Automatic Transcription of Lecture Speech using Language Model Based on Speaking-Style Transformation of Proceeding Texts

Yuya Akita *Makoto Watanabe* *Tatsuya Kawahara*

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

Abstract

For language modeling of spontaneous speech recognition, we propose a style transformation approach, which transforms written texts to a spoken-style language model. Since these two styles are largely different and thus direct transformation is difficult, we cascade two transformation methods; rule-based transformation to rewrite written-style texts to intermediate "verbatim" texts, and statistical transformation of language model from the verbatim style to the spoken style which is suitable for ASR. In an experimental evaluation on real lecture speech, the proposed transformation approach achieved higher performance than the conventional linear interpolation method.

Index Terms: automatic speech recognition, lecture speech, language model, style transformation

1. Introduction

For automatic speech recognition (ASR) of spontaneous speech such as academic lectures and classroom lectures, spoken-style expressions should be appropriately modeled along with domain-relevant topics. Although a large amount of well-matched data is needed to construct language model, the amount of available spoken-style text, especially faithful transcript, is usually limited because of transcription costs. In contrast, a large amount of written texts are available, for example, proceedings of academic conferences and textbooks of classroom lectures. However, these texts hardly contain spoken-style expressions such as filler words.

Therefore, the conventional approach to language modeling is to combine topic-relevant document texts with some spontaneous speech corpus. For example, news articles and web texts were combined with transcripts of conversational telephone speech (CTS) for recognition of meetings and speeches [1, 2]. Also, materials of lectures such as slides were used with the CTS transcripts for ASR of classroom lectures [3]. This synthesis-based approach is simple and effective, while the resulting model contains irrelevant linguistic expressions and inconsistency in N-gram entries, which may degrade ASR performance.

For better language modeling of spontaneous speech, we have been proposing a language model transformation framework [4]. In this framework, transformation patterns and their probabilities are estimated as a "transformation model" using a small amount of parallel aligned corpus of faithful transcripts and document-style texts. This model is then applied to a large amount of document-style input texts, transforming them to spoken-style N-gram entries and statistics. We have successfully demonstrated the effectiveness of this framework in a parliamentary meeting transcription task [5], where the transformation model was applied to a large amount of verbatim transcripts which were officially made in Parliament.

When applying it to a lecture transcription task, we need to train a different transformation model, because speaking styles are much different between meetings and monologue speeches. In contrast to parliamentary meetings, verbatim records of lectures are not made in a large scale. We can often access to proceeding papers, but they are not in spoken style. Furthermore, direct transformation from proceeding texts to spoken style is not straightforward, because word-by-word alignment between written documents and faithful transcripts is hardly obtained.

Therefore, in this paper, we introduce another transformation method to rewrite written-style texts to the verbatim style. Specifically, we enhance rule-based text rewriting [6] to generate verbatim texts. Verbatim texts are also made by editing faithful transcripts, thus a parallel aligned corpus can be obtained by the edit, so we can train a statistical transformation model to estimate spoken-style N-gram entries. By cascading these two transformation methods, spoken-style N-gram language model can be constructed from written-style texts.

2. Style transformation for spontaneous speech recognition

2.1. Texts and styles

Figure 1 illustrates the concept of the proposed framework for ASR of lectures. We classify texts for language model into three types: faithful transcripts, verbatim texts and proceeding texts. Faithful transcripts

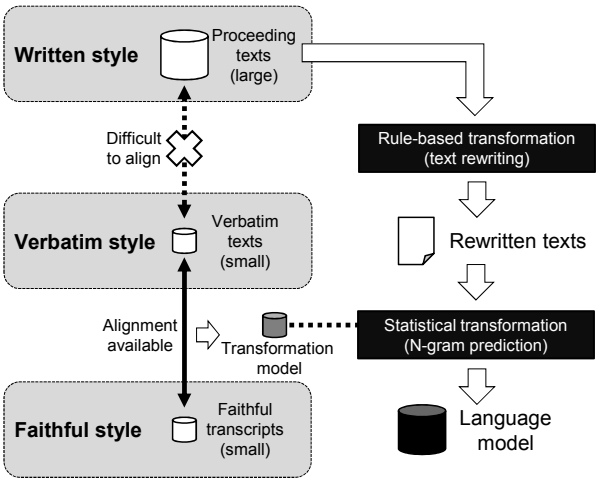

Figure 1: Conceptual image of the proposed framework

contain various spoken-style expressions including colloquial expressions and disfluencies such as fillers, and thus they are ideal for language model training. However, preparing a large amount of transcripts is virtually impossible because of transcription costs. Verbatim texts include lecture notes and closed captions. Even though these verbatim texts are often available, the size of texts is not sufficient for language model training. Moreover, their style does not completely match spoken utterances, since disfluency phenomena and colloquial expressions are often edited. In contrast to these types of texts, written documents such as lecture proceedings and books are available in a large scale, while their style is much more different from faithful transcripts. This is particular to Japanese language, for which we are conducting ASR. For example, plain-style end-of-sentence (EOS) words such as "*de-aru*" and "*da*" are used mainly in written sentences, while polite-style EOS words such as "*desu*" and "*masu*" are used in spoken utterances. Classic conjunctive words are also paraphrased, e.g., "*soreyue*" (therefore) is changed to "*sorede*" or simply "*de*."

2.2. Transformation to spoken style

In our previous work on parliamentary meeting transcription [5], we conducted speaking style transformation from verbatim transcripts to the faithful style. A huge amount of verbatim records were available, since the Japanese Diet (Parliament) creates an official verbatim record for every meeting. Thus, we could successfully apply a transformation model that was trained with a parallel corpus of verbatim records and corresponding faithful transcripts.

For lecture transcription, we can perform transformation of verbatim texts to the faithful style in the same manner, as a parallel aligned corpus of faithful transcripts and verbatim texts can be obtained by editing the former into the latter. Our transformation method is domain-independent, hence we can use a general lecture corpus for this purpose. However, we need to transform written documents, since the size of verbatim texts is limited as mentioned above. The word-based alignment between written documents and verbatim texts is difficult, because corresponding sentences are not always found in the written materials. This means that transformation from the written style is difficult with our statistical framework.

To fill the gap between the written style and the verbatim style, we introduce a rule-based text rewriting method [6]. Since people simply paraphrase typical written-style expressions when making utterances, transformation can be modeled with paraphrasing rules. On the contrary, disfluency phenomena such as fillers can be observed at any point in a sentence. This cannot be modeled by rules, and must be modeled in a statistical manner, thus we combine the rule-based transformation with the statistical transformation.

3. Language modeling based on cascaded style transformations

When we focus on ASR of lectures, we can exploit text resources of proceedings of academic meetings and conferences in the past. As a pre-processing, we first perform removal of foreign language sentences and unification of terms and expressions. Then, the rule-based text transformation is applied to generate verbatim-style rewritten texts. Finally we perform the statistical transformation to predict spoken-style N-gram entries and their occurrence statistics.

3.1. Rule-based text rewriting

To transform written-style text into the verbatim style, we adopt the rule-based transformation [6] which uses handcrafted rewriting rules for several types of expressions. Here, we incorporate rules rewriting to a polite form, and rules paraphrasing classic expressions to casual expressions. These rules are mainly applied to functional expressions. They are frequently observed, but their variety is limited. Therefore, transformation rules can be described by hand for respective written-style expressions. These rules use contextual information such as preceding and following words and morphological information such as part-of-speech (POS) tags, hence the rules cover a wide range of written expressions.

The original rewriting method [6] was developed to generate sentences for speech synthesis, where colloquial expressions are not preferred. Thus, we extended the rule set to cover colloquial expressions, by counting expressions in real lectures. Specifically, we counted trigram entries in proceeding texts and transcripts of speech in a collection of academic lectures, which will be described

in Section 4, then we extracted frequent N-gram entries found only in transcripts. Based on these N-gram entries, we made rewriting rules by hand.

For single input sentence, multiple rewriting rules may be applicable. In these cases we apply all applicable rules to the input sentence and generate multiple sentences.

3.2. Statistical transformation of language model

N-gram statistics are calculated for verbatim texts, which are generated by rule-based transformation. Then, statistical transformation is performed to predict occurrence counts in faithful transcripts. The statistical transformation method [4] is based on the framework of statistical machine translation [7], 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 the verbatim style and the faithful style as different ones, denoted by X and Y, respectively, and estimate faithful-style language model $P(Y)$, which is formulated as Equation (2) by rewriting Equation (1).

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

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

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

Here, x and y are individual patterns that are transformed, and $N_{LM}(x)$ and $N_{LM}(y)$ are N-gram entries including them. Transformation patterns x and y contain preceding and following words as contexts. To alleviate the data sparseness problem, part-of-speech (POS) contexts are also introduced. Using the estimated N-gram entries and occurrence counts, the spoken-style language model is trained in a standard manner.

In this work, we use the Corpus of Spontaneous Japanese (CSJ) [8], which is a collection of academic lectures and public speeches, to train the transformation model. Since the CSJ does not have edited documents for transcripts of speech, we prepared the documents by editing transcripts. The edit includes removal of fillers and paraphrasing of colloquial expressions. We prepared this kind of documents for 177 lectures, and aligned them with transcripts to form a parallel corpus.

Table 1: Specifications and performance of language models

LM	Vocab. size	Perplexity	%OOV	WER
NLP	9.94K	245	2.55%	26.1%
NLP-rule	11.0K	287	2.13%	25.3%
NLP-stat	11.0K	105	1.73%	16.8%
CSJ_E	19.9K	210	4.48%	26.5%
CSJ_E+NLP	24.0K	109	1.08%	16.1%
CSJ_E+NLP-rule	24.5K	128	0.97%	16.0%
CSJ_E+NLP-stat	24.5K	100	0.97%	15.6%

4. Experimental evaluations

We evaluated the proposed approach in ASR of real lectures. For the test set, we chose 10 Japanese lectures in workshops on spoken document processing which were held in years 2007, 2008 and 2009. The topic of the lectures is natural language processing such as language modeling, information retrieval and machine translation. The average duration of lectures is 20 minutes, and the average number of words is 4.4K. For the analysis to create transformation rules, which was described in Section 3.1, we used speech data of other lectures in the workshops.

4.1. Language models and prediction performance

Table 1 lists the language models tested in this experiment. As written-style input texts, we used proceedings of annual meetings of the Association for Natural Language Processing in years 2004 to 2009. Using these texts, we trained three language models; a written-style "NLP" model trained directly with the proceeding texts, partly transformed model "NLP-rule" trained from the texts transformed only by the rule-based method, and fully transformed model "NLP-stat" with the rule-based and statistical transformation methods. For comparison, we also conducted a conventional approach, i.e., linear interpolation of domain-relevant and spoken-style models. Here, NLP model was used as the former. For the latter, a general spoken-style language model ("CSJ_E") was trained from extemporaneous public speeches in the CSJ. We preliminarily examined the best interpolation weights and determined it as 0.5:0.5. The total numbers of words in training data of NLP and CSJ_E models were 2.7M and 4.1M, respectively.

Perplexity and out-of-vocabulary (OOV) rate on the test set by these models are also shown in Table 1. Compared with the written-style NLP model, the NLP-rule model had a larger vocabulary, which reduced the OOV rate while perplexity was not improved by this model. The NLP-stat model had almost the same vocabulary as the NLP-rule model because only a small number of filler

words were added, nevertheless both perplexity and the OOV rate were largely improved. The statistical transformation could successfully cover fillers and spoken-style N-gram entries. This NLP-stat model had comparable performance to the linearly interpolated CSJ_E+NLP model in terms of perplexity. The latter achieved lower OOV rate, since it covers much more variations in the spoken style.

Then, we investigated interpolations of the transformed model with the CSJ_E model, i.e., CSJ_E+NLP-rule and CSJ_E+NLP-stat models. The tendency of reduction on perplexity and OOV rates over CSJ_E+NLP model was almost same as the case with NLP, NLP-rule and NLP-stat models. After the spoken-style model was interpolated, the proposed transformation achieved further improvement on perplexity and OOV rate.

4.2. Evaluation on ASR

We tested these language models by ASR. As an acoustic model, we prepared triphone HMM trained with 257-hour lecture speech in the CSJ. Minimum phone error (MPE) training [9] was conducted for the model. The number of shared states was 3,000, and each state had 16 Gaussians. We used 38-dimensional acoustic features which consisted of MFCC, ΔMFCC and $\Delta\Delta$MFCC, together with ΔEnergy and $\Delta\Delta$Energy. We applied cepstrum mean and variance normalization (CMN/CVN) and vocal tract length normalization (VTLN) to the features. Then, we performed unsupervised MLLR-based speaker adaptation of the acoustic model to each speaker in the test set. The decoder was Julius rev.4.1.5.

The average word error rates (WER) on the test set are listed in Table 1. With the NLP model, the WER was high (26.1%) because the model hardly covered spoken-style expressions. WER was reduced to 25.3% by applying the rule-based transformation, and succeeding statistical transformation drastically improved it to 16.8%. As for interpolated models, WER of 16.1% was obtained by the CSJ_E+NLP model, and the CSJ_E+NLP-stat model further improved WER by 0.5% (15.6%). The effect of the proposed method was demonstrated in real ASR.

5. Conclusions

We have proposed an approach to build a language model for speech recognition of spontaneous speech such as lectures, without the use of rich amount of transcripts. The approach consists of two transformation methods. First, rule-based transformation is applied to written-style texts to generate verbatim texts, then statistical transformation is conducted on the texts to generate spoken-style N-gram entries and statistics. We evaluated the proposed approach in ASR of real lectures, and demonstrated improvement of WER over the conventional approach.

6. Acknowledgements

The authors are grateful to Prof. Sadao Kurohashi and Dr. Tomohide Shibata of Kyoto University for providing the rule-based text transformation tool. This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

7. References

[1] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An Audio Indexing System for Election Video Material," in *Proc. ICASSP*, 2009, pp. 4873–4876.

[2] S. Kombrink, T. Mikolov, M. Karafiat, and L. Burget, "Recurrent Neural Network based Language Modeling in Meeting Recognition," in *Proc. Interspeech*, 2011, pp. 2877–2880.

[3] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," in *Proc. Interspeech*, 2007, pp. 2553–2556.

[4] Y. Akita and T. Kawahara, "Statistical Transformation of Language and Pronunciation Models for Spontaneous Speech Recognition," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 18, no. 6, pp. 1539–1549, 2010.

[5] Y. Akita, M. Mimura, and T. Kawahara, "Automatic Transcription System for Meetings of the Japanese National Congress," in *Proc. Interspeech*, 2009, pp. 84–87.

[6] S. Kurohashi, D. Kawahara, N. Kaji, and T. Shibata, "Automatic Text Presentation for the Conversational Knowledge Process," in *Conversational Informatics: an Engineering Approach*, T. Nishida, Ed. John Wiley & Sons Ltd., 2007, pp. 201–216.

[7] 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.

[8] 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, pp. 518–521.

[9] D. Povey and P. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *Proc. ICASSP*, vol. 1, 2002, pp. 105–108.