Construction and Utilization of Bilingual Speech Corpus for Simultaneous Machine Interpretation Research

Hitomi Tohyama†, Shigeki Matsubara‡, Nobuo Kawaguchi‡, and Yasuyoshi Inagaki∗

†Graduate School of Information Science, Nagoya University, Japan
‡Information Technology Center, Nagoya University, Japan
∗Faculty of Information Science and Technology, Aichi Prefectural University, Japan
hitomi@el.itc.nagoya-u.ac.jp

Abstract

This paper describes the design, analysis and utilization of a simultaneous interpretation corpus. The corpus has been constructed at the Center for Integrated Acoustic Information Research (CIAIR) of Nagoya University in order to promote the realization of the multi-lingual communication supporting environment. The size of transcribed data is about 1 million words, and the corpus would deserve to be called the simultaneous interpretation corpus of the largest-in-the-world class. The discourse tag and the utterance time tag were given to the corpus, and some software tools for corpus analysis in order to support the practical use of the corpus have been developed. Therefore, the corpus is expected to be useful not only for the development of simultaneous interpreting systems but also for the construction of an interpreting theory.

1. Introduction

Recently, spoken language corpora have been constructed for the purpose of studying on speech processing in many research organizations (for example [1, 2]). The large-scale corpora are recognized to be important widely, and used in various research areas, such as speech recognition, natural language processing, linguistics, language education, and dictionary compilation. At the Center for Integrated Acoustic Information Research of Nagoya University (following, CIAIR), a corpus of simultaneous interpretation between Japanese and English has been constructed over five years. We aim at the realization of the multi-lingual communication supporting environment. The recording time is 182 hours in total. The speech data has been all transcribed and visualized. Furthermore, we have completed language analysis of the corpus. The size of transcribed data is about 1 million words, and the corpus would deserve to be called the simultaneous interpretation corpus of the largest-in-the-world class. Additionally, we have developed some software tools for corpus analysis in order to support the practical use of the corpus. They have been developed as software which can be performed on the Web server, and a user can refer to the corpus easily by using a browser.

This paper describes the design, collection, construction, analysis, and utilization of the simultaneous interpretation corpus. In the following section, we describe the purpose and the design of the simultaneous interpretation corpus. In Section 3, we describe the recording of the corpus. In Section 4, the construction of the corpus is explained in full detail. Section 5 discusses the use of the corpus.

Figure 1: Recording environment

2. Design of the Corpus

2.1. Aim of Data Collection

Machine interpretation has become one of the most important research topics with the advance of technologies for speech processing and language translation. Several experimental systems of spoken dialogue translation for specific task domains have been developed [3, 4, 5]. The interpreting style of them is within so-called consecutive interpretation. In order to provide an environment to support natural and smooth cross-lingual communication, to develop a technique for simultaneous machine interpretation has been awaited and tried out recently. Not only a generation of translation but an outputting timing of translation is required for a simultaneous interpreting system. It would be effective to investigate and analyze the interpreting process of professional simultaneous interpreters. The CIAIR is constructing and maintaining various types of speech and language database for the purpose of the advancement of robust speech information processing technology [6]. Moreover, a bilingual database of simultaneous interpretation has also been constructed as a part of this project. We aim to develop speech translation technologies and to construct interpreting theories.

2.2. Policy of Corpus Design

The large-scale corpus needs to be equipped with flexibility, because many researchers are expecting to utilize the corpus for their purposes. Therefore, we collected both monologue and dialogue data. The contents of the database are daily topics. The database targets English and Japanese.

3. Recording of Speech Data

3.1. Recording Environment

One of the purposes of CIAIR is to collect a large quantity of speech data which were generated under natural circumstances, so the recording took place in a classroom. Facial expressions and conversational behaviors of speaker's are also important information for simultaneous interpretation, thus, the interpreters were in booths from which they could see the speakers as Fig.1 shows. Both speakers and interpreters used the same cross-talking microphones. The speeches were digitized by sampling frequencies of 16 kHz and 16 bits, and recorded onto digital audio tapes (DAT) in multiple channel environments. All interpreters are professional interpreters who are active in the front lines.

3.2. Recording of Monologue Speech Data

Simultaneous interpreters go into a booth, and interpret the lecturer's speeches from the headphone. Although the lecturers face to the audience, they cannot hear the interpreter' speech. It enables them to speak at their own paces. The contents of speeches are economics, history, and, culture, etc. Moreover, each monologue speech is interpreted by two or four professional interpreters. Their degree of experience differs from one another (5 years over or not). The flexibility of a database is raised by using four interpreters. Therefore, it becomes possible to compare two or more interpretation examples in a sample of utterance. Moreover, we can compare interpreters' utterance speed, speaking timing, and strategy of interpretation. The speech was recorded for about 10 minutes per lecture.

3.3. Recording of Dialogue Speech Data

Travel conversation was selected as a domain of conversation, which includes popular topics during overseas travel at airports and hotels, and simulated conversations are recorded. To put it concretely, the following topics were selected: "airport check in", "hotel check-in/check-out", "booking of a room at a hotel", and "booking of a seat in an airplane", and so on. In order to enhance the quality of interpretations for both English speakers and Japanese speakers, each speaker was accompanied by one interpreter. To ensure all the participants' speech pretension, speakers can listen only to the output from the other speaker's interpreter, and the interpreters can listen only to the speech that they are assigned to interpret. Please note that these dialogues are simulative, in which the contents of the speeches can be limited. In attempting to collect utterances as unfettered as possible, such background information as the speaker's roles and conversational tasks were informed to speakers in advance. For a speaker who is a customer of a hotel, for example, the kind and number of rooms that should be reserved, and for a speaker as front desk clerk, rooms that can be reserved, etc. We set up "airport" and "hotel" as the typical situations of dialogue communications doing overseas travel. The recoding time per one dialogue was from 1 minute to 16 minites, and dialogues of various types were collected.

4. Construction of the Corpus

4.1. Transcription of Speech Data

The transcription was produced based on the standard transcription rules of t he Corpus of Spoken Japanese (CSJ) developed by the National Japanese Language Research Institute [7]. Figure 2 and 3 show the sample data of English speaker's utterances and that of English-Japanese interpreter's utterances, respectively. All the speech data (182 hours) were transcribed into a text. The standardization is shown as follows:

```
0001 - 00:05:264-00:09:399 N:
The theme for this speech is going to be the American
0002 - 00:09:840-00:11:032 N:
Presidential debate
0003 - 00:11:424-00:13:391 N:
and who would be the
0004 - 00:13:640-00:15:215 N:
better president for America<SB>
0005 - 00:16:272-00:18:327 N:
(F um) Let's see, today is
0006 - 00:18:640-00:20:400 N:
December fifteenth
0007 - 00:20:696-00:24:407 N:
and it's been about a month and a half since
```

Figure 2: Sample of the transcribed text: English speaker's utterance

```
0001 - 00:06:660-00:08:355 I:
次のスピーチのテーマは
0002 - 00:09:344-00:10:071 I:
アメリカの
0003 - 00:10:816-00:13:439 I:
大統領選に対するディベートです<SB>
0004 - 00:14:048-00:14:623 I:
誰が
0005 - 00:15:184-00:17:055 I:
アメリカにとって良い大統領なのか<SB>
0006 - 00:19:224-00:21:648 I:
今日が十二月の十五日です<SB>
0007 - 00:22:784-00:25:247 I:
約一か月半経ちました<SB>
0008 - 00:27:152-00:27:983 I:
アメリカ
0009 - 00:28:432-00:32:303 I:
の(W 大(D りょ)統領選;大統領選)が始まってから一か月半です<SB>
```

Figure 3: Sample of the transcribed text: English-Japanese interpreter's utterance

- Utterance unit
  Utterance units were set by 200ms-or-longer pauses in the speech of speakers and interpreters.
- Notation
  Recorded Japanese speech is transcribed in two different ways: orthographic and phonetic transcriptions.
- Tag annotation
  - Utterance ID
    A serial number was given to each utterance unit.
  - Time tag
    The beginning time and end time of the utterance units were tagged.
  - Discourse tag
    Language tags were also added onto fillers, hesitations, and corrections.

4.2. Visualization of Speech Data

We developed a timing information visualization tool. The speaking time of English lecturers, English-Japanese interpreters, Japanese lecturers and Japanese-English interpreters and their overlapping relations are displayed as Fig.4 shows.

Table 1: Statistics of monologue data

item		data size (words)	data size (utterances)	recording time (min.)
speaker	English	90249	8422	695
	Japanese	84278	6529	597
	Total	174527	14951	1292
interpreter	EJ	266050	25507	1639
	JE	127991	16083	1265
	Total	394041	41590	2904
Sum total		568568	56541	4196

Table 2: Statistics of dialogue data

item		data size (words)	data size (utterances)	recording time (min.)
speaker	English	107850	14223	1678
	Japanese	106258	16485	1678
	Total	214108	30708	3356
interpreter	EJ	116776	15286	1678
	JE	91743	13719	1678
	Total	208519	29005	3356
Sum total		422627	59713	6712

Thereby, we can visually observe a overlap of a lecturer and an interpreter utterance.

4.3. Alignment of the Bilingual Speech

For a detailed analysis of interpreters' speech, such as extraction of temporal characteristics of interpretation, the acquisition of translation patterns, the detection of translation units and so forth, it is necessary to align utterances of speakers and interpreters with relatively small units. An alignment support tool working on the internet, which Fig.5 shows, has been developed by using CGI script. The users can align the utterance units by the clicking the mouse on the bilingual text displays. The aligned data can be used for analysis of interpreting units and timing. We have aligned the corpus using the tool according to the following conditions:

- The unit should be the minimum alignment unit
- Utterances should be aligned as small as possible
- Utterance units such as fillers or non-language phenomena and the utterances with no appropriate counterparts can have no correspondence.

As stated above, a detailed alignment standard was established, so, the annotation work was made uniform.

4.4. Environment Information

A large-scale corpus has various availabilities. Information that doesn't appear at speech data and text data might be important. Therefore, we provided the Environment information of recording data file for every session. Environment information includes date, location, recording time, audio-video equipment, topic, type of speech, the speaker's roles and conversational tasks, the information on speaker, the information on interpreter (years of experience, etc).

Figure 4: Visualization of the binlingual speech

Figure 5: Alignment of the binlingual speech

4.5. Statistics on the Corpus

The large-scale corpus involving 1 million words have been developed; we have finished recording the data of 182 hours of speech, transcribing it into text, attaching discourse tags, and matching source utterances to their target utterances, so far.

5. Utilization of the Corpus

To develop the simultaneous machine interpretation, it is necessary to analyze the interpreting process of professional simultaneous interpreters. We have researched simultaneous interpreters' utterance timing, interpretation unit, and generation of translation and compared the simultaneous interpretation with the consecutive one in cross-lingual communication. We have proved the effectiveness of simultaneous interpretation technology. Furthermore, if the data for analysis is large-scale, it is also possible to verify a qualitative analysis result still more a quantitative one. This section describes the main research results performed by using the corpus.

5.1. Analysis of Interpreter's Speaking Timing

Simultaneous interpretation may overlap with the corresponding native speech. It is expected that an interpreter recognizes

a part of a lecturer's utterance as an interpreting unit and interprets it at an early stage. We have investigated the interpreting units and speaking timing of professional interpreter by analyzing the aligned corpus. The summary of results is shown below:

- Since a subject appears at the beginning of a sentence in both Japanese and English, the subject can be interpreted immediately.
- By controlling the out putting speed of system based on the quantity of the input utterance. It is possible to reduce the difference between the beginning time of the interpreter's utterance and that of the lecturer's utterance.

5.2. Analysis of Temporal Features

Interpretation has two styles: consecutive interpretation and simultaneous interpretation. One of major differences between them is whether an interpreter starts to speak after the speaker completed his/her utterances, which is consective interpretation, or before, which is simulteneous interpretaion [8]. Simulteneous interpretaion occurs that the speaker's utterance and interpreter's utterance temporally overlap each other; however in the consective interpretaion, those utterances doesn't. We have done the further research on these two styles of interpretation.

We have compared simultaneous interpretation with consecutive one in cross-lingual communication and we have proved the possibility that simultaneous interpretation technology performs more effectively. Our study proved how effective the simultaneous interpretation is by analyzing the actual simultaneous interpretation data. It focused the on the efficiency and the smoothness of cross-language conversation through simultaneous interpretation. The summary of results is shown below:

- The growth of average dialogue time growth rate on simultaneous interpretation was twice as much as consecutive interpretation, which indicates that conversational efficiency through simultaneous interpretation has been raised considerably in comparison with consecutive interpretation. This tendency can be seen in English-Japanese interpretation.
- The average of speakers' waiting time on conversations through interpreting is 4.4 seconds for English speakers on simultaneous interpretation, and 15.4 seconds on consecutive interpretation. That is 4.3 seconds for Japanese speakers on simultaneous interpretation, and 14.5 on consecutive interpretation, which indicates that smoothness increases drastically.

The result proved the usability of simultaneous interpreting technology as a support environment for cross-language dialogues because in the dialogues through simultaneous interpretation.

5.3. Collection of Interpretation Strategies

Simultaneous interpretation is advanced language processing activities of human. Simultaneous interpreters have to generate their translations simultaneously with original speech. However, they have the restrictions on speaking timing (when-to-say) and how to translate speaker's utterance (how-to-say). However, they have various kinds of strategies to raise simultaneity. In this investigation, the interpreting patterns used frequently and having both/either high flexibility and simultaneity were extracted from a bilingual spoken monologue corpus. The CIAIR corpus has as many of four interpreter data per one monologue section. Therefore, it is possible to collect two or more interpretation patterns from one speaker's utterance. Those extracted interpretation patterns can be used as a rule of the interpretation system. It is possible to develop the system by using the rule.

6. Conclusion

This paper has described the design, analysis and utilization of the CIAIR simultaneous interpretation corpus of Nagoya University. We expected that the corpus will be used for not only the development of simultaneous interpreting systems but also the construction of an interpreting theory. We are going to develop the CIAIR corpus further by giving it more detail tags and constructing alignment data. In those days, the spoken language technology has progressed. Therefore, the demand for large-scale spoken database is rising in not only speech processing but also cognitive science, phonology, and linguistics. We hope that CIAIR corpus will be used in the various research fields. It is preferable to exchange the opinion between different areas and to progress overall.

The CIAIR corpus has been already exhibited. For more details, please refer to the following URL:

`http://slp.el.itc.nagoya-u.ac.jp/sidb/`

7. Acknowledgments

The authors would like to thank Prof. Dr. Toyohide Watanabe and all members of EL project for their precious advices. This work is partially supported by the Grand-in-Aid for Exploratory Research (No. 17652040) of JSPS.

8. References

[1] Maekawa, K., Koiso, H., Furui, S. and Isahara, H., "Spontaneous Speech Corpus of Japanese," Proc. of 2nd International Conference on Language Resources and Evaluation, pp.947–952, 2000.

[2] Takezawa, T., Nakamura, A. and Sumita E., "Database for Conversational Speech Translation Research at ATR," The Phonetic Society of Japan, Vol.4, pp.16–23, 2000.

[3] Mima, H., Iida, H. and Furuse, O., "Simultaneous Interpretation Utilizing Example-based Incremental Transfer," Proc. of 17th Computational Linguistics and 36th Association for Computational Linguistics, pp.855–861, 1998.

[4] Watanabe, T., Okumura, A., Sakai, S., Yamabana, K., Doi, S. and Hanazawa, K., "An Automatic Interpretation System for Travel Conversation," Proc. of 6th International Conference on Spoken Language Processing, Vol. IV, pp. 44–68, 2000.

[5] Amtrup, J. "Incremental Speech Translation," Lecture Note in Artificial Intelligence, Vol. 1735, 1999.

[6] Kawaguchi, N., Matsubara, S., Takeda, K. and Itakura, F., "Multi-Dimensional Data Acquisition for Integrated Acoustic Information Research," Proc. of 3rd International Conference on Language Resources and Evaluation, pp. 2043–2046, 2002.

[7] Maekawa, K., "Corpus of Spontaneous Japanese: Its design and evaluation," Proc of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.

[8] Gile, D., "Consecutive vs. Simultaneous: which is more accurate?," The Journal of the Japan Association for Interpretation Studies, No.1, pp. 8–20, 2001.