, and J. Peters, "Filled-Pause Modeling for Medical Transcriptions," in *Proc. SSPR*, 2003, pp. 143–146.

[6] P. Brown, S. Pietra, V. Pietra, and R. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[7] 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, vol. 3, pp. 518–521.

[8] Y. Akita and T. Kawahara, "Generalized Statistical Modeling of Pronunciation Variations using Variable-length Phone Context," in *Proc. ICASSP*, 2005, vol. 1, pp. 689–692.