Automatic Detection of Edited Parts in Inexact Transcribed Corpora Based on Alignment between Edited Transcription and Corresponding Utterance

KENGO OHTA
Toyohashi University of Technology
Dept. of Computer Sciences and Engineering
Japan
kohta@slp.cs.tut.ac.jp

MASATOSHI TSUCHIYA
Toyohashi University of Technology
Information Media Center
Japan
tsuchiya@imc.tut.ac.jp

SEIICHI NAKAGAWA
Toyohashi University of Technology
Dept. of Computer Sciences and Engineering
Japan
nakagawa@slp.cs.tut.ac.jp

Abstract: The availability of a large-scale spontaneous speech corpora is crucially important for various domains of spoken language processing. However, the available corpora are usually limited because of its cost to prepare. On the other hand, inexact transcribed corpora have been widely produced in the form of shorthand notes, meeting records, or closed captions. Although these inexact transcribed corpora are more freely available than faithful/exact ones, these are not faithfully transcribed but contains edited transcriptions. Under this background, we are considering to build an efficient semi-automatic framework for converting inexact transcripts to faithful ones or exact transcriptions. This framework consists of two steps: the first step is to automatically detect positions of edited parts, and the second step is to manually transcribe the edited parts. This paper proposes an automatic detection method of edited parts in edited transcribed corpora for this framework. In our proposed method, an automatic alignment between edited transcription and its corresponding utterance is performed, and then a support vector machine based detector is applied to detect edited parts using some features obtained by the automatic alignment. As a result of evaluation on the Japanese National Diet Record, a reasonable result was obtained in speaker-closed condition.

Key–Words: edited part detection, edited corpora, speech alignment, spoken language processing

1 Introduction

A number of applications in the field of spoken language processing could benefit from the availability of training material based on a large-scale spontaneous speech corpora. For example, a speech recognition system for spontaneous speech needs a language model that covers spoken-style expressions as well as domain-relevant topics. The simplest approach to constructing such a model is to train it from a large-scale corpus consisting of many faithful transcripts of spontaneous speech in the relevant domain. However, the available corpora are usually limited because they are quite expensive to prepare.

On the other hand, inexact transcribed corpora are widely produced in the form of shorthand notes, meeting records, or closed captions. These inexact transcribed corpora are more freely available than faithful ones because their inexactness reduces transcription costs. For instance, the Japanese National Diet Library\(^1\) publishes the Japanese National Diet Record, which contains numerous transcriptions of debates in the Japanese National Diet for the past 60 years. However, these corpora are not faithfully transcribed but contains edited transcriptions. An example is shown in Figure 1. As shown in Figure 1, in the inexact transcription to help the readability, redundant expressions (ex."ですね/desune", "と/to") and disfluencies such as filled pause (ex."え/-eel", "い/-il") and hesitations (ex."ka/kel") are removed. Additionally, colloquial expressions (ex."てる/teru") are replaced by literary expressions (ex."ついている/teiru"), and lack of particles (ex."を/wol") are compensated. Some commas are added or removed according to the shorthand writers’ judgement, which ignores the speaker’s intention.

Under this background, we are considering to build an efficient semi-automatic framework for converting inexact transcripts to faithful ones or exact transcriptions. This framework consists of two steps: the first step is to automatically detect positions of edited parts, and the second step is to manually transcribe the edited parts. And then, we collect the parallel corpora of the spoken-style(exact) expressions and the written-style(edited) expressions.

\(^1\)http://kokkai.ndl.go.jp/
Figure 1: An Example of exact/inexact transcription (in Japanese)

(i) Exact Transcription

( tokoro ga desu ne, ee, kono shiryou, mi te mi masu to kanagawa ken no baai wa, ke, kekkoto shi te zaisei teki ni ii yutaka ni naq teru to. )

(ii) Inexact Transcription for readability

Figure 2: Bigram Constraint for Automatic Alignment

\[ P(w_{i+1} | w_i) = 0.5 \]

\[ P(sp_i | w_i) = 0.5 \]

Figure 2: Bigram Constraint for Automatic Alignment

As the first step of our proposed method, we perform an automatic alignment between exact transcription and its corresponding utterance. In this alignment, the words in the exact transcription and the corresponding utterances are aligned with allowing the insertion of short pause between words. Such alignment is implemented by an automatic speech recognition with the constraint based on a bigram language model as shown in Figure 2. Here, \( w_i \) represents the \( i \)-th word in the transcription and \( sp_i \) represents the short pause occurred immediately after the word \( w_i \).

In this paper, we propose a method to detect positions of edited parts from inexact transcribed corpora. In our proposed method, an automatic alignment between edited transcription and its corresponding utterance is performed, and then a support vector machine based detector is applied to detect edited parts using some features obtained by the automatic alignment.

There are several previous works utilizing an automatic alignment between text and speech for spoken language processing. For example, Lamel et al.[2] applied an automatic alignment for lightly supervised acoustic model training. In their work, an automatic alignment is used to filter out unreliable training data for acoustic model training. In other cases, Roy et al.[3] utilized an acoustic score obtained from automatic alignment to estimate accuracy and difficulty of transcription of speech recordings. Additionally, in [4], Maruyama et al. suggested to utilize an automatic alignment for timing detection of closed captioning in documentary program. On the other hand, the main idea of our proposed method is to detect edited parts based on degradation of acoustic score and variation of syllable durations in automatic alignment due to the mismatch between edited transcription and its corresponding utterance.

The remainder of this paper is organized as follows. In Section 2, we introduce an automatic alignment method between edited transcription and its corresponding utterance. Section 3 describes an edited part detector based on a support vector machine. The evaluation experiment on the Japanese National Diet Record is presented in Section 4. Finally, Section 5 concludes the paper.

2 Automatic Alignment between Edited Transcription and Its Corresponding Utterance

As the first step of our proposed method, we perform an automatic alignment between edited transcription and its corresponding utterance. In this alignment, the words in the edited transcription and the corresponding utterances are aligned with allowing the insertion of short pause between words. Such alignment is implemented by an automatic speech recognition with the constraint based on a bigram language model as shown in Figure 2. Here, \( w_i \) represents the \( i \)-th word in the transcription and \( sp_i \) represents the short pause occurred immediately after the word \( w_i \). As a result of this constraint, the output of the automatic speech recognition is restricted to the same word sequence in the transcription.
is distorted and the acoustic score in the alignment degrades because of the mismatch between the syllable and the aligned model. In the example of Figure 3, the duration of syllable /si/ is inappropriately extended and the model /te/ is forced to align with the frames of the syllable /ne/. Besides, the short pause model is also forced to align with the frames of the syllables /te de su/. Hence, if syllable durations are overly long or short compared with their inherent values, or if acoustic scores are worse compared with a standard value, it may suggest that there are mismatches between text and utterance due to edit it.

In the most of inexact transcribed corpora, the correspondence between an utterance in speech and a sentence in corpora is not available. Considering this, we estimated the number of words in each utterance based on the number of syllables in automatic continuous syllable recognition result. Our preliminary experiment showed that there are almost no differences in the accuracy of automatic alignment either when the number of words in each utterance was precisely available, or when it was estimated based on the continuous syllable recognition.

3 Automatic Detection of Edited Part

As the second step of our proposed method, we apply a support vector machine based detector to detect edited parts using some features obtained by the automatic alignment which is described in Section 2.

In this study, we formalize the detection of edited parts as a binary classification problem for each word in inexact transcribed corpora. Each word is classified into edited word or non-edited word based on the features obtained by the automatic alignment.

3.1 Implementation

We chose TinySVM (ver 0.09)[5] as an implementation of support vector machine with polynomial kernel.

3.2 Features

In our method, the seven categories of features are adopted. We used the features [6] as the reference for feature selection. Details of each feature are described in the following sections. All these features for the focused word, the preceding two words, and the succeeding two words are combined as a feature vector for each word.

3.2.1 Acoustic Score

The acoustic score obtained by automatic alignment is utilized for the first feature. This score is normalized by duration of word, and then subtracted by an acoustic score obtained from HMM based continuous syllable recognition. This corresponds to the posterior log-probability as the feature.

3.2.2 Mean of Syllable Duration

The mean of duration of syllables in the word is used as the second feature. The duration of each syllable is normalized by two factors, that is, local mean and global mean.

The mean of syllable duration normalized by local mean is defined as follows:

$$Local\_d = \frac{1}{N} \sum_{i=1}^{N} \frac{dur(s_i)}{\sum_{j=i-3}^{i+3}dur(s_j)}$$  \hspace{1cm} (1)
where \( N \) is the number of syllables in the word, and \( \text{dur}(s) \) is the duration of the syllable \( s \). The duration of each syllable is normalized based on the preceding three syllables and the succeeding three syllables.

The mean of syllable duration normalized by global mean is defined as follows:

\[
Global_d = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{dur}(s_i)}{\sum_{s \in U} \text{dur}(s)}
\]  

(2)

where \( U \) is the set of syllables in the utterance. The duration of each syllable is normalized based on the all syllables in the utterance.

### 3.2.3 Variance of Syllable Duration

The variance of duration of syllables in the word is used as the third feature. The duration of each syllable is normalized based on all syllables in the utterance (\( Global_d \)). The definition is as follows:

\[
\text{Var}_d = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{\text{dur}(s_i)}{\sum_{s \in U} \text{dur}(s)} - Global_d \right)^2
\]  

(3)

### 3.2.4 Score of Statistical Syllable Duration Model

In [7], a statistical phone duration model is utilized to filter out corrupted user utterances for computer-assisted pronunciation training system. As motivated by their work, we build a statistical syllable duration model and use the score of this model as the fourth feature.

The definition of this feature is below:

\[
\text{Score}_d = \frac{1}{|W|} \log \left( \prod_{s \in W} \frac{P(\text{dur}(s)|s)}{P_{\text{anti-model}}(\text{dur}(s))} \right)
\]  

(4)

where \( W \) is the set of syllables in the word. Here, the statistical syllable duration model \( P(\text{dur}(s)|s) \) is the Gamma distribution trained from syllable durations obtained by the automatic alignments between exact transcriptions and their corresponding utterances. On the other hand, the anti-model \( P_{\text{anti-model}}(\text{dur}(s)) \) is also the Gamma distribution, which models the distortion of syllable duration in automatic alignment between inexact transcription and utterance. This anti-model is trained from syllable durations obtained by the automatic forced alignments between shuffled transcriptions and utterances (as this result, each transcription is aligned to uncorresponding utterance), which simulate the mismatches between transcriptions and utterances.

### 3.2.5 Word Identity

The word identity is used as the fifth feature. There are several words which have a greater tendency to be edited, such as auxiliary verbs "ます/masu" or particles "が/ga". This feature captures such tendency.

### 3.2.6 Word Length

The total number of syllables in the word is used as the sixth feature. We consider this feature acts as the negative feature based on the hypothesis that the longer word have a greater tendency to involve an edited part. For example, the word "わけど/keredomo" is often replaced by "けど/kedo", and the word "やっぱり/yaqpari" is often replaced by "はり/yahari".

### 3.2.7 Word Duration

The word duration obtained by automatic alignment is used as the seventh feature. This feature is also based on the previous hypothesis that the longer word have a greater tendency to involve an edited part.

### 3.3 Target Value for Detection Performance

In this study, we set the target value for the detection performance as 33% in precision and 50% in recall. Here, to take an example, consider a short transcription consisting of 100 words in which the 10 words are edited. In this target, we can detect 15 words in which the 5 words are truly edited with the target performance. As a result, human transcribers can find and correct the 5 edited words by checking the only 15 words. In other words, if user want to find 10 edited words, he/she needs to check only 30 candidate words. This is three times more efficient than without our detection method.

On the other hand, in this setting, 90 words in the transcription are exact part. We can detect 85 words in which the 80 words are exact part by applying our detection method. This improves the exact ratio from 90% to 94%. Such detection of exact parts can be utilized for the lightly supervised acoustic model training [2].

### 4 Experiment

In this section, we explain our evaluation experiments using the Japanese National Diet Record.

#### 4.1 Experimental Setup

The Japanese National Diet Record contains numerous inexact transcriptions in which spoken-style ex-
pressions and disfluencies such as filled pause and hesitations are edited. In this experiment, disfluencies are excluded from the target of detection, because our purpose is not to recover the disfluencies but to convert the written-style(edited) expressions to the spoken-style(exact) expressions.

We conducted the evaluation experiments in two conditions, that is, the speaker semi-closed condition and the speaker open condition. In the speaker semi-closed experiment, two speakers are included both in training set and in test set Note that the training set and the test set are parts of the same meeting. On the other hand, in the speaker open experiment, speakers are completely separated into each set. The data statistics for each condition are shown in Table 1.

As the decoder for automatic alignment and continuous syllable recognition, we used the in-house large vocabulary continuous speech recognition system, SPOJUS++ (SPoken Japanese Understanding System)[8]. The 116 Japanese context-independent syllable-based acoustic models[9] (a left-to-right topology, 4 emitting states, and single Gaussian mixture with a full covariance matrix) was trained from academic presentation speech data and simulated public speech data in the Corpus of Spontaneous Japanese[10]. The input speech was analysed with the condition as shown in Table 2.

4.2 Experimental Results

4.2.1 Accuracy of Automatic Alignment

In our preliminary experiment, we evaluated the accuracy of the automatic alignment between exact transcription and its corresponding utterance. As a result, 97.4% of accuracy was achieved with allowing errors within 30msec. Most of the alignment errors are caused by lazy or weak pronunciations by the speaker. In other case, the adjacent noise, cough, and breathe also distorted the alignments. Additionally, the alignment became comparatively difficult in the low power segment.

4.2.2 Speaker Semi-closed Experiment

The recall-precision curves of edited part detection for all 4 speakers, 2 closed speakers, 2 open speakers, are shown in Figure 4. As shown in Figure 4, there are a large difference in performance between closed speakers and open speakers. Specifically, a reasonable performance(recall=60%, precision=30%) was obtained for the closed speakers, while much worse result was shown for the open speakers. This result suggests that each speaker has its own specific characteristic of edits or the amount of the training data was insufficient. More training data are not desired for our purpose which is to detect edited parts from large-scale corpora with small amount of supervision.

A sharp degradation in the precision was observed at the range of 0% to 5% in the recall. Similar tendency was also found in other conditions as shown in the following sections. This results indicate that there are some negative instances(exact parts) which are false alamed with highly positive score for any cause. Further analyses should be performed to these degradations in our future work.

![Figure 4: Recall-Precision Curve of Detection of Edited Parts in Speaker Semi-closed Condition](image)

4.2.3 Speaker Open Experiment

The recall-precision curve for the speaker open experiment is shown in Figure 5. As shown in Figure 5, the performance is significantly lower than the speaker semi-closed condition. This results emphasizes the previous suggestion that each speaker has its own specific characteristic of edits. We consider the difference between the result in Figure 5 and the result for open-speakers in Figure 4 is caused by two factors: the difference of edited ratio, and the difference of the recording environment.

4.2.4 Feature Analysis

We analysed the effect of each feature set both in speaker semi-closed condition and speaker open condition. Here, we divided the features into three sets.

- Feature set 1: acoustic score feature.
### Table 1: Data Statistics

<table>
<thead>
<tr>
<th></th>
<th>Speaker Semi-closed</th>
<th>Speaker Open</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>Speech Length (min)</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td># of Speakers</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td># of Words</td>
<td>3.6k</td>
<td>3.6k</td>
</tr>
<tr>
<td># of Edited Words</td>
<td>347</td>
<td>257</td>
</tr>
<tr>
<td>Edited Ratio (%)</td>
<td>9.6</td>
<td>7.1</td>
</tr>
</tbody>
</table>

### Table 2: Conditions of acoustic analysis for input speech

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Rate</td>
<td>16kHz</td>
</tr>
<tr>
<td>Preemphasis</td>
<td>0.98</td>
</tr>
<tr>
<td>Analysis Window</td>
<td>Hamming Window</td>
</tr>
<tr>
<td>Analysis Frame Length</td>
<td>25ms</td>
</tr>
<tr>
<td>Analysis Frame Shift</td>
<td>10ms</td>
</tr>
<tr>
<td>Feature Parameter</td>
<td>MFCC + △MFCC + △△MFCC + △Pow + △△Pow (38 dimensions)</td>
</tr>
</tbody>
</table>

- Feature set 2: mean of syllable duration, variance of syllable duration, score of statistical syllable duration model.
- Feature set 3: word identity, word length, word duration.

The effects of adding each feature set are shown in Figure 6 and Figure 7. As shown in Figure 6 and Figure 7, the feature set 2 gave no improvement in the performance. On the other hand, adding the word-related features (features set 3) gave consistent improvement in both experiments.

#### 4.2.5 Detection of Exact Parts

We also evaluated the other experiment: the detection of exact parts. The results are shown in Figure 8 and Figure 9. As shown in Figure 8, the exact ratio without our detection method (recall=100%) is 93%. This improves to 98% by filtering the 60% of the whole transcription (recall=60%). Similarly, as shown in Figure 9, the exact ratio improves from 96% to 98% by filtering the 60% of the whole transcription for the open speaker condition. Such detection of exact parts can benefit some applications such as lightly supervised acoustic model training [2].

## 5 Conclusion

In this paper, we proposed an automatic detection method of edited parts in inexact transcribed corpora.
Figure 6: Recall-Precision Curve with Different Feature Sets in Speaker Semi-closed Condition

Figure 7: Recall-Precision Curve with Different Feature Sets for Closed Speakers

Figure 8: Recall-Precision Curve of Detection of Exact Parts in Speaker Semi-closed Condition

Figure 9: Recall-Precision Curve of Detection of Exact Parts in Speaker Open Condition
The proposed method consists of two stages. In the first stage, an automatic alignment between edited transcription and its corresponding utterance is performed. In the second stage, a support vector machine based detector is applied to detect edited parts using some features obtained by the automatic alignment in the first stage.

The evaluation experiment using the Japanese National Diet Record showed that our proposed method achieved a reasonable performance in the speaker-closed condition.

As the future work, we plan to conduct a larger-scaled experiment, which includes evaluation of speaker-dependent and speaker-independent edited part detector. Additionally, we are going to evaluate how human transcriber can find and correct the edited part efficiently by applying our proposed method in the experiment with human subjects.

As a by-product of correction of edited part in inexact transcribed corpora, we can collect the parallel corpora of the spoken-style (exact) expressions and written-style (edited) expressions. This parallel corpora can be utilized for many applications such as speaking style transformation of language model[1] and lightly supervised training of acoustic model[11].

Acknowledgements: We would like to thank the Global COE program "Frontiers of Intelligent Sensing" for supporting our research.

References:


