Automatic Transformation of Lecture Transcription into Document Style using Statistical Framework

†*Kazuya Shitaoka,* ‡*Hiroaki Nanjo,* †*Tatsuya Kawahara*

†Guraduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{shitaoka,kawahara}@ar.kyoto-u.ac.jp
‡Faculty of Science and Technology, Ryukoku University
Ohtsu 520-2194, Japan
nanjo@ryukoku-u.jp

Abstract

This paper addresses automatic transformation from spoken style texts to written style texts. Exact transcriptions and speech recognition results of live lectures include many spoken language expressions, and thus, are not suitable for documents and need to be edited. In this paper, we present a method of applying of the statistical approach used in machine translation to this post-processing task. Specifically, we implement the correction of colloquial expressions, the deletion of fillers, the insertion of periods, and the insertion of particles in an integrated manner. A preliminary evaluation confirms that the statistical transformation framework works well and we achieved high recall and precision rate of period and particle insertion.

1. Introduction

Under the Science and Technology Agency Priority Program in Japan (1999-2004) [1], a large scale spontaneous speech corpus, which is called the "Corpus of Spontaneous Japanese (CSJ)", has been collected and studies of spontaneous speech recognition have been explored. Our main goal is to realize automatic transcriptions (automatic speech recognition) and post-processing for an archive of live lectures such as oral presentations in conferences.

Transcriptions of lecture speech include many colloquial expressions peculiar to spoken language. The Japanese spoken language in particular is quite different from the written language. Thus, Japanese spoken language is not suitable for documents in terms of readability, and it is necessary to transform transcriptions and recognition results into document style for practical archives. This process is also important as a pre-process of automatic summarization [2][3]. In this paper, we consider spoken and written Japanese language to be different languages and apply the translation methodology to the automatic transformation of the former to the latter.

There are a number of software programs which do this transformation task, but they only perform one-to-one transformation based on pre-defined rules and simple pattern matching, and do not consider the consistency and correctness of the output. Moreover, the simple method can not deal with cases in which one word can be mapped in multiple ways depending on the context.

In this paper, we approach the problem using a statistical framework that has become popular in machine translation [4][5][6]. With this framework, we perform correction of colloquial expressions, deletion of fillers, insertion of periods (end-of-sentence symbols) and insertion of particles in an integrated manner.

2. Framework of Statistical Machine Translation

The statistical machine translation framework is formulated in the same way as statistical speech recognition. It is formulated by finding the best output sequence Y for an input sequence X, such that the *a posteriori* probability $P(Y|X)$ is maximum. According to Bayes's rule, maximization of $P(Y|X)$ is equivalent to the maximization of the product (sum in log scale) of $P(Y)$ and $P(X|Y)$, where $P(Y)$ is the probability of the source language model and $P(X|Y)$ is the probability of the transformation model. The transformation model represents correspondence of input and output word sequences. In this paper, we perform left-to-right decoding, that is, we do not address the swapping of word positions even though this is usually taken into account in conventional machine translation studies.

In the task of style conversion, the input X is a word sequence of spoken language transcriptions that do not have periods (i.e., end-of-sentence symbols) but include pause duration. The output Y is a word sequence of the written language. For $P(Y)$ calculation, we use a word 3-gram model trained with a written language corpus. Since applying the conversion of one word affects

Table 1: Example of conversion pairs and their probabilities

Written language Y	Spoken language X	$P(X\|Y)$
donoyo: (how)	*donoyo:*	0.54
	do:yu:fu:	0.46
(-shi) teiru (doing something)	*(-shi) teru*	0.12
	(-shi) teiru	0.88

Table 2: Example of patterns and probabilities of particle deletion

Pattern Y	Deletion probability
Noun wa Noun	0.073
Noun o Noun	0.032
Noun wa Verb	0.056
Noun o Verb	0.040
Noun ga Verb	0.012
Noun wa Adjective	0.20
Noun ga Adjective	0.024
Noun wa Conjunction	0.16

"wa", *"o"* and *"ga"* are Japanese particles.

neighbor words in an N-gram model, the decoding is performed for a whole input word sequence with beam pruning.

3. Transformation Procedures

This section describes the transformation procedures and transformation model that gives $P(X|Y)$ in detail.

3.1. Deletion of Fillers

Japanese spoken language include many fillers and interjections such as *"ano:"* and *"e:to"*, which must be deleted in transcribing to written text. Since none of these are observed in written language corpus with which the source language model is trained, the equation $P(Y) = P(Y|X) = 0$ holds where Y includes such words. This suggests that all fillers and interjections in the input transcription X are automatically deleted by the source language model.

3.2. Correction of Colloquial Expressions

In transforming Japanese spoken language to written language, colloquial expressions peculiar to spoken language should be converted into formal expressions. $P(X|Y)$ represents the probability that colloquial expression X arises for written expression Y. We estimate $P(X|Y)$ from the parallel corpus of exact transcriptions of spoken language and texts after correction by a human editor.

We define 64 conversion pairs and estimate their probabilities with a parallel corpus of 18 lectures of CSJ. Examples of transformation pairs and their probabilities are listed in Table 1.

Table 3: Test-set specification

	Duration (min.)	#Words in transcription	
		exact	cleaned
A01M0035	28	5557	5378
A01M0007	30	3899	3802
A01M0074	12	2509	2451
A05M0031	27	5371	4854

3.3. Insertion of Particles

Since in spoken Japanese particles are often omitted, they must be complemented with alternatives. As the particle phenomena is dependent on adjacent words, we define the deletion probabilities of particles $P(X|Y)$ for the triplet of the preceding part of speech, the particle itself, and the following part of speech, such as *"Noun Particle Noun"*, *"Noun Particle Verb"* and *"Noun Particle Adjective"*. Examples are listed in Table 2.

3.4. Insertion of Periods

In recognizing read speech, periods are conventionally assigned to pauses at the end of utterances because an utterance is assumed to be a sentence. In spontaneous speech, however, pauses are put not only at the end of sentences but at arbitrary places. Thus, CSJ has pause marks with their duration instead of periods and speech recognizers using a language model trained with CSJ do not output periods. However, periods are needed in document-style text for better readability.

In this paper, we convert pauses into periods selectively, considering duration information and the adjacent parts of speech in the statistical framework. Specifically, the pause duration thresholds of X with which pauses are converted to periods are set up depending on the contextual words of Y.

4. Experimental Evaluation

4.1. Task and Test-set

The CSJ developed by the Science and Technology Agency Priority Program project consists of a variety of oral presentations at technical conferences and informal monologue talks on given topics. The test-set for evaluation consists of four lecture presentations specified in Table 3, which have been commonly used in speech recognition tasks [7][8]. They were given by experienced lecturers who did not prepare drafts.

In the statistical machine translation framework, the source language model score $P(Y)$ has significant influence on the candidate selection from possible hypotheses. In written Japanese, there are basically two styles regarding end-of-sentence expressions: the normal i.e., *"dearu"* style and the polite i.e., *"desu/masu"* style. For these, we use two language models: (1) one trained with lecture notes available on the World Wide Web [7] and

Table 4: Training data of source language models

	Lecture notes on Web	Newspaper corpus
#Morphemes	1.7M	76.5 M

Table 5: Result of period insertion with several pause duration thresholds

Pause duration threshold	Recall	Precision	F.
Zero	83.2%	75.4%	0.791
Average	64.4%	93.7%	0.763
Depending on expressions	76.3%	92.3%	0.835

source language model: lecture notes

(2) one trained with a Japanese newspaper corpus [9], respectively. Training data amounts for these models are shown in Table 4. Thus, the end-of-sentence expression styles are converted to be consistent according to the selected model.

4.2. Results

We have implemented the above-described procedures. Fillers are completely deleted automatically. Conversion from colloquial expressions to formal expressions is almost successful.

In this section, we evaluate the results of period and particle insertion in terms of recall rate, precision rate and F-measure. They are defined as follows.

$$\textit{recall rate} = \frac{\textit{number of correctly inserted parts}}{\textit{number of parts that should be inserted}}$$

$$\textit{precision rate} = \frac{\textit{number of correctly inserted parts}}{\textit{number of all inserted parts}}$$

$$\textit{F-measure} = \frac{\textit{2 * recall rate * precision rate}}{\textit{recall rate + precision rate}}$$

4.2.1. Decoding Parameters

In the statistical machine translation framework, the output sequence Y is found such that $P(Y) \cdot P(X|Y)$ is maximized. For practical use, we introduce parameters in the following expression for improving performance that are familiar in statistical speech recognition.

$$\max_{Y}\{\log P(X|Y) + \alpha \log P(Y) + \beta N_Y\}$$

where α is a language model weight used to adjust the dynamic ranges of $P(Y)$ and $P(X|Y)$, and β is a word insertion penalty used to normalize the number of words in Y (N_Y) since the language score $P(Y)$ becomes smaller according to the number of words.

Figure 1 plots the average of F-measures for several values of α and β. The optimal values of α and β were 5 and 8, respectively, and these values are used in the following experiments.

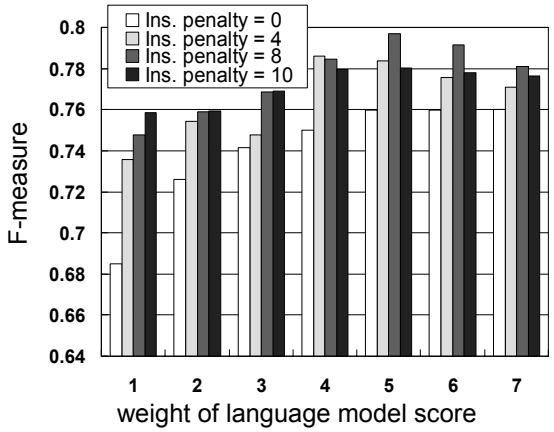

Figure 1: Average F-measure (period and particle insertion) for different values of weight and insertion parameters

Table 6: Particle insertion results

source LM	Recall	Precision
Lecture Note	89.4% (42/47)	65.9% (81/123)
Newspaper	87.2% (41/47)	63.3% (124/196)

4.2.2. Result of Period Insertion

We investigated several methods that convert pauses of spoken language into periods of written language. A pause whose duration is longer than a given threshold can be converted into a period. Specifically, three thresholds are compared: (1) zero, (2) average of pause duration in a lecture and (3) use of different thresholds depending on the context, the latter of which is proposed in this paper. Specifically, a pause following a typical Japanese end-of-sentence expression such as *"- desu (pause)"* and *"- masu (pause)"* can be converted to a period even if the duration is short. On the other hand, a pause at the beginning or end of sentence expressions peculiar to spoken Japanese such as *"- to (pause)"*, *"(pause) de -"* and *"- ta (pause)"* can be converted when the duration is longer than the average. The overall results are shown in Table 5.

When we set zero as the threshold, that is, we allow any pause to be converted to a period, erroneous insertion is caused and the precision rate is degraded. In contrast, setting the threshold to the average value degrades the recall rate. Using a context dependent threshold, we were able to achieve high recall and precision rates.

4.2.3. Result of Particle Insertion

Next, we evaluate the performance of inserting particles (See Table 6). In this experiment, we used two different source language models to compute $P(Y)$. Statistically, there is no significant difference between them. Since selected particles differ among human editors, we set multiple correct particles, the recall rate of which was 89.4%.

Table 7: Comparison of statistical and rule-based transformation models

method	Period insertion			Particle insertion		
	Recall	Precision	F.	Recall	Precision	F.
Statistical	76.3% (281/371)	92.3% (283/306)	0.835	89.4% (42/47)	65.9% (81/123)	0.759
Rule-based	76.3% (281/371)	92.3% (283/305)	0.835	89.4% (42/47)	58.3% (109/187)	0.706

source language model: lecture notes

On the other hand, we did not obtain a high precision rate. We investigated the false insertions in detail and found that errors were mainly caused by the insertion of the particle *"no"* (meaning "of") between compound nouns, such as *"Kyoto Daigaku"* (the proper noun "Kyoto University") v.s. *"Kyoto no Daigaku"* (a university located in Kyoto). This problem can be solved by adding entries of proper nouns to the lexicon. Excluding these error segments, the precision rate should be about 79.4%.

4.2.4. Unification of End-of-Sentence Style

The conversion process involves unification of end-of-sentence expression styles. The results we obtained with this process are shown in Table 8. With a lecture model, 86.5% of end-of-sentence expressions were unified to the polite *"desu/masu"* style. With the use of a newspaper language model, 65.2% of end-of-sentence expressions were unified to the normal *"dearu"* style. In this paper, we do not deal with transformations that require conjugations, such as *"shi masu"* (polite style of verb " do") $\rightarrow$ *"suru"* (normal style of verb "do") and *"ki masu"* (polite style of verb "come") $\rightarrow$ *"kuru"* (normal style of verb "come"). These must be also described in the transformation models.

4.2.5. Comparison of Transformation Models

To verify the effect of the transformation model, we set the transformation scores $P(X|Y)$ to 1 or 0, which are equivalent to those of rule-based transformation. A comparison is given in Table 7. There are no differences for period insertion. As for particle insertion, the statistical transformation model reduced 36 false alarms without increase of false rejections and thus improved the precision rate.

5. Conclusions

In this paper, we have presented a statistical method of transforming spoken language to written language for automatic archiving of lectures. This method enabled us to successfully achieve filler deletion, correction of colloquial expressions, and the high accuracy of both period insertion (F-measure: 0.835) and particle insertion (F-measure: 0.759). The results showed that the proposed approach is more effective than the conventional rule-based approach.

Table 8: Results of end-of-sentence expression style unification

Source LM	Lecture Note	Newspaper
Unification Ratio	86.5%	65.2%

Acknowledgment: The authors are grateful to Prof. S. Furui of the Science and Technology Agency Priority Program on "Spontaneous Speech: Corpus and Processing Technology". The authors are thankful to Prof. H. G. Okuno of Kyoto University for useful comments.

6. References

[1] S.Furui, K.Maekawa, and H.Isahara, "Toward the realization of spontaneous speech recognition – introducing of a japanese priority program and preliminary results –," in *Proc. ICSLP*, 2000, vol. 3.

[2] C.Hori and S.Furui, "Automatic speech summarization based on word significance and linguistic likelihood," in *Proc. IEEE-ICASSP*, 2000, vol. III, pp. 1579–1582.

[3] Y.Yamashita and A.Inoue, "Extraction of Important Sentences Using F0 Information for Speech Summarization," in *Proc. IEEE-ICASSP*, 2002, pp. 1181–1184.

[4] I.Garcia-Varea, F.Casacuberta, and H.Ney, "An Iterative, DP-Based Search Algorithm For Statistical Machine Translation," in *Proc. ICSLP*, 1998, vol. 4, pp. 1135–1138.

[5] Y.Wang and A.Waibel, "Fast Decoding For Statistical Machine Translation," in *Proc. ICSLP*, 1998, vol. 6, pp. 2775–2778.

[6] T.Watanabe and E.Sumita, "Bidirectional Decoding for Statistical Machine Translation," in *Proc. COLING*, 2002.

[7] H.Nanjo, K.Kato, and T.Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," in *Proc. EUROSPEECH*, 2001, pp. 2531–2534.

[8] T.Shinozaki and S.Furui, "Towards automatic transcription of spontaneous presentations," in *Proc. EUROSPEECH*, 2001, pp. 491–494.

[9] A.Lee, T.Kawahara, K.Takeda, M.Mimura, A.Yamada, A.Ito, K.Itou, and K.Shikano, "Continuous Speech Recognition Consortium — an open repository for CSR tools and models —," in *In Proc. IEEE Int'l Conf. on Language Resources and Evaluation*, 2002.