TOPIC-INDEPENDENT SPEAKING-STYLE TRANSFORMATION OF LANGUAGE MODEL FOR SPONTANEOUS SPEECH RECOGNITION

Yuya Akita *Tatsuya Kawahara*

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

ABSTRACT

For language modeling of spontaneous speech, we propose a novel approach, based on the statistical machine translation framework, which transforms a document-style model to the spoken style. For better coverage and more reliable estimation, incorporation of POS (part-of-speech) information is explored in addition to lexical information. In this paper, we investigate several methods that combine POS-based model or integrate POS information in the ME (maximum entropy) scheme. They achieve significant reduction in perplexity and WER in a meeting transcription task. Moreover, the model is applied to different domains or committee meetings of different topics. As a result, even larger perplexity reduction is achieved compared with the case tested in the same domain. The result demonstrates the generality and portability of the model.

Index Terms— language model, statistical transformation, spontaneous speech, automatic speech recognition

1. INTRODUCTION

Language modeling is one of key issues in spontaneous speech recognition. Spoken-style expressions appeared in spontaneous speech should be appropriately modeled along with domain-relevant topics. Although a large amount of well-matched data is needed, the amount of available spoken-style text, especially faithful transcript, is usually limited because of transcription costs. Therefore, the conventional approach is combination of topic-relevant document databases with some spontaneous speech corpus. This synthesis-based approach is simple and effective. However, the resulting model also contains irrelevant linguistic expressions which may degrade ASR performance.

For better language modeling of spontaneous speech, approaches transforming a corpus or a model to the spoken-style have been investigated. For example, Schramm et al.[1] proposed generation of simulated spoken-style text by randomly inserting fillers into written-style text. Hori et al.[2] proposed transformation of language model using a weighted finite-state transducer (WFST) framework. We have also proposed language model transformation[3] based on the statistical machine translation (SMT)[4] framework. Transformation patterns and their probabilities are estimated using a parallel aligned corpus of the faithful transcripts and their document-style texts. This method directly estimates N-gram statistics without generating "virtual" transcripts, thus generates effective entries and reliable probabilities via statistical framework. In [3], we reported a preliminary result in a meeting transcription task.

The transformation or translation model is usually trained with a small corpus, therefore, efficient modeling is necessary. To realize better coverage and more reliable estimation, a model based on part-of-speech (POS) tags is introduced, and integrated as a back-off scheme. In this paper, we investigate the interpolation scheme of word-based and POS-based models as well as the maximum entropy (ME) scheme which integrates all available information.

The transformation model is expected to be domain-independent and applicable to other domains than that of the training data, although in the previous report[3] we only tested on the same domain. In this paper, we also investigate the generality and portability of the model by applying it to different domains or meetings of various topics.

2. STATISTICAL TRANSFORMATION OF LANGUAGE MODEL

Figure 1 shows a concept of the proposed statistical language model transformation. It is based on the framework of statistical machine translation, where sentence Y of the target language is generated from sentence X of the source language, which maximizes posterior probability $P(Y|X)$ based on Bayes' rule.

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

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

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

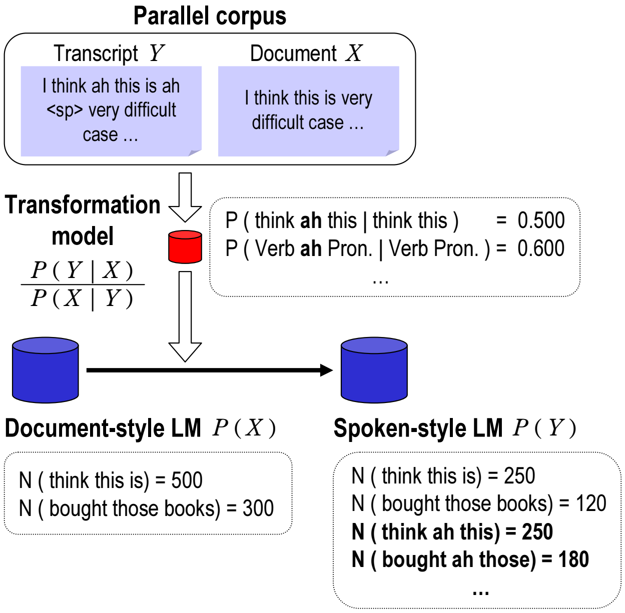
Parallel corpus
Transcript Y
Document X
I think ah this is ah <sp> very difficult case ...
I think this is very difficult case ...
Transformation model
$\frac{P(Y|X)}{P(X|Y)}$
P (think **ah** this | think this) = 0.500
P (Verb **ah** Pron. | Verb Pron.) = 0.600
...
Document-style LM $P(X)$
Spoken-style LM $P(Y)$
N (think this is) = 500
N (bought those books) = 300
N (think this is) = 250
N (bought those books) = 120
N (think ah this) = 250
N (bought ah those) = 180
...

Fig. 1. Conceptual image of language model transformation

The conditional probabilities $P(Y|X)$ and $P(X|Y)$, that is, transformation model, can be estimated using a parallel corpus of faithful transcripts and their corresponding document-style texts. For N-gram language model, transformation is actually performed on N-gram occurrence counts (N_{LM}).

$$N_{LM}(y_1^n) = N_{LM}(x_1^n)\frac{P(y|x)}{P(x|y)} \tag{3}$$

Here, x and y are individual patterns that are transformed and $N_{LM}(x_1^n)$ and $N_{LM}(y_1^n)$ are N-gram entries including them.

In our previous report[3], we estimated $P(y|x)$ and $P(x|y)$ using simple statistics of transformation patterns observed in a parallel corpus. Here, word contexts (neighboring words) are included in x and y to generate transformation patterns and estimate probabilities. For example, to model insertion of filler "ah," we include neighboring words to generate a pattern: x = "*think this*" and y = "*think ah this*." This context-dependent model will improve the precision of the model, but encounter the data sparseness problem, as the size of the parallel corpus is usually small. To solve the problem, in this paper, we introduce more elaborate models, based on linear interpolation and maximum entropy (ME) schemes, which also consider POS (part-of-speech) information.

2.1. Training of transformation model

The basic word-based transformation model is trained directly using occurrence counts of word sequences. First, statistics of document-style word sequence $N(x)$ and corresponding spoken-style word sequence $N(y \leftarrow x)$ are calculated using an aligned corpus. Then, word-based transformation probabilities $P_{\text{word}}(y|x)$ are estimated as following:

$$P_{\text{word}}(y|x) = \frac{N(y \leftarrow x)}{N(x)} \tag{4}$$

The word-based statistics will easily encounter the data sparseness problem. For better coverage and more reliable estimation, we also introduce a model based on POS tags. A POS tag is given to every contextual word using a morphological analyzer. Transformation probabilities $P_{\text{POS}}(y|x)$ for POS-based patterns are estimated in the same way as Equation (4). Conditional probabilities $P_{\text{word}}(x|y)$ and $P_{\text{POS}}(x|y)$ are also estimated accordingly.

2.2. Back-off scheme

The word-based and POS-based transformation models can be applied to N-gram entries of a document-style language model using Equation (3). As the simplest method to combine word-based and POS-based models, a back-off scheme is used:

$$P(y|x) = \begin{cases} P_{word}(y|x) & \text{if } x \rightarrow y \text{ exists} \\ P_{POS}(y|x) & \text{else if } x_{POS} \rightarrow y \text{ exists} \end{cases} \tag{5}$$

In this method, each word-based pattern is first applied to N-gram entries in turn, and transformation is performed with $P_{\text{word}}(y|x)$ when the pattern is matched. If not matched, then POS-based pattern is applied.

2.3. Linear interpolation scheme

As an alternative to the back-off method, we introduce the linear interpolation scheme of the two models. In this method, weighted sum of the word-based and POS-based probabilities are used as a transformation probability, and transformation is performed if either word-based or POS-based pattern is matched with the original N-gram entries.

$$P(y|x) = \lambda P_{word}(y|x) + (1-\lambda)P_{POS}(y|x) \tag{6}$$

2.4. Maximum entropy scheme

In the above methods, the word-based and POS-based models are estimated separately and combined afterwards. To better count lexical and POS information in an integrated manner, we introduce the ME scheme. In this model, conditional probability $P(y|x)$ is determined by

$$P(y|x) = \frac{1}{Z}\exp\left[\sum_i \lambda_i f_i(x,y)\right] \tag{7}$$

where f_i is a feature function, λ_i is a weight of a feature and Z is a normalization factor. As features $\{f_i\}$ for ME, we use preceding and following words and their POS tags. The ME model is applied to every N-gram entry of the document-style model, and spoken-style N-gram is generated if the transformation probability is larger than a threshold.

Table 1. Perplexity and WER on Budget committee test-set

Model		PP	WER
Baseline		80.0	—
Mixture	+CSJ	75.9	19.84%
	+Transcript	56.6	19.44%
Transformed (Proposed)	Back-off	52.7	19.43%
	Linear	53.0	19.47%
	ME	55.4	19.73%

Table 2. Specifications of language models

Model	Minutes ("Baseline")	Mixture ("Mixture")	Transformed ("Proposed")
Training corpus	Minutes of the National Congress	Minutes + Transcript	Minutes + Parallel
Style	Document	Spontaneous	Spontaneous
#Words	86.7M	86.7M+0.66M	86.7M+0.66M
Vocabulary size	51,671	52,093	52,093
#Trigrams	4.19M	4.32M	4.64M

3. COMPARISON OF TRANSFORMATION MODELS

We have evaluated the back-off, linear interpolation and ME-based transformation models on a meeting transcription task. We collected four-year (1999-2002) archives of the minutes of the National Congress of Japan for training of the baseline trigram 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 (mostly from the Budget committee in year 2003) to make a parallel corpus for training of the transformation model. Total amount of the training transcripts is 666K, which is too small to directly train a language model.

For the test-set, we prepared transcripts of yet another meeting from the Budget committee (in year 2003), which is more spontaneous than plenary sessions in a parliament[5, 6]. The total number of distinct speakers is 23, so various spontaneous expressions are observed. The size of the test-set is 63K words.

For comparison, we prepared a language model ("+CSJ") by combining the baseline with the Corpus of Spontaneous Japanese[7]. The CSJ consists of a large number of extemporaneous public speeches (2.9M words), which are not matched to the task domain. The mixture model was used in our previous works[8]. We also prepared another mixture model ("+Transcript") using the baseline model and a model trained with transcripts in the parallel corpus.

Table 1 shows perplexity by three models. All transformed models significantly reduced perplexity. The perplexity reduction over the baseline and our previous ("+CSJ") models are 35.0% and 30.6%, respectively, by the back-off method. The reduction over the "+Transcript" model is 6.9%. Performance of back-off and linear interpolation model is almost same, while ME model is slightly worse. The number of trigram entries in the ME method is fewer than those of other two transformed models, because transformation probabilities generally get smaller by the ME estimation.

Then, we evaluated the transformed language model by ASR experiments. In this experiment, the transformation model based on the back-off method was used. Acoustic model was triphone HMM trained using academic presentations in the CSJ. The amount of training data is about 257 hours. Speaker adaptive training (SAT) and unsupervised speaker adaptation by MLLR were performed. As a decoder, our Julius rev-3.5.2 was used.

Word error rates (WER) are shown in Table 1. Compared to our previous mixture model ("+CSJ"), the transformed model improved WER by relatively 2.0%. The improvement is statistically significant ($p<0.05$).

In these experiments, the "+Transcript" mixture model is almost comparable to the proposed transformation model. We presume this is because the parallel corpus and the test-set are chosen from the meeting of the same (Budget) committee and the same time period.

4. EVALUATION ON DIFFERENT TOPIC DOMAINS

In order to evaluate the generality of the proposed transformation model, we applied it to different topic domains from the training data. We prepared various congressional committee meetings held in year 2005, excluding the Budget committee, as new test-sets. The total number of meetings is 25 and the size of the test-set transcripts is 1.1M words. The average text size per meeting is 45K words.

In this experiment, the baseline language model was enhanced by adding the training texts of year 2003. Its specification is shown in Table 2. The mixture model was made by combining with the transcript of the parallel corpus. The transformed model used was based on the back-off method.

Figure 2 shows perplexity by the baseline, mixture and transformed models. In these cases, perplexity was significantly reduced in all meetings, even compared with the "+Transcript" mixture model. Average reduction by the transformed model is 20.5% against the baseline model, and 11.8% against the mixture model. Particularly, perplexity reduction against the mixture model is larger than that for the Budget committee (6.9%), i.e., topically same meeting shown in the previous section. The result shows that the proposed approach is general and even more effective for different topic domains.

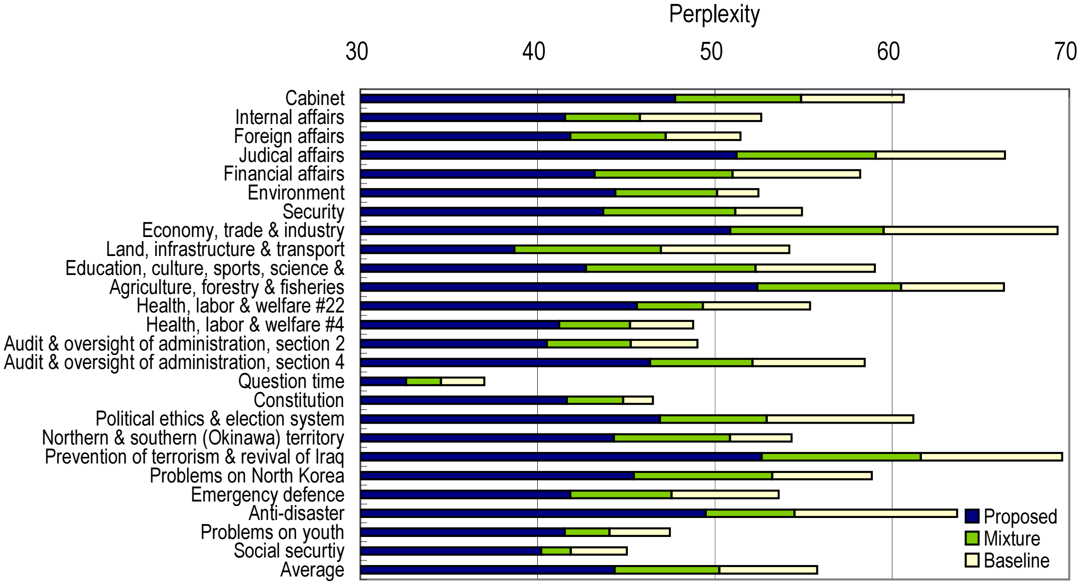

Fig. 2. Perplexity reduction for respective committee meetings

At this moment, we have made evaluation only with the perplexity reduction. However, we expect that ASR accuracy will also be improved, even compared with the result in the previous section, since the perplexity reduction is much larger.

5. CONCLUSIONS

We have proposed a novel statistical transformation approach which efficiently generates a language model for spontaneous speech recognition. For better coverage and more reliable estimation of the transformation model, incorporation of POS information is explored. Specifically, we investigated three methods: back-off method, linear interpolation, and ME method. In the experimental evaluation, the simple back-off and linear interpolation methods were more effective than the ME method both in perplexity reduction and WER improvement.

Furthermore, we have investigated the generality and portability of the approach by applying the model to different domains, or committee meetings of different topics. As a result, even larger perplexity reduction is achieved compared with the case tested in the same domain. The result is encouraging for further application and evaluation of the approach.

6. REFERENCES

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

[2] T. Hori, D. Willett, and Y. Minami, "Language Model Adaptation using WFST-based Speaking-style Translation," in *Proc. ICASSP*, 2003, vol. 1, pp. 228–231.

[3] Y. Akita and T. Kawahara, "Efficient Estimation of Language Model Statistics of Spontaneous Speech via Statistical Transformation Model," in *Proc. ICASSP*, 2006, vol. 1, pp. 1049–1052.

[4] 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.

[5] J. Lööf, M. Bisani, Ch. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schlüter, and H. Ney, "The 2006 RWTH Parliamentary Speeches Transcription System," in *Proc. ICSLP*, 2006, pp. 105–108.

[6] B. Ramabhadran, O. Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro, "The IBM 2006 Speech Transcription System for European Parliamentary Speeches," in *Proc. ICSLP*, 2006, pp. 1225–1228.

[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.