Construction of Chunk-Aligned Bilingual Lecture Corpus for Simultaneous Machine Translation

Masaki Murata †, Tomohiro Ohno ‡‡, Shigeki Matsubara † and Yasuyoshi Inagaki ‡‡‡

†Graduate School of Information Science, Nagoya University,
‡‡Graduate School of International Development, Nagoya University,
‡‡‡Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan
murata@el-itc.nagoya-u.ac.jp, {ohnoh,matubara}@nagoya-u.jp

Abstract

With the development of speech and language processing, speech translation systems have been developed. These studies target spoken dialogues, and employ consecutive interpretation, which uses a sentence as the translation unit. On the other hand, there exist a few researches about simultaneous interpreting, and recently, the language resources for promoting simultaneous interpreting research, such as the publication of an analytical large-scale corpus, has been prepared. For the future, it is necessary to make the corpora more practical toward realization of a simultaneous interpreting system. In this paper, we describe the construction of a bilingual corpus which can be used for simultaneous lecture interpreting research. Simultaneous lecture interpreting systems are required to recognize translation units in the middle of a sentence, and generate its translation at the proper timing. We constructed the bilingual lecture corpus by the following steps. First, we segmented sentences in the lecture data into semantically meaningful units for the simultaneous interpreting. And then, we assigned the translations to these units from the viewpoint of the simultaneous interpreting. In addition, we investigated the possibility of automatically detecting the simultaneous interpreting timing from our corpus.

1. Introduction

With the development of speech and language processing, speech translation systems have been developed (Frederking et al., 2002; Arranz et al., 2005; Gao et al., 2006; Nakamura et al., 2006). These studies target spoken dialogues, and employ consecutive interpretation, which uses a sentence as the translation unit. On the other hand, there exist a few researches about simultaneous interpreting (e.g. (Ryu et al., 2006)), and recently, the language resources for promoting simultaneous interpreting research, such as the publication of an analytical large-scale corpus (Matsubara et al., 2002), has been prepared. For the future, it is necessary to make the corpora more practical toward realization of a simultaneous interpreting system.

In this paper, we describe the construction of a bilingual corpus which can be used for simultaneous lecture interpreting research. Simultaneous lecture interpreting systems are required to recognize translation units in the middle of a sentence, and generate its translation at the proper timing. We constructed the bilingual lecture corpus by the following steps to develop a simultaneous lecture interpreting system.

1. We segmented sentences in the lecture data into semantically meaningful units for the simultaneous interpreting.
2. We assigned the translations to these units from the viewpoint of the simultaneous interpreting.

In addition, we investigated the possibility of automatically detecting the simultaneous interpreting timing from our corpus.

1. Segmentation of an input sentence into suitable translation units (segmentation).

This paper is organized as follows: Section 2 discusses a method of simultaneous lecture interpreting. Section 3 describes the design and the construction of a chunk-aligned bilingual lecture corpus. Section 4 reports corpus-based analyses on automatic detection of interpreting timing.

2. Simultaneous lecture interpretation

A simultaneous lecture interpreting system is required to output the translation result simultaneously with the input utterance. Since a sentence in a lecture tends to be long basically, it is necessary for the system to adopt shorter language units than sentences as translation units. We are suggesting a method of simultaneous lecture interpreting, as shown in Figure 1, consisting of the following three steps:

1. Segmentation of an input sentence into suitable translation units (segmentation).
2. Translation of each translation unit (translation).
3. Concatenation of these translated segments so that the translations form a natural English sentence (concatenation).

These steps work simultaneously with the speech input. Figure 2 shows an example of simultaneous translation process based on this approach. In this example, the following input Japanese sentence:

しはなきとれよすでしの出發が遅れる可能性がありますので、ご了承くだ

is segmented into four units “しはなきとれよすでしの,” “出発が遅れる可能性がありますので,” and “ご了承くださいませ,” in the middle of the input. At the same time as segmentation, the system translates the units into English phrases “for now,” “it is on time,” “the departure might be delayed,” and “please understand it,” respectively. Then, the system concatenates each translation result and generates the following English sentence:

For now, it is on time, but the departure might be delayed. Please understand it.

Table 1: Size of segmented Japanese data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sentences</td>
<td>1,935</td>
</tr>
<tr>
<td>morphemes</td>
<td>60,829</td>
</tr>
<tr>
<td>bunsetsus</td>
<td>23,598</td>
</tr>
<tr>
<td>clauses</td>
<td>9,664</td>
</tr>
<tr>
<td>chunks</td>
<td>8,644</td>
</tr>
<tr>
<td>chunks per sentence</td>
<td>4.47</td>
</tr>
<tr>
<td>bunsetsus per chunk</td>
<td>2.73</td>
</tr>
</tbody>
</table>

In addition, this database includes the speech of interpretations by professional interpreters and their transcribed texts. However, such the interpretations are not always suitable as the data for current machine translation technologies because simultaneous interpretations under real environment may include loose translations of original sentences. Therefore, we assigned the renewed translations to this data.

3.1. Segmentation of the lecture text

We have segmented lecture texts into several shorter units than sentences by hands. In this paper, we call this unit a chunk. We set the following concepts as the chunk:

- Not so long: If the length of chunks gets long, the simultaneity is decreased because it takes much time to start the translation process. Also, it is desired that the length of chunks is uniform so that the delay of translation is kept constant.
- Semantically meaningful: It is desired that a chunk is semantically meaningful because the translation needs to be generated for each chunk.

We defined the maximum length of a chunk as 4.3 sec by considering the delay in the actual interpretations by the professional interpreters (Ono et al., 2008), and we segmented sentences into chunks according to this restriction. Table 1 shows the size of the segmented Japanese spoken monologue data. As a result, 1,935 sentences in the database were segmented into 8,644 chunks (4.47 chunks per sentence). In addition, we have already confirmed that the chunk boundaries can be detected automatically with about 80% of precision (Ohno et al., 2009).
3.2. Translation of chunks
We constructed the bilingual corpus by assigning the translations to each chunk. The translations were provided by professional translators who are familiar with interpretations. Though it is ideal to assign one translation to each chunk, every chunk can not be always translated by itself. The translators provided a translation to each chunk basically, but if a chunk was not able to be translated by itself, the translators translated such chunk together with chunks following it.

Figure 4 shows an example of the bilingual corpus. In this example, a chunk “それから千九百五十六年には” was translated into “Then, in 1956” by itself. The other hand, for example, a chunk “より強くなったということが” was not translated by itself. So, this chunk was translated into “can be said, I think, to have become stronger.” together with a chunk “いえると思います” following it.

Table 2 shows the size of the translation data.

4. Analysis of interpreting timing
We tried to assign one translation to one chunk at the construction of our corpus. However, there existed chunks which were not able to be translated by itself. So, it is not always appropriate that the system adopts a chunk as a translation unit. We investigated when the translation of a certain chunk had been generated. Concretely, we measured the number of chunks that had been observed by the time the translations were generated. Figure 5 shows the result. There exist 5,662 chunks which were able to be translated when these were observed, and its percentage of total is 65.50%. Also, 85.67% of all chunks was able to be translated in case that the next chunk was observed.

To identify translation units based on our corpus, it is necessary to decide whether to generate the translation whenever a chunk boundary is detected. We analyzed the timing with which the translation was generated at chunk boundaries. In this paper, we call this timing simultaneous interpreting timing. We focused on the pause, clause boundary and dependency relation as the available information in the automatic analysis of simultaneous interpreting timing. Here, 65.50% (5,662/8,644) of all chunk boundaries were simultaneous interpreting timing in our corpus. This is the standard ratio of simultaneous interpreting timing on chunk boundaries.

4.1. Pauses and interpreting timing
Pauses can be detected automatically when these were inserted. Since the pauses could correspond to syntactic boundaries. Therefore, pauses might be useful for detecting simultaneous interpreting timing. Table 3 shows the relation between pauses and simultaneous interpreting timing. The ratio that a chunk boundary having a pause was a simultaneous interpreting timing was 75.65% (4,526/6,444), and this ratio was higher than that of chunk boundaries (65.50%). This indicates that pauses are useful to detect
それから千九百五十六年には
ソ連との共同宣言によりまして
日ソ間の国交が回復しております
平和条約の締結は
その後に残されておりますが
一応
戦争状態は終わったわけでございます
それから千九百六十年には
日米の新しい安保条約が締結されております
安保条約の上でわが国の発言権が
より強くなったということが
いえます
それから千九百六十四年になりますと
オーウーシーに加盟しております
これは純粋に戦後処理というよりは
その次の新しい飛躍の時期への
助走と申しますか
準備の時期だったと思います
それから千九百六十五年には
韓国との国交が正常化しております
そして千九百六十八年には
それまでアメリカが占領しておりました
小笠原が返還され
千九百七十二年には沖縄が戻されました
先程申し上げましたように
中国との国交正常化もできた
そういう時期でございます

それから千九百五十六年には
ソ連との共同宣言によりまして
日ソ間の国交が回復しております
平和条約の締結は
その後に残されておりますが
一応
戦争状態は終わったわけでございます
それから千九百六十年には
日米の新しい安保条約が締結されております
安保条約の上でわが国の発言権が
より強くなったということが
いえます
それから千九百六十四年になりますと
オーウーシーに加盟しております
これは純粋に戦後処理というよりは
その次の新しい飛躍の時期への
助走と申しますか
準備の時期だったと思います
それから千九百六十五年には
韓国との国交が正常化しております
そして千九百六十八年には
それまでアメリカが占領しておりました
小笠原が返還され
千九百七十二年には沖縄が戻されました
先程申し上げましたように
中国との国交正常化もできた
そういう時期でございます

Then, in 1956
under the Japan-Soviet Union joint declaration
diplomatic relations between Japan and the Soviet were normalized.
The conclusion of the Peace Treaty
was later postponed, but
provisionally,
the state of the war ended.
Then, in 1960
the Japan-U.S. Security Treaty, the new one, was concluded.
Our nation’s right to speak during the Security Treaty
can be said, I think, to have become stronger.
And in 1964
Japan joined the OECD.
It wasn’t purely the disposal of the postwar period,
rather
an approach run for the next step up, or
I suppose, a preparation period.
In 1965
diplomatic relations between Japan and Korea were normalized.
In 1968
Ogasawara, having been occupied by the U.S., was returned.
In 1972, Okinawa was returned from the U.S. to Japan.
As I mentioned before,
the diplomatic relations between Japan and China were also normalized.
That was exemplary of such a period.

Figure 4: Example of the bilingual corpus

Figure 5: Relation between the number of observed chunks and generation of translations

4.2. Clause boundaries and interpreting timing

A clause is one of semantically meaningful language units including one verb phrase and corresponds to a simple sentence. The variation in the length of clauses is smaller (for example, than that of sentences), and clause boundaries can be detected using the local morphological information with high accuracy (Kashioka et al., 2003). Therefore, the clause boundaries may be useful for detecting simultaneous interpreting timing. Table 4 shows the relation between clause boundaries and simultaneous interpreting timing. Among chunk boundaries which were also clause boundaries, 77.61% of them were the simultaneous interpreting timing. Thus, we confirmed the usefulness of clause boundaries for detecting simultaneous interpreting timing. However, there exist several types of clause boundary, and the role of each clause on a sentence is different by the types of clause boundaries. We investigated the ratio that a chunk boundary which was also a clause boundary was a simultaneous interpreting timing. Table 5 shows the top 10 clause boundary types about the occurrence frequency and their
There existed clause boundaries which were simultaneous interpreting timing with the ratio over 85%, such as “compound clause -keredomo,” and “condition clause -to,” besides “end of a sentence.” On the other hand, in case of “supplement clause” and “adnominal clause,” the ratio that such clause boundaries which simultaneous interpreting timing was less than 40%. This means that the likelihood that the chunk boundaries were simultaneous interpreting timing is different according to the types of the clause boundary.

### 4.3. Dependency structure and interpreting timing

A dependency relation is a modification relation in which a modifier bunsetsu depends on a modified bunsetsu. We focused on the dependency relation in which a bunsetsu depends on the next bunsetsu. In case that a bunsetsu depends on the next bunsetsu, the chunk boundaries existing between them may be hard to be a simultaneous interpreting timing because the sequence of such the bunsetsus forms a semantically meaningful unit. Table 6 shows the relation between dependency structure and simultaneous interpreting timing. In case that a bunsetsu did not depend on the next bunsetsu, the ratio that the chunk boundaries between them were simultaneous interpreting timing was 72.93% (5,235/7,178). This indicates that the dependency structure could be used for detecting the simultaneous interpreting timing.

### 5. Conclusion

This paper has described the construction of a simultaneous lecture interpreting corpus. We have constructed the corpus by assigning translations for simultaneous interpreting to chunks, which are made by segmenting the Japanese lecture data. Among all chunks, about 65% of them were translated when they occurred. Therefore, we confirmed that this data is useful for developing simultaneous lecture interpreting systems.

In the future, we will study on techniques for deciding simultaneous interpreting timing and concatenating the translation result by using our corpus.

### 6. Acknowledgments

This study was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 20300058) of JSPS and by the Continuation Grants for Young Researchers of The Asahi Glass Foundation.

### 7. References
