ABSTRACT

The utterance units that serve as input to speech translation and/or spoken dialogue systems that handle spontaneous speech are not always sentences. However, the processing units of language translation are sentences. Since we do not have enough knowledge about the sentences of spoken languages, we use the term “language processing units” instead of “sentences.” First, using conventionally interpreted dialogue data, we show that utterance units sometimes need to be divided into several language processing units, and sometimes need to be connected to make up a single language processing unit. Next, we propose a method of transforming utterance units into language processing units based on pause information and the N-gram of fine-grained part-of-speech subcategories. We confirm in experiments that our method yields good results.

Section 2 shows the necessity of transforming language processing units by dividing and connecting utterance units. Section 3 describes the proposed method, which is based on pause information and the N-gram of fine-grained part-of-speech subcategories. Section 4 describes experiments and results, section 5 is a brief discussion, and section 6 gives our conclusions.

2. NECESSITY

2.1. Definition of Terms

We assumed that “utterance units” mean inputs to speech translation and/or spoken dialogue systems that handle spontaneous speech as well as computational processing units by a streaming speech detector. Our end-point detection (EPD) module was a streaming speech detector able to detect the start of speech within about 50 ms, but the detection of the end of the speech was much longer: almost one second [3].

We assumed that “language processing units” mean computational processing units by automatic language translation systems. We introduced a partial translation mechanism [9] into ATR-MATRIX [5], and our speech-to-speech translation system accepted speech recognition results including recognition errors. The processing units of our language translation subsystem [10] in ATR-MATRIX were expected to be sentences.

2.2. Overview of Our Bilingual Travel Conversation Database

We collected conversations between Japanese and English speakers through human interpreters in speech translation research [7, 8]. We assigned an interpreter to each translation direction (J to E or E to J) when collecting one conversation in order to gather good quality data (upper side of figure 1). The human interpreters successively interpreted each utterance so we could gather basic data for developing “a speech translation system in the near future.”

The definition of utterance and turn in our database was as follows (lower side of figure 1).

1. A speaker could only speak when it is his/her turn. When a speaker wanted to stop speaking and listen to a response, he/she could transfer the right of utterance to the other speaker.
2. Each utterance had to be concluded in ten seconds or less, and was then sent to an interpreter.
3. The interpreter translated the speaker’s utterance and conveyed it to the other speaker. The speaker’s utter-
2.3. Necessity of Utterance Division

Here we show examples of utterance divisions from our bilingual travel conversation database. Pause information is one of the most important factors for utterance division so the information on pause lengths is described in [] . The utterance consists of two language processing units in both examples (1) and (2).

1) O-mata-se itashishi mashi ta. [440 ms] Shinguru ippaku ichi-man-gen no o-haya deshi ta ne.
   Thank you for waiting. [440 ms] You'd prefer a single room for ten thousand yen per night, right?

2) O-haya wo shirabe masu. [170 ms] Shiyaraku o-machi kudasai.
   I will check the rooms. [170 ms] Please wait for a moment.

We carried out a preliminary experiment on automatic pause detection using a part of our bilingual travel conversation database. By using two kinds of features, i.e., the logarithm of powers and the number of zero crossings, and assuming pauses to be silent periods more than or equal to 300 ms, we were able to detect all of the pauses without confusing them with double long consonants.

We can divide example (1) into two language processing units using only automatically detected pause information if we assume the length of pauses to be more than or equal to 300 ms. However, we cannot divide example (2) into two language processing units using only pause information because 170 ms is less than 300 ms. Therefore, both pause information and additional information are necessary to divide utterances into language processing units.

2.4. Necessity of Utterance Connection

Here we show examples of utterance connections from our bilingual travel conversation database.

3) (a) Kaado no bangoo ga. [680 ms] go ni nana kyuu.
   The card number is [680 ms] five two seven nine.
   (b) San kyuu ni zero.
   Three nine two zero.
   (c) Ni yon roku kyuu.
   Two four six nine.
   (d) Zero zero kyuu hachi [410 ms] de gozai masu ne.
   Zero zero nine eight [410 ms] right?

   In example (3), one turn consists of four utterances. The human interpreter translated each English equivalent for each Japanese utterance. Human interpreters were able to translate this type of speaking style. However, it may be necessary for automatic translation systems to connect such types of utterance into single units.

4) O-mata-se itashishi mashi ta. [1600 ms] Youshitsu wa, [1130 ms] ippaku ni-shoku tsuki, [450 ms] ni-man-gen de, hoyo beddo ga hea masu.
   Thank you for waiting. [1600 ms] The western room <topic marker> [1130 ms] dinner and breakfast are included for each night, [450 ms] It will be twenty thousand yen, and an extra bed is available.

   Example (4) is one utterance. Our EPD module in ATR-MATRIX assumes that the utterance ends if there is a pause of more than one second. Therefore, one utterance in example (4) consists of three utterance units using our EPD module as follows. Utterance unit (4) (b) must be connected to utterance unit (4) (c).
Table 2. Transformation from utterance units into language processing units

<table>
<thead>
<tr>
<th></th>
<th>Long pause exists</th>
<th>Long pause does not exist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical model succeeds</td>
<td>(A) insert a period</td>
<td>(B) insert a period</td>
</tr>
<tr>
<td>Statistical model fails</td>
<td>(C) insert a comma</td>
<td>(D) do nothing</td>
</tr>
</tbody>
</table>

(1) O-mata-se utter. Thank you for waiting.
(b) Youshita wa.
The western room <topic marker>
(c) iipaku ni-shoku tsuki. [450 ms] ni-man-yen de, hojo beddo ga hari masu.
dinner and breakfast are included for each night, [450 ms] It will be twenty thousand yen, and an extra bed is available.

3. PROPOSED METHOD

We propose a method of transforming utterance units into language processing units based on pause information and the N-gram of fine-grained part-of-speech subcategories. Table 2 shows a summary of our model. We insert a comma in (C) in table 2, and do nothing in (D) in table 2.

For the statistical model, we adopt a modified model of the statistical clause boundary predictor proposed in [5]. The original model proposed in [5] refers to two words before the current position and two words after the current position. The model calculates the clause boundary score according to the following equation, where $\bullet$ indicates the current position, $w_1$ and $w_2$ indicate the two words before the current position, and $w_3$ and $w_4$ indicate the two words after the current position.

$$F([w_1w_2\bullet w_3w_4]) = \frac{C([w_1w_2\bullet]) + C([w_2\bullet w_3]) + C([\bullet w_3w_4])}{C([w_1w_2]) + C([w_2w_3]) + C([w_3w_4])}$$

where $C([w_1w_2\bullet])$ is the count of clause boundaries on the right side of the bigram $[w_1w_2]$, $C([w_1w_2])$ is the total counts of the bigram $[w_1w_2]$ in the training data, and so on.

The statistical clause boundary predictor assumes the clause boundary if the value of $F$ is bigger than a threshold. The value of the threshold is adjusted to make the best performance using a development test set.

We adopt a modified model as follows.

$$F([w_1w_2\bullet w_3]) = \frac{C([w_1w_2\bullet]) + C([w_2\bullet w_3])}{C([w_1w_2]) + C([w_2w_3])}$$

This statistical model refers to two words before the current position and one word after the current position. In the following section, we concentrate on Japanese and examine the following three kinds of combinations.

1. Only part-of-speech (POS)
2. Part-of-speech (POS), inflection forms and types of inflection words
3. Surface expressions, part-of-speech (POS), inflection forms and types of inflection words

4. EXPERIMENTS AND RESULTS

4.1. Utterance Division Using Japanese Transcription Files

According to our bilingual travel conversation database [7, 8], utterance units often need to be divided into several

Table 3. Utterance division using Japanese transcription files

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical model</td>
<td>85.6%</td>
<td>65.7%</td>
</tr>
<tr>
<td>Statistical model with heuristics</td>
<td>97.7%</td>
<td>99.4%</td>
</tr>
</tbody>
</table>

Table 4. Utterance division using Japanese speech recognition results

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical model with heuristics</td>
<td>95.6%</td>
<td>94.8%</td>
</tr>
</tbody>
</table>

language processing units. As a result, we first conducted utterance division experiments using Japanese transcription files.

Our bilingual travel conversation database consists of 618 conversations as shown in table 1. We selected nine conversations for the test set. The test set contained 166 turns of 216 utterances. We used the remaining 609 conversations to train the statistical model. The training was conducted in turns.

Table 3 shows a summary of the experiments. The number of target boundaries was 123. Equation (2) was better than equation (1). The performance of the model of part-of-speech (POS), inflection forms and types of inflection words was the best among the three above-mentioned combinations. The statistical model in table 3 indicates the best performance. After analyzing errors and false alarms, we found that they were covered by patterns of statistically rare cases. We introduced these patterns as heuristics. The recall and precision performance using the statistical model with heuristics improved significantly as shown in table 3. The heuristics that we introduced are as follows.

- If a conjunctive postpositional particle does not follow an interjection, we set a boundary to the position.
- If a conjunctive postpositional particle follows an interjection, we do not set a boundary to the position.
- We do not set a boundary to the position between a finished form of auxiliary verb and a sentence final postpositional particle.

4.2. Utterance Division Using Japanese Speech Recognition Results

We conducted Japanese speech recognition experiments using the speech recognizer reported in [11]. The test set was the same as the one with nine conversations. The inputs to the speech recognizer were the utterances because speech wave data files were digitized per utterance. Table 4 shows the results of the utterance division using speech recognition results. The number of target boundaries was 73. The word recognition accuracy of top candidates was 77.4% (the word correctness rate was 83.6%).
6. CONCLUSIONS

We have reported a transformation of utterance units into language processing units to build a speech-to-speech translation system. First, we showed that utterance units sometimes need to be divided into several language processing units, and sometimes need to be connected to make up a single language processing unit. Next, we propose a method of transforming utterance units into language processing units based on pause information and the N-gram of fine-grained part-of-speech subcategories. We conducted experiments and confirmed that our method yields good results. The technique reported in this paper has been introduced into the ATR-MATRIX speech-to-speech translation system from Japanese to English [3].

ACKNOWLEDGMENTS

The author wishes to thank Ms. Naoko Ohtsuki for her contribution to the experiments.

REFERENCES


Table 5. Preliminary experiment of Japanese utterance connection

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical model</td>
<td>81.0%</td>
<td>85.4%</td>
</tr>
<tr>
<td>Statistical model with heuristics</td>
<td>90.3%</td>
<td>86.3%</td>
</tr>
</tbody>
</table>

Table 6. Preliminary experiment of English utterance division

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical model</td>
<td>70.0%</td>
<td>64.1%</td>
</tr>
</tbody>
</table>

4.3. Utterance Connection Using Japanese Transcription Files

Our EPD module in ATR-MATRIX assumes that the utterance ends if there is a pause of more than one second. Therefore, one utterance from the above example (4) was automatically divided into the above example (4’) by our EPD module. In the same test set of nine conversations, there were 20 samples that had to be connected, but long pauses of more than one second existed. When we applied our statistical model to them, we classified all of them into (C) in table 2.

An additional experiment was also carried out. Utterances of the same test set of the nine conversations were artificially divided into utterance units using an automatic pause detector. There were 99 samples that had to be connected but small pauses of more than 10 ms existed. Table 5 shows the result of the utterance connection experiment using these artificial samples.

4.4. Utterance Division Using English Transcription Files

We conducted preliminary experiments of utterance division in English using the parallel parts of the same test set of nine conversations. We used the remaining 699 conversations to train the statistical model. The training was conducted in turns. The targets were limited to sentence boundaries. We examined the following four types of combination.

1. Only part-of-speech (POS)
2. Part-of-speech (POS), inflection forms and types of inflection words
3. Surface expressions, part-of-speech (POS), inflection forms and types of inflection words
4. Only surface expressions

Table 6 shows the results of this experiment. Equation (1) was better than equation (2). The performance of the model of surface expressions, part-of-speech (POS), inflection forms and types of inflection words was the best among the four types of combination.

5. DISCUSSION AND FUTURE RESEARCH

The performance of Japanese is better than that of English even when using equation (2). This is because a sentence’s final expression in Japanese has more distinguishing characteristics than that in English. We will conduct further investigations on utterance connections of Japanese and transformation from utterance units to language processing units of English.