Collection of a Simultaneous Translation Corpus for Comparative Analysis

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan
{hiroaki-sh,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

Abstract

This paper describes the collection of an English-Japanese/Japanese-English simultaneous interpretation corpus. There are two main features of the corpus. The first is that professional simultaneous interpreters with different amounts of experience cooperated with the collection. By comparing data from simultaneous interpretation of each interpreter, it is possible to compare better interpretations to those that are not as good. The second is that for part of our corpus there are already translation data available. This makes it possible to compare translation data with simultaneous interpretation data. We recorded the interpretations of lectures and news, and created time-aligned transcriptions. A total of 387k words of transcribed data were collected. The corpus will be helpful to analyze differences in interpretations styles and to construct simultaneous interpretation systems.

Keywords: simultaneous translation, simultaneous interpretation, corpus collection

1. Introduction

While the translation performance of automatic speech translation (ST) has been improving, ST has mainly been used in consecutive translation situations such as conversation, where ST translates only after the speaker has finished speaking. To move beyond this scenario, there are several works about real-time ST (Ryu et al., 2004; Fügen et al., 2007; Bangalore et al., 2012; Fujita et al., 2013; Sridhar et al., 2013b) that automatically divide longer sentences up into a number of shorter ones. On the other hand, there have also been some works on constructing simultaneous interpretation databases (Toyama et al., 2004; Paulik and Waibel, 2009; Sridhar et al., 2013a). Toyama et al. (2004) constructed a simultaneous interpretation database (SIDB) that has a total of 182 hours of voice recordings including English-Japanese/Japanese-English simultaneous interpretation. Paulik et al. (2009) collected simultaneous interpretation data from European Parliament Plenary Sessions (EPPS), and conducted experiments using simultaneous interpretation data from English to Spanish. Rangarajan et al. (2013a) also used the European Parliament Interpreting Corpus (EPIC) (Claudio and Annalisa, 2005) which is a trilingual (Italian, English and Spanish) corpus of European Parliament speeches. These databases are useful not only to analyze how simultaneous interpreters translate but also to construct real-time ST systems.

This paper describes the collection of a corpus of simultaneous interpretation data[1] for use in analysis and development of real-time ST systems. There are two features of this corpus that distinguish it from related works. First, we collect data from three interpreters with different amounts of experience. Paulik et al. and Rangarajan et al. did not consider the amount of experience, and Toyama et al. have many interpreters data of different amounts of experience, but each interpreter covers only one lecture. In our database, all lectures have interpretation data from all three interpreters. As a result, it is easy to compare the interpretation of interpreters of different levels. Second, for part of the data, we can compare interpretations to translations. We use English lectures that have been subtitled in Japanese, making it possible to compare the translation data (i.e. subtitles) with the simultaneous interpretation data.

Our corpus now contains lectures and news in English-Japanese/Japanese-English with speech data and transcripts. The size of transcribed data totals 387k words. In this paper, we describe the collection of the source-language materials, the interpretation process, and the recording and transcription of the resulting data.

2. Material

The simultaneous interpreters interpret four kinds of material: TED, CNN, CSJ and NHK. The details of the material are shown in Table 1.

TED: TED[2] is a series of talks that address a wide range of topics within the research and practice of science and culture. We focus on TED because its format and breadth make it an attractive testbed for broad-coverage speech translation systems. Another reason is that many of the TED talks already have Japanese subtitles available. This makes it possible to compare data created by translators (i.e. the subtitles) with simultaneous interpretation data we collect.

We took two precautions to maintain interpretation performance. First, we chose the TED talks from nine to sixteen minutes, as if the simultaneous interpreters interpret a long speech, they may lose their concentration, damaging performance. Second, we adjusted the topics of the TED talks. Each interpreter interpreted six talks per day, five of which being general domain

[1] The corpus is available at http://ahclab.naist.jp/resource/stc/
[2] http://www.ted.com/

Table 1: A summary of the data. For TED, the S rank interpreter interpreted more lectures.

Data	Domain	Format	Lang	Number	Minutes (avg.)	Words (avg.)
TED (S rank)	Lectures	Video	English	46	558 (12.1)	98,034 (2,131)
TED (A, B rank)	Lectures	Video	English	34	415 (12.2)	70,228 (2,066)
CSJ	Lectures	Voice	Japanese	30	326 (10.9)	85,042 (2,835)
CNN	News	Voice	English	8	27 (3.4)	4,639 (580)
NHK	News	Voice	Japanese	10	16 (1.8)	4,121 (412)

topics and one of which being a specialized topic. This is because the content of the more specialized topics are high level with many technical terms.

CSJ: The corpus of spontaneous Japanese (CSJ) (Maekawa, 2003) is a corpus of academic lectures and staged talks on more general topics. The interpreters interpret in real time from Japanese to English while listening to the lecture. To maintain the interpretation performance, as with TED, we have the interpreters interpret six lectures per day, five of which are mock lectures and one of which is an academic lecture.

CNN: CNN radio news[3] is an American news station's radio channel. We chose news because it is generally more difficult than lectures, allowing us to compare performance of each interpreter under highly difficult situations.

NHK: We also use the data of NHK radio news[4]. NHK is Japan's national public broadcasting organization. NHK is similar to CNN, being the most representative news channel in Japan.

3. Recording of Simultaneous Interpretation Data

3.1. Interpreters

Three simultaneous interpreters cooperated with the recording. The profile of the simultaneous interpreters is shown in Table 2. The most important element of the interpreter's profile is the length of their experience as a professional simultaneous interpreter. Each interpreter is assigned by rank decided by years of experience. By comparing data from simultaneous interpreters of each rank, it is likely that we will be able to collect a variety of data allowing us to compare better translations to those that are not as good. Note that all of the interpreters work as professionals of both directions between English and Japanese, and have a mother tongue of Japanese. They interpret meetings and lectures in their actual work.

3.2. Environment

The simultaneous interpreters go into a booth, and interpret speech coming in from an earphone. A shotgun microphone is used for recording the interpreter's voice.
For TED, the interpreters interpret in real time from English to Japanese while watching and listening to the TED videos. The reason we prepare the video is that it makes the translation quality better when interpreters have not only the audio information (content of talks and voice) but also the visual information (expressions, gestures, and slides). In particular, the interpreters noted that viewing the slides improved the quality of interpretation. For all data other than TED, there is no associated video, so we use only voice data.
We give the simultaneous interpreters a document related to the talks in advance. The motivation for this is that in real interpretation situations, interpreters are almost always given materials to study. To approximate this, we give a document with the summary of the talk and specialized terminology for TED and CSJ. For CNN and NHK the documents include a complete transcript of the news instead of just a summary, as the news does not deviate from the transcript and just reads the transcript quickly.

[3]http://cnnradio.cnn.com/
[4]http://www3.nhk.or.jp/

Table 2: Profile of simultaneous interpreters

Experience	Rank
15 years	S rank
4 years	A rank
1 year	B rank

0001 - 00:20:393 - 00:25:725
So I'm going to present, first of all, the background of my research and purpose of it
0002 - 00:26:236 - 00:27:858
and also analytical methods.
0003 - 00:28:397 - 00:30:828
Then (F ah) talk about my experiment.

0001 - 00:44:107 - 00:45:043
本日は<H>
0002 - 00:45:552 - 00:49:206
みなさまに(F え)難しい話題についてお話ししたいと思います。
0003 - 00:49:995 - 00:52:792
(F え)みなさんにとっても意外と身近な話題です。

Figure 1: Example of a transcript in English and Japanese with annotation for time, as well as tags for fillers (F) and disfluencies (H)

4. Transcription and Annotation

4.1. Transcript

After recording the simultaneous interpretation, a transcript is made from the recorded data. The Japanese transcript is made by transcription criteria for the corpus of spontaneous Japanese (Koiso et al., 2000). The English transcript

Table 3: Examples of comparing the translation and simultaneous interpretation data in TED

	Sentence
Source	but this understates the seriousness of this particular problem because it doesn't show the thickness of the ice
Reference (translator)	しかし / もっと深刻な / 問題 / というのは / 実は / 氷河の厚さなのです *but / more serious / problem / is / in fact / the thickness of the ice*
Reference (S rank)	しかし / これ本当は / もっと深刻で / 氷の厚さまでは / 見せてないんですね *but / this is really / more serious and / the thickness of the ice / it isn't shown*
Reference (A rank)	この / 本当に / 問題に / なっているのは / 氷の厚さです *this / real / problem / becoming is / the thickness of the ice*
Reference (B rank)	この / 問題は *this / problem is*

Table 4: Examples of comparing among simultaneous interpretation data in CSJ

	Sentence
Source	私共は乳児が音楽をどのように聞いているかまた聴取に発達齢差が見られるかを検討しております
Reference (S rank)	what we research on is how infants listen to music and if there's any age difference in terms of listening abilities
Reference (A rank)	we would like to introduce how the important to listen to music and is there any difference according to the age
Reference (B rank)	how the infants listen to the music or that there is a differences of the development ages we this is the research object

is made by the following rules.

- Filled pause should be enclosed by filler tags.
- Mark where utterances are stretched out.
- Sentences always be closed with periods or question marks. Commas may be used when necessary.
- Speech errors or unclear statement is made, add the correct form and the actual utterance.

An example of the transcript is shown in Figure 1. Each talk is divided into utterances using pauses of 0.5 seconds or more, and each talk is annotated with content, an ID, the start/end time and discourse tags (e.g. fillers and disfluencies).

An example of transcripts of the three interpreters is shown in Table 3 and 4. In Table 3, we can see that the higher rank interpreter can interpret most of details. For example, the S rank interpreter can interpret the phrase "seriousness," but the A and B rank interpreters cannot. Especially, for this sentence, the B rank interpreter has trouble translating at all. In Table 4, we can see that the higher ranked interpreter can generally achieve more accurate interpretation results than those of the others. For example, the B rank interpreter cannot interpret grammatically. Looking at the S and A rank, the S rank is better than the A rank, because the S rank can interpret the phrase "乳児" to "infants," but the A rank cannot. The data enables to analyze these and similar differences in interpretation.

4.2. Automatic Evaluation

The automatic evaluation accuracy of the three simultaneous interpreters with respect to the translation data is shown in Table 5. We use three lectures (the number of sentence is 523, and number of words in the reference is 13,864) and manually align the sentences. We can see that the translation accuracy of the S rank is best in the three from the view of all automatic evaluation measures. However, we can also see that the B rank interpreter unexpectedly exceeds the A rank interpreter . Basically, the performance of simultaneous interpretation can be evaluated by various factors (not only quality of translation but also acoustic information and the starting timing of interpretation). If we would like to know the real whole performance of simultaneous interpreters, subjective evaluation for simultaneous interpretation (Hamon et al., 2009) is necessary.

Table 5: Translation accuracy of each interpretation data in TED

Interpreter	Words (JA)	BLEU	RIBES	WER	TER
S rank	12,968	11.35	59.30	89.13	85.20
A rank	10,818	6.12	48.09	92.45	89.84
B rank	10,700	8.31	48.75	91.71	88.78

Table 6: The number of words in target data of three interpreters

Rank	TED	CSJ	CNN	NHK
S rank	66,307	52,202	6,495	3,075
A rank	69,451	47,941	9,825	3,516
B rank	68,654	45,484	10,807	3,075

4.3. Data Size

The size of the transcript for each rank is shown in Table 6. For CSJ, as the rank increases, the number of words tends to increase. This is because the S rank interpreter can interpret sentences that the A and B rank interpreters cannot. In TED, the number of words for the S rank interpreter is the least, but this is due to the fact that several TED talks have movies played by the speakers in addition to the actual talk. We informed the interpreters that it was acceptable to either interpret the movies or not, and the S rank interpreter did not interpret the movies at all, while the A and B rank interpreters did. For CNN as the rank is lower, the number of words is higher. This is because of the difficulty of the news task, and due to the fact that we gave the interpreters full transcripts in advance. The lower rank interpreter just translated in advance and did not interpret, while the higher rank interpreters chose to browse the material and interpret on the fly.

5. Conclusion

This paper describes the collection of a simultaneous interpretation corpus. Professional simultaneous interpreters with the different amounts of experience cooperated with the collection. In the future, we would like to analyze and quantify the differences between the translation styles of interpreters with different amounts of experience.

6. Acknowledgments

Part of this work was supported by JSPS KAKENHI Grant Number 24240032.

7. References

Bangalore, Srinivas, Sridhar, Vivek Kumar Rangarajan, Golipour, Prakash Kolan Ladan, and Jimenez, Aura. (2012). Real-time incremental speech-to-speech translation of dialogs. In *Proc. NAACL*.

Claudio, Bendazzoli and Annalisa, Sandrelli. (2005). An approach to corpus-based interpreting studies. In *Proc. MuTra*.

Fügen, Christian, Waibel, Alex, and Kolss, Muntsin. (2007). Simultaneous translation of lectures and speeches. *Machine Translation*, 21(4):209–252.

Fujita, Tomoki, Neubig, Graham, Sakti, Sakriani, Toda, Tomoki, and Nakamura, Satoshi. (2013). Simple, lexicalized choice of translation timing for simultaneous speech translation. In *Proc. InterSpeech*.

Hamon, Olivier, Fugen, Christian, Mostefa, Djamel, Arranz, Victoria, Kolss, Muntsin, Waibel, Alex, and Choukri, Khalid. (2009). End-to-end evaluation in simultaneous translation. In *Proc. EACL*.

Koiso, Hanae, Tsuchiya, Naoko, Mabuchi, Yoko, Saito, Miki, Kagomiya, Takayuki, Kikuchi, Hideaki, and Maekawa, Kikuo. (2000). Transcription criteria for the corpus of spontaneous Japanese. *IEICE Technical Report*.

Maekawa, Kikuo. (2003). Corpus of spontaneous Japanese: Its design and evaluation. In *Proc. ISCA/IEEE Workshop on Spontaneous Speech*.

Paulik, Matthias and Waibel, Alex. (2009). Automatic translation from parallel speech: Simultaneous interpretation as MT training data. In *Proc. ASRU*.

Ryu, Koichiro, Mizuno, Atsushi, Matsubara, Shigeki, and Inagaki, Yasuyoshi. (2004). Incremental Japanese spoken language generation in simultaneous machine interpretation. In *Proc. Asian Symposium on Natural Language Processing to Overcome language Barriers*.

Sridhar, Vivek Kumar Rangarajan, Chen, John, and Bangalore, Srinivas. (2013a). Corpus analysis of simultaneous interpretation data for improving real time speech translation. In *Proc. InterSpeech*.

Sridhar, Vivek Kumar Rangarajan, Chen, John, Bangalore, Srinivas, Ljolje, Andrej, and Chengalvarayan, Rathinavelu. (2013b). Segmentation strategies for streaming speech translation. In *Proc. NAACL*.

Toyama, Hitomi, Matsubara, Shigeki, Ryu, Koichiro, Kawaguchi, Nobuo, and Inagaki, Yasuyoshi. (2004). CIAIR simultaneous interpretation corpus. In *Proc. Oriental COCOSDA*.