Correct parts extraction from speech recognition results using semantic distance calculation, and its application to speech translation

Yumi WAKITA, Jun KAWAI[*] Hitoshi IIDA
ATR Interpreting Telecommunications Research Laboratories

Abstract

This paper proposes a method for extracting the correct parts from speech recognition results by using an example-based approach for parsing those results that include several recognition errors. Correct parts are extracted using two factors: (1) the semantic distance between the input expression and example expression, and (2) the structure selected by the shortest semantic distance. We examined the correct parts extraction rate and the effectiveness of the method in improving the speech understanding rate and the speech translation rate. The examination results showed that the proposed method is able to efficiently extract the correct parts from speech recognition results. About ninety-six percent of the extracted parts are correct. The results also showed that the proposed method is effective in understanding misrecognition speech sentences and in improving speech translation results. The misunderstanding rate for erroneous sentences is reduced about half. Sixty-nine percent of speech translation results are improved for misrecognized sentences.

1 Introduction

In continuous speech recognition, N-grams have been widely used as effective linguistic constraints for spontaneous speech [1]. To reduce the search effort, N of a high-order can be quite powerful; but making the large corpus necessary to calculate a reliable high-order N is unrealistic. For a realistic linguistic constraint, almost all speech recognition systems use a low-order N-gram, like a bi-gram or tri-gram, which can be constrainted only to the local parts. However this is one of the reasons why many misrecognized sentences using N-grams are strange on long parts spanning over N words. During the recognition process, several candidates have to be pruned if the beam width is too small, and the pruning cannot but use only those local parts already recognized. Even if we could get a large enough corpus to train a high-order N-gram, it would be impossible to determine the best recognition candidate in consideration of the whole sentence. To put a speech dialogue system or a speech translation system into practical use, it is necessary to develop a mechanism that can parse the misrecognized results using global linguistic constraints.

Several methods have already been proposed to parse ill-formed sentences or phrases using global linguistic constraints based on a context-free-grammar (CFG) framework, and their effectiveness against some misrecognized speech sentences have been confirmed [2, 3]. Also these parsings are used for translation (see for example the use of the GLR parser in Janus[4]). In these studies, even if the parsing was unsuccessful for erroneous parts, the parsing could be continued by deleting or recovering the erroneous parts. The parsing was done on the assumption that every input sentence is well-formed after all erroneous parts are recovered. In reality, however spontaneous speech contains a lot of ill-formed sentences and it is difficult to analyze every spontaneous sentence by the CFG framework. Concerning the CFG framework, syntactic rules written by subtrees are proposed [5]. Even if a whole sentence can not be analyzed by CFG, the sentence can be expressed by combining several subtrees. The subtrees are effective in parsing spontaneous speech parts. Still, because the subtrees can deal only with local parts like in N-gram modeling basically, parsing is not sufficient for parsing misrecognized sentences. Furthermore, the subtrees are not sufficient in extracting suitable meaningful candidate structures, because that these linguistic constraints are based on the grammatical constraint without semantics.

[*] Now working at Toyo Information Systems Co., Ltd

To parse misrecognized sentences of spontaneous speech, we propose a correct parts extraction (CPE) method that uses global linguistic and semantic constraints by an example-based approach.

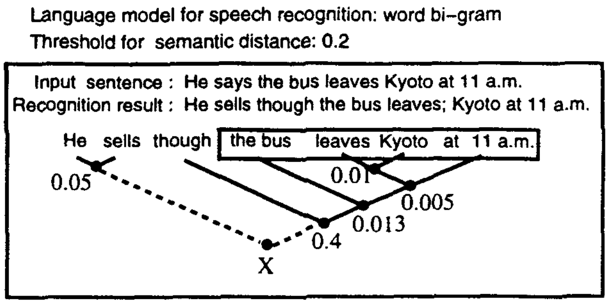
In the next section, we describe the CPE method. In the following section, we show evaluation results of CPE applied to Japanese-to-English speech translation experiments.

2 Correct Parts Extraction using Constituent Boundary Parser

2.1 Constituent Boundary Parser (CB-parser)

For effective and robust spoken-language translation, a speech translation system called Transfer Driven Machine Translation (TDMT) which carries out analysis and translation in an example-based framework has been proposed[6]. TDMT which refers to as Example-Based Machine translation(EBMT)[7] does not require a full analysis and instead defines patterns on sentences/phrases expressed by "variables" and "constituent boundaries". These patterns are classified into several classes, for example a complex sentence pattern class, an embedded clause pattern class, and phrase class. A long-distance dependency structure can be handled by complex sentence patterns. The process employs a fast nearest-matching method to find the closest translation example by measuring the semantic conceptual distance of a given linguistic expression from a set of equivalents in the example corpus.

In general, the EBMT method is particularly effective when the structure of an input expression is short or well-defined and its bounds have been recognized. When applying it in translation of longer utterances, the input must first be chunked to determine potential patterns by analyzing it into phrases after adding part-of-speech tags. In TDMT, translation is performed by means of stored translation examples which are represented by "constituent boundary patterns". These are built using limited word-tag information, derived from morphological analysis, in the following sequence[6]: (a) insertion of constituent boundary markers, (b) derivation of possible structures by pattern matching, and (c) structural disambiguation using similarity calculation[8].



Language model for speech recognition: word bi-gram
Threshold for semantic distance: 0.2

Input sentence : He says the bus leaves Kyoto at 11 a.m.
Recognition result : He sells though the bus leaves; Kyoto at 11 a.m.

Figure 1: Example of correct part extraction

If the process of the similarity calculations for candidate phrase patterns were executed top-down & breadth-first, then the calculation cost would be too expensive and the decision on the best phrase would have to be postponed. The translation cost are reduced in TDMT and phrases or partial sentences are analyzed because that the current TDMT uses instead on incremental method to determine the best structure locally in a bottom-up & best-only way to constrain the number of competing structures. This means that even TDMT fails for a whole sentence analysis, substructures partially analyzed can be gotten.

2.2 Correct Parts Extraction

Our proposed correct parts extraction (CPE) method obtains correct parts from recognition results by using the CB-parser. CPE uses the following two factors for the extraction: (1) the semantic distance between the input expression and an example expression, and (2) the structure selected by the shortest semantic distance.

The merits of using the CB-parser are as follows.

- The CB-parser can analyze spontaneous speech which can not be analyzed by the CFG framework, only if the example expressions are selected from a spontaneous speech corpus. With more expressions in spontaneous speech, there is an increased ability to distinguish between erroneous sentences and correct ones.

- The CB-parser can deal with patterns including over N words which can not be dealt with during speech recognition. (see Table 5).

- The CB-parser can extract some partial structures independently from results of

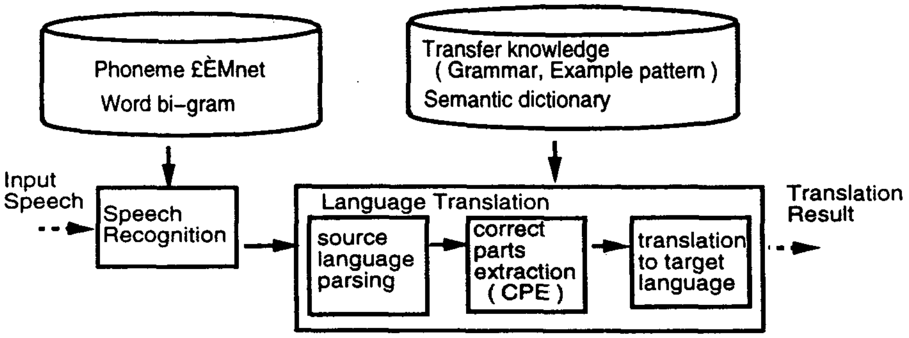


Figure 2: Speech translation system using CPE

parsing, even if the parsing fails for a whole sentence.

Correct parts are extracted under the following conditions:

- When expressions including erroneous words show big distance values to the examples. When the distances are over the distance threshold, the parts are defined as "erroneous parts".
- Correct parts are extracted only from global parts consisting of over N words. If local parts including less than N words can not have a relation to other parts, the parts are defined as "erroneous parts", even if the semantic distances are under the threshold.

Figure 1 shows an example of CPE. The input sentence /He says the bus leaves Kyoto at 11 a.m./ is recognized as /He sells though the bus leaves Kyoto at 11 a.m./ by continuous speech recognition using a word bi-gram. The solid lines in Figure 1 indicate partial structures and the number for each structure denotes the corresponding semantic distance value. The dotted line indicates the failure analysis result. In this example, the analysis for the whole sentence is unsuccessful because the part /He says/ is misrecognized as /He sell though/. At first, the distance value of the longest part, /though the bus leaves Kyoto at 11 a.m./, is compared with the threshold value . The part is considered to include erroneous words because the distance value 0.4 is larger than the threshold value 0.2 . Secondly, the next longest part /the bus leaves Kyoto at 11 a.m./ is evaluated. This part is extracted as a correct part because the distance 0.005 is under the threshold value. Thirdly, the remaining part /He sells/ is evaluated. The distance of the part /He sells/ is under the threshold value, but the part includes only two words which are under N, so the part /He sells/ is regarded as an erroneous part.

3 Evaluation

We evaluated CPE using the speech translation system shown in Figure 2. CPE has already been integrated into TDMT as explained in the previous section. At first, the obtained recognition results were analyzed and then partial structures and their semantic distances were output. Next, the correct parts were extracted and only the extracted parts were translated into target sentences.

We evaluated the following three things: (1) the recall and precision rates of the extracted parts , (2) the effectiveness of the method in understanding misrecognized results, and (3) the effectiveness of the method in improving the translation rate. For the evaluations, we used 70 erroneous results output by a speech recognition experiment using the ATR spoken language database on travel arrangement [10].

3.1 Rate of correct parts extraction

To evaluate CPE, we compared the recall and precision rates after extraction to the same rates before extraction. Recall and precision are defined as follows:

$$\text{recall} = \frac{\text{number of correct words in extracted parts}}{\text{number of words in the correct sentence}}$$

$$\text{precision} = \frac{\text{num. of correct words in extracted parts}}{\text{num. of words in the recognition results}}$$

The extraction defines the threshold for the number of words in the structure to be N+1, on the assumption that the semantic distances of the local parts consisting of under N words

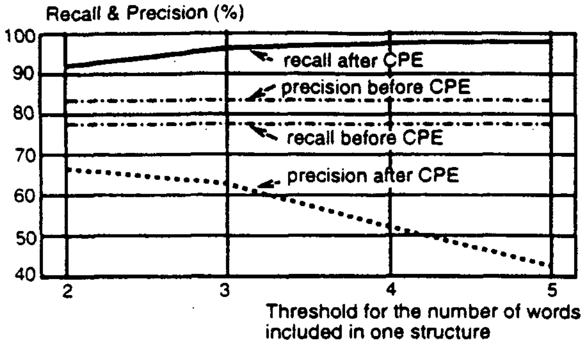


Figure 3: Relationship between the extraction rate and the number of words in a structure

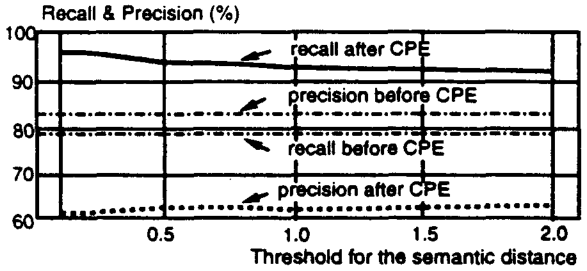


Figure 4: Relationship between the extraction rate and the threshold of the semantic distance

are not useful for determining whether the parts are correct or not. To confirm whether the assumption is true or not, extraction experiments were performed under variable threshold conditions for the number of words in the structure. Figure 3 shows the obtained recall and precision rates.

- The recall rates under all conditions are over 92% and the best recall rate is 97%. This indicates that the rates increased over 15% from before the extraction.
- The precision rates show a decrease of over 20% from before the extraction. This means that some correct parts could not be extracted.
- When the threshold is two, the recall rates decrease much more than when the threshold is over three.
- When the threshold is over four, the precision rate deceases a lot.

Furthermore, extraction experiments were performed under variable threshold values of the semantic distance for examining the relation between the threshold for the semantic distance and the rate of correct parts extraction. The recall and precision rates are shown in Figure 4.

- There is a general trend that when the threshold increases, the recall rate decreases and the precision rate increases. But the differences of these rates are less than the differences by changing the threshold of the number of words as shown in Figure 3. In particular, the precision rate changes only slightly.
- When the threshold is defined as below 0.2, the recall and precision rates do not change.

These results show the following;

- Words extracted by CPE are almost the real correct words.
- The threshold for the number of words should be defined as over three when a "BI" gram is adopted, because the recall rates decrease when the threshold is two. It therefore seems the assumption is true that local parts consisting of under N words are not useful for determining the correct parts.
- The best threshold condition for the number of words is three in consideration of both the recall and the precision. Under this condition, the recall rate is typically 96% and the precision rate is typically 63%.
- The best threshold condition for the semantic distance is 0.2 , because when the threshold is defined as over 0.2, the recall rate decreases.

3.2 Effect to speech understanding

To confirm the effectiveness of CPE in understanding speech recognition sentences, we compared the understanding rate of extracted parts using CPE with the rate of the recognition results before extraction. The same 70 erroneous sentences as in the previous experiments were used. The threshold for the number of words was defined as three and the threshold for the semantic distance was defined as 0.2, which were confirmed to be the best values in Figure 3 and Figure 4. The recognition results were evaluated by five Japanese. They gave one of the following five levels (L1)-(L5) to each misrecognition result before extraction and after extraction, by comparing the results with the corresponding correct sentence before speech recognition. The five levels were:

(L1) Able to understand the same meaning as the correct sentence.

(L2) Able to understand, but the expression is slightly awkward.

(L3) Unable to understand, but the result is helpful in imagining the correct sentence.

(L4) Understanding of the wrong meaning. CPE is not helpful.

(L5) Output of the message "Recognition impossible."

Each of the average rates of the five evaluators is shown in Table 1. CPE was effective in reducing the misunderstanding rate over half (35.5% to 15.2%). The results able to be understood which are given (L1) and (L2) increased but only a little (19.6% to 20.3% for (L1), 22.0% to 22.6% for (L2)) by using CPE. The tendency was that most of the misrecognition sentences including only negligible errors could be understood even without CPE, because the evaluators could see the errors themselves while reading the misrecognition results. On the other hand, most of the misrecognition sentences that included many erroneous parts were understood incorrectly. The proposed CEP was very effective here in preventing misunderstandings. Nonetheless, other additional mechanisms seem necessary, like an error recovering mechanism that increases the number of understandable sentences.

3.3 Effect to speech translation

We evaluated the effectiveness of CPE in Japanese-English speech translation experiments using the speech translation system shown in Figure 2. The conditions for the database, and the threshold values for the CPE method were the same as in the previous experiments. The translation results were evaluated by three Japanese each with a high ability to converse in the English language. They gave one of five levels (L1)-(L5) to each translation result of the misrecognized sentences, by comparing the result with the corresponding translation result of the correct sentence before speech recognition. (L1)-(L4) for the evaluations were the same as in the previous experiments and (L5) meant "Cannot translate".

Each of the average rates of the three evaluators is shown in Table 2.

Without CPE, 85.7% of the recognition results could not be translated. It seems that CPE is good for (L1)-(L3) but poor for (L4); (L5) shows negligible effect. The correctness rate for translation after CPE is more than double the rate before CPE (11.9% to 25.7%). The sum of (L1)-(L3) is 69%. This means that the proposed CPE is effective in improving the translation performance. However, we cannot ignore the fact that 21% of the recognition results were translated to erroneous sentences.

4 Discussions

Some deletion errors of function words are solved by TDMT even without CPE. This is because the translation trains a lot of the spontaneous speech in which identical function words had been deleted. On the other hand, CPE is effective for many erroneous sentences. Important misrecognition characteristics effectively handled by CPE are as follows:

(a) Some insertion errors between words

(b) Errors at the tail parts of sentences

(c) Strange expressions including over N words

(d) Expressions not similar to examples

(e) Input too complicated to parse (but not errors)

In contrast, characteristics not effectively handled by CPE are as follows:

(f) Errors of final parts causing ambiguity, e.g, of a person, of a situation, whether a sentence is negative or positive, or whether a sentence is interrogative or affirmative. In these cases, the translation results are incorrect even if CPE is used.

Table 3 - Table 7 show examples for each of the characteristics. The top sentence of each table is the input sentence and the second sentence is the recognition result; the final word sequences are only parts extracted from the recognition results. All of the words are Japanese words expressed in Roman characters and the words or sentences in brackets are the translated English equivalents.

4.1 Insertion errors

Filled-pauses, e.g., "umm" or "well", are often spoken in spontaneous speech. Many speech recognition systems deal with filled-pauses as

Table 1: The effect of CPE toward understanding misrecognition results

Levels	(L1)	(L2)	(L3)	(L4)	(L5)
without CPE	19.6%	22.0%	23.0%	35.5%	0.0%
after CPE	20.3%	22.6%	36.8%	15.2%	5.4%

Table 2: The effect of CPE toward translating misrecognition results

Levels	(L1)	(L2)	(L3)	(L4)	(L5)
without CPE	11.9%	0%	0%	2.4%	85.7%
after CPE	25.7%	16.7%	26.6%	21.0%	10.0%

recognized words. Many Japanese filled-pauses consist of only one phoneme, e.g., "e","q", or "n", and it is easy for mismatches to parts of other words to occur. Furthermore, filled-pauses have no strong relations to any words and it is difficult to constrain them with an N-gram framework. These are the reasons why insertion errors of filled-pauses are often found in misrecognized results.

Table 3 is an example of insertion errors by filled-pauses. For this example, a structure analysis for the whole sentence failed. However, the parts before and after the filled-pauses, /deNwa(telephone) baNgou(number) wa/ and /go(five) ni(two) nana(seven)/ could be extracted as correct parts. The two words /kyuu(nine)/ and /desu(is)/ could not be extracted because the part /kyuu desu/ included only two words.

4.2 Errors at the tail parts of sentences

For an indirect expression or an honorific expression, several function words are often spoken successively at the final part of the sentence. Misrecognition often occurs at this part. When the words necessary for understanding an utterance have been spoken before the final part, it is possible to perform translation to an understandable sentence by extracting only the beginning parts. Table 4 shows an example of an error occurring at a final part /N desu keredomo/. The part /N desu keredomo/ is part of an honorific expression and all of the words in this part are function words. The proposed extraction selects only the beginning part /heya no yoyaku wo onegai sitai(would like to reserve a room)/. The translation result is a little strange but it can be understood and almost has the correct meaning. Actually, only /I/ could not be translated because the misrecognized part /N desu keredomo/ included a keyword to determine the person.

4.3 Strange expression consisting of over N words

Table 5 shows an example of a strange expression consisting of over N words. In this example, every word pair is not strange because all of them have already been constrained by bigram modeling. But the expression consisting of three words i.e., /oyako(parent and child) no gokibou(preference)/ is strange. The part /oyako no/ can be said to be an erroneous part because it can be connected to other parts and consists only of two words.

4.4 Expressions not similar to examples

The important merit of the example-based approach is that any structural ambiguity or semantic ambiguity can be reduced in consideration of the similarity to examples. The recognition result shown in Table 6 was misrecognized in the part /ii(am)/ to /i(stay)/. But the misrecognized result /Suzuki Naoko to i masu (I am staying with Suzuki Naoko)/ is very natural in general. It seems therefore that CFG can parse an erroneous sentence without any problem and the sentence can be understood although with a different meaning. (/I am staying with Suzuki Naoko/ which is different from the correct meaning /I am Suzuki Naoko/). However, this is rare for a travel arrangement corpus and the semantic distance value of the whole sentence is over the threshold. As a result of CPE, only /Suzuki Naoko/ can be extracted and translated to /Naoko Suzuki/.

4.5 An utterance including several sentences

Even if a recognition result is correct, when one utterance includes several sentences, TDMT without CPE sometimes fails because the boundary of the sentences can not be understood, for example. /waka ri masi ta (I see). doumo arigatou (Thank you)/. Though the translation fails without CPE, CPE can extract each sentence one by one and the translation result after CPE is correct.

4.6 Expression of bad effect by CPE

The keywords for determining whether a sentence is negative or positive, or whether a sentence is interrogative or affirmative, are often spoken at the final part of the sentence. When these keywords are misrecognized, the translation result is quite different from the correct translation result. The input sentence in Table 7 is a negative sentence. The keyword determining the sentence to be negative is /naku/, but is misrecognized. As a result of the translation after CPE, a positive sentence is translated and the meaning is opposite to the intended meaning.

5 Conclusion

This paper proposed a method for extracting correct parts from speech recognition results in order to understand recognition results from speech inputs which may include erroneous parts. Correct parts are extracted using (a) the semantic distances between the input expression and an example expression and (b) the structure selected by the shortest semantic distance.

We examined three things: (1) the correct parts extraction rate, (2) the effectiveness of the method in improving the speech understanding rate. and (3) the effectiveness of the method in improving the speech translation rate. Results showed that the proposed method is able to efficiently extract the correct parts from speech recognition results; ninety-six percent of the extracted parts are correct. The results also showed that the proposed method is effective in preventing the misunderstanding of the erroneous sentences and in improving the speech translation results. The misunderstanding rate for erroneous sentences is reduced over half and sixty-nine percent of the speech translation results can be improved for misrecognized sentences.

In the future, we will try to feed the extraction results back into the speech recognition process for re-recognizing only the non-extracted parts and to improve the speech recognition performance. By repeating the correct parts extraction and the feedback, we will confirm whether there is an improvement in the understanding and translation performance. Furthermore, we will confirm the effectiveness of the proposed method using other languages.

References

[1] L.R.Bahl, F.Jelinek and R.L.Mercer: "A Maximum Likelihood Approach to Continuous Speech Recognition," In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp.179-190. 1983.

[2] C.S.Mellish:"Some chart-based techniques for parsing ill-formed input.", In *proc. of the Annual Meeting of the ACL*, pp.102–109, 1989.

[3] H.Saitou,M.Tomita:"Parsing noisy sentences,", In *proc. of COLING'88*, pp.561–566. 1988.

[4] A.Lavie, D.Gates, M.GAvalda, L.Mayfield, A.Waibel, and L.Levin:"Multilingual Translation of Spontaneously Spoken Language in a Limited Domain" In *Proc.of 16th ICCL*, pp.442–447, 1996.

[5] T.Takezawa,T.Morimoto:"Dialogue Speech Recognition Method using Rules based on Subtrees and Preterminal Bigrams" In *IEICE Trans in Japanese*, D-II Vol.J79-D-II No.12 pp.2078-2085. 1996.

[6] O.Furuse, H.Iida:"Constituent Boundary Parsing for Example-Based Machine Translation" In *proc. of COLING'94*, pp.105–111. 1994.

[7] E.Sumita, H.Iida:"Experiments and Prospects of Example-based Machine Translation" In *Proc. of 29th ACL*, pp.185–192, 1991.

[8] E.Sumita, H.Iida:"An Example-Based Disambiguation of English Prepositional Phrase attachment" In *Proc. Syst. and Com. in Japan* Vol.26, No.4, pp.30-41, 1995.

[9] O.Furuse, H.Iida:"Incremental Translation Utilizing Constituent Boundary Patterns" In *proc. of COLING'96*, pp.412–417. 1996.

[10] T.Morimoto et al.: "A Speech and language database for speech translation research" In *Proc. of ICSLP'94*, pp.1791–1794, 1994.

Table 3: Example of insertion errors between words

Input sentence	deNwa baNgou wa go ni nana kyuu desu (The telephone number is five two seven nine)
Recognition result / / : insertion errors	deNwa baNgou wa /q/ /o/ go ni nana /aq/ kyuu desu (telephone)(number) (five) (two) (seven) (nine) (is)
Result after CPE ... : non-extracted parts	deNwa baNgou wa .... go ni nana ...... (The telephone number .... five two seven .....)

Table 4: Example of errors at the final part of a sentence

Input sentence	heya no yoyaku wo onegai sitai N desu keredomo. (I would like to reserve a room.)
Recognition result _ : erroneous parts	heya no yoyaku wo onegai sitai ne su tomo (room) (reserve) (would like to)
Results after CPE ... : non-extracted parts	heya no yoyaku wo onegai sitai .... (... would like to reserve a room)

Table 5: Example of a strange expression over N words

Input sentence	oheya no gokibou wa gozai masu ka ? (room) (preference) (Do you have any preference for a room ?)
Recognition result _ : erroneous parts	oyako no gokibou wa gozai masu ka ? (parent and child) (preference)
Result after CPE ... : non-extracted parts	.... gokibou wa gozai masu ka (Do you have any preference ......?)

Table 6: Example of an expression not similar to the example sentences

Input sentence	Suzuki Naoko to ii masu (Suzuki)(Naoko) (I am) (I am Naoko Suzuki)
Recognition result _ : erroneous parts	Suzuki Naoko to i masu (Suzuki)(Naoko) (stay)
result after CPE ... : non-extracted parts	Suzuki Naoko to ..... (..... Naoko Suzuki)

Table 7: Example of bad effect by CPE

Input sentence	tsugou de tomare naku natta (reason) (stay) (can't) (I can't stay for some reason)
Recognition result _ : deletion errors	tsugou de tomare ___ natta (reason) (stay)
result after CPE ... : non-extracted parts	tsugou de tomare .... can stay for some reason