A Comparative Study on Human Communication Behaviors and Linguistic Characteristics for Speech-to-Speech Translation

Toshiyuki Takezawa and Genichiro Kikui

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
{toshiyuki.takezawa, genichiro.kikui}@atr.jp

Abstract

A large bilingual corpus of English and Japanese is being built at ATR Spoken Language Translation Research Laboratories in order to improve speech translation technology to the level where people can use a portable translation system for traveling abroad, dining and shopping, and hotel situations. As a part of these corpus construction activities, we have been collecting spoken dialogue data by using an experimental translation system between English and Japanese. In a previous study, we found that humans communicate as part of their daily social life, so they prefer using complex sentences and saying than one sentence per utterance. However, corpus-based machine translation systems for conversational expressions tend to be limited to dealing with short simple sentences. To find a way to bridge the gap between human communication behaviors and system performance, we examined the relationship between instructions and linguistic expressions. The experimental results suggest that a state-of-the-art translation system may be useful for subjects who can make their utterance length short by following instructions.

1. Introduction

Because of its great potential for industrial and social applications, speech-to-speech translation (S2ST) technology has received considerable attention. Corpus-based technologies have provided successes in speech processing such as HMM, N-gram, and corpus-based speech synthesis. As for natural language processing such as machine translation, corpus-based technologies are also promising, as seen by the recent launching of several corpus-based S2ST projects such as the European TC-STAR (Technology and corpora for speech-to-speech translation) project (Höge, 2002) and the DARPA Babylon project (Eurospeech, 2003).

A large bilingual corpus of English and Japanese is being built at ATR Spoken Language Translation Research Laboratories in order to improve speech translation technology to the level where people can use a portable translation system for traveling abroad, dining and shopping, and hotel situations. In order to expand the coverage to a wider variety of domains, we have been collecting sentences that bilingual experts consider useful for people going-to/coming-from another country. The resulting English-Japanese aligned corpus is called BTEC (Basic Travel Expression Corpus) (Takezawa et al., 2002), which is now being translated into several other languages (Kikui et al., 2003).

We also require a corpus that correctly reflects the utterances to be spoken to the system for use in evaluating the performance of the entire system and component technologies. The Spoken Language DataBase (SLDB) (Takezawa et al., 2004) provides ideal data because professional human interpreters help communication between people speaking different languages. The Machine-translation-Aided bilingual spoken Dialogue corpus (MAD) (Takezawa and Kikui, 2003) has realistic data for studying communication behaviors and linguistic expressions that are helpful in front of S2ST systems.

According to the previous study (Takezawa and Kikui, 2003), we found that humans communicate as part of their daily social life, so they prefer using complex sentences and saying more than one sentence per utterance. However, the corpus-based machine translation systems at ATR tend to be limited to dealing with short simple sentences because BTEC contains basic travel expressions. In order to investigate how to bridge the gap between human communication behaviors and expressions in BTEC, we examined the relationship between instructions and linguistic expressions. This paper presents an overview and gives discussions on our work.

Section 2 describes the experimental system's construction. Section 3 presents dialogue experiments. Section 4 offers some discussions on the results. Finally, section 5 gives our conclusions.

2. Experimental System Construction

We use human typists to transcribe the users' utterances and input them into a machine translation system between English and Japanese instead of using speech recognition systems because we want to first focus on a particular component technology, i.e. Machine Translation (MT).

Figure 1 shows the experimental system configuration. An English typist transcribes an English utterance and inputs it into a machine translation system operating from English to Japanese. The translated Japanese text and its synthesized speech are sent to a Japanese speaker. Likewise, a Japanese typist transcribes a Japanese utterance and inputs it into a machine translation system operating from Japanese to English. The translated English text and its synthesized speech are sent to an English speaker. By repeating this, an MT-aided bilingual dialogue continues. Speech waves, transcriptions, and translated texts are stored in log files.

A combined system of an extended Transfer Driven Machine Translation (TDMT) system (Sumita et al., 1999) and a DP-matching Driven transDucer (D3) system (Sumita, 2001) was used as the Japanese-to-English translation system. If the value of a distance measure between input and

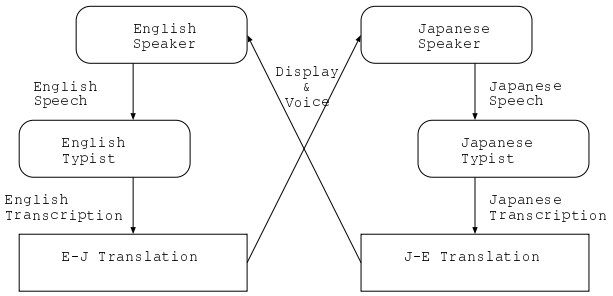

Figure 1: Experimental System Configuration

translation examples was less than 0.2, the result of the D3 system was selected; otherwise, the result of the extended TDMT system was selected. Furthermore, an extended TDMT system was used for an English-to-Japanese translation system. CHATR (Campbell, 1996) was used as a Japanese speech synthesis system. For the English speech synthesis system, AT&T Labs' Natural Voices^{TM} was used.

As a part of this dialogue data collection, we employed the Japanese speech recognition system SPREC (Naito et al., 2001), which is a component of the ATR-MATRIX speech translation system between Japanese and English (Sugaya et al., 1999).

3. Dialogue Data Collection

3.1. Instructions to Subjects

In a previous study (Takezawa and Kikui, 2003), we employed the following instructions to subjects based on the experience of SLDB, which involved human-interpreter-aided bilingual spoken dialogue data collection.

(1) Speak loudly and clearly.

(2) One utterance must be made within ten seconds.

(3) Errors sometimes occur. In such cases, try to make the dialogue continue by confirming or repeating unclear parts.

(4) The system sometimes needs time. Please wait for it.

Items (1) and (2) are based on the experience of SLDB. Items (3) and (4) were added after we employed a machine translation system instead of human interpreters. We assume that less constrained instructions are good for prospective users of S2ST systems, so we employed such instructions. We call this set of instructions "Instruction A."

According to the experience of ATR-MATRIX (Sugaya et al., 1999), both speech recognition and machine translation tended to yield better performance with shorter utterances. Moreover, corpus-based machine translation systems at ATR tend to be able to deal only with short simple sentences because BTEC contains basic travel expressions. Therefore, we tried the following more constrained instructions to consider the limitations of machine translation systems.

(5) Speak briefly and concisely.

(6) Japanese side: Try to use "*watashi-wa* (I)," "*watashi-no* (my)," "*anata-ni* (you)" and so on rather than less standard variants.

MAD4
Period: 12 days from June to July 2003
Japanese speakers: 12 people
English speakers: 12 people
Task settings: 16 patterns
Utterances: 3,666
Dialogues: 166

Table 1: Overview of Dialogue Experiment

	First 6 days	Final 6 days
Morning	AT	BT
Afternoon	BT	CT, CR

Table 2: Schedule of Instructions to Subjects

These items were added to Instruction A. Item (5) considers both speech recognition and machine translation. Item (6) considers machine translation from Japanese to English. We call this enlarged set of instructions "Instruction B."

We also tried the following items to consider speech recognition.

(7) Use a monotone voice.

(8) Speak at a fixed rate.

These items were added to Instruction B, and we call this set of eight instructions "Instruction C."

3.2. Conducting Experiments

Sufficiently skilled typists were able to carry out their work accurately and quickly enough for these experiments. For the user interface, we used headset microphones with headphones and small-sized portable PCs. Table 1 shows an overview of the experiment. Speaker pairs were changed everyday.

During the first 6 days, Instruction A was given in the morning sessions, and Instruction B was given in the afternoon sessions. The tasks of dialogues were balanced by changing the morning parts and afternoon parts everyday. During the latter 6 days, Instruction B was given in the morning sessions, and Instruction C was given in the afternoon sessions. As a part of Instruction C, we employed a Japanese speech recognition system instead of a Japanese typist. In order to distinguish these two conditions, T is used to represent the typist and R is used to represent the speech recognition system. As described above, A, B, and C are used to distinguish the instructions given to subjects. Therefore, the resulting dialogue data were classified into four types: AT, BT, CT, and CR. Table 2 shows these classifications.

4. Discussions

4.1. Basic Characteristics

As basic characteristics, Table 3 shows the average number of words per utterance, Table 4 shows the average number of sentences per utterance, and Table 5 shows the percentages of simple and complex sentences in Japanese.

	BTEC	SLDB	MAD4/AT	MAD4/BT	MAD4/CT	MAD4/CR
Japanese	6.87	13.30	11.13	9.78	9.01	8.02
English	5.87	11.27	12.60	9.56	9.54	8.97

Table 3: Average Number of Words Per Utterance

	BTEC	SLDB	MAD4/AT	MAD4/BT	MAD4/CT	MAD4/CR
Japanese	1.07	1.35	1.35	1.41	1.33	1.23
English	1.08	1.38	2.19	1.78	1.84	1.74

Table 4: Average Number of Sentences Per Utterance

All of these tables include the values of BTEC and SLDB as well as MAD4/AT, MAD4/BT, MAD4/CT and MAD4/CR.

According to these tables, except for the average number of sentences per utterance for English, the basic characteristics of MAD4/AT are similar to those of SLDB, that is, transcriptions of bilingual conversations through human interpreters. This is because Instruction A is based on that of SLDB. The reason why the average number of sentences per utterance for English increased is because some of the English speakers tended to say "okay" or "yes" at the beginning of their utterances when responding.

By comparing MAD4/AT with MAD4/BT, the average number of words per utterance decreased both for Japanese and English. The reduction rates on average are about 88% for Japanese and about 76% for English.

As for the percentage of simple and complex sentences in Japanese, approximately 70% of the sentences for Instruction A were simple sentences, which is almost the same as that of SLDB because both instructions were the same. However, approximately 80% of the sentences for Instruction B were simple sentences, which is almost the same as that of the BTEC corpus.

There were no significant differences between MAD4/BT and MAD4/CT; however, MAD4/CR, in which a Japanese speech recognition system was employed, tended to make the utterance length shorter.

4.2. Analysis by Each Subject

Depending on the instructions, the average number of words per utterance decreased for both Japanese and English. During the first 6 days, all speaker pairs were first given Instruction A in the morning sessions and Instruction B in the afternoon sessions. Speaker pairs were changed everyday, so there were six English speaker participants and six Japanese speaker participants. We call these E1, E2, E3, E4, E5 and E6 for English and J1, J2, J3, J4, J5 and J6 for Japanese. Table 6 shows the average number of words per utterance for each speaker.

According to Table 6, half of the Japanese speakers (J1, J3, and J5) were able to make their utterances shorter, according to the change in the instructions, but the other half (J2, J4, and J6) could not change their utterance length along with the change in instructions, while all of the English speakers were able to make their utterances shorter. The reason may be based on differences in the oral education systems, although there may be some other biases because about half of the English speakers were tutors of English conversation.

Speakers	MAD4/AT	MAD4/BT	Rate
J1	13.0	10.4	80.0%
J2	9.6	9.6	100.0%
J3	12.1	10.2	84.3%
J4	10.8	11.5	106.5%
J5	11.1	8.9	80.2%
J6	10.9	10.6	97.2%
E1	9.0	8.0	88.9%
E2	13.3	10.4	78.2%
E3	11.5	9.6	83.5%
E4	15.8	13.4	84.8%
E5	13.8	8.4	60.9%
E6	11.5	9.7	84.3%

Table 6: Average Number of Words Per Utterance for Each Speaker

Words/Terms	MAD4/AT	MAD4/BT	Rate
Watashi (I)	0.99%	4.58%	462.6%
Watakushi (I)	0.00%	0.00%	—
Kochira (I)	0.79%	1.83%	231.6%
Anata (you)	0.00%	2.14%	∞
Sochira (you)	0.59%	0.15%	25.4%
Okyakusama (you)	0.00%	0.15%	∞

Table 7: Occurrence of Pronouns Per Japanese Sentence

4.3. Occurrence of Pronouns Per Japanese Sentence

With Instruction B, we asked Japanese subjects to try to use "*watashi-wa* (I)," "*watashi-no* (my)," "*anata-ni* (you)" and so on. Table 7 shows the occurrences of these six words/terms per Japanese sentence.

According to Table 7, the occurrence of "*watashi*" increased by about 4.6 times, but "*watashi*" was used originally in only about 1% of the sentences. Even after Instruction B, "*watashi*" was not used in the remaining 95% of the sentences. This suggest that it is too difficult for Japanese speakers to use such words/terms frequently even if explicit instructions are given.

	BTEC	SLDB	MAD4/AT	MAD4/BT	MAD4/CT	MAD4/CR
Simple sentences	82.8%	65.9%	69.5%	79.8%	79.9%	83.5%
Complex sentences	17.2%	34.1%	30.5%	20.2%	20.1%	16.5%

Table 5: Simple and Complex Sentences in Japanese

	MAD4/AT	MAD4/BT
Total	31.4%	35.0%
Simple sentences	43.2%	42.0%

Table 8: Coverage Based on BTEC for Japanese

4.4. Coverage Based on BTEC for Japanese

We expect improvement in speech translation performance if many more expressions in MAD were covered by BTEC. Table 8 shows coverage based on BTEC for Japanese. We assume that the coverage is the rate of similar expressions in the BTEC corpus, since our purpose is corpus-based speech translation, so we define coverage as the rate of expressions, that is, 80% of the words are matched and the length is from 0.8 times to 1.2 times the original. Under this definition, the coverage increased from about 31% to about 35% according to the change in instructions. This is because simple sentences increased according to the change in instructions. However, if it is limited to simple sentences, the coverage is almost the same or there may be a slight decrease with the change in instructions. This may be due to the effect of the instruction items requiring subjects to try to use "*watashi-wa* (I)," "*watashi-no* (my)," "*anata-ni* (you)" and so on.

Currently, even a machine translation system for spoken dialogue can deal with a large translation dictionary, but it tends to be difficult for a state-of-the-art translation system to deal effectively with complex sentences. However, according to the experimental results, a state-of-the-art translation system may be useful for subjects who can make their utterance length short.

5. Conclusions

As part of corpus construction activities for future speech translation research, we have been collecting dialogue data using an experimental translation system between English and Japanese. The purpose of this data collection is to study the communication behaviors and linguistic expressions preferred in advance of developing such systems. We will continue to analyze the collected dialogue data. We have already selected some test sets from the data and are currently preparing performance evaluation tests of basic component technologies, such as speech recognition and machine translation.

In the near future, after conducting an experiment to investigate the effect of the other participant's original voice and display, we plan to collect field data after employing a speech recognition system instead of human typists.

6. Acknowledgments

The authors wish to thank Ms. Yayoi Suzuki, Mr. Atsushi Nishino, Mr. Kouji Takashima, Ms. Tomoko Somekawa, Mr. Gen Itoh and Dr. Mitsunori Mizumachi for their help in conducting the experiment.

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

7. References

Nick Campbell. (1996). CHATR: A high-definition speech resequencing system. In *Proceedings of the ASA/ASJ Joint Meeting*, pages 1223–1228.

Eurospeech. (2003). Special session: Multilingual speech-to-speech translation. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, pages 361–384.

Harald Höge. (2002). Project proposal TC-STAR: Make speech to speech translation real. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 136–141.

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. (2003). Creating corpora for speech-to-speech translation. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, pages 381–384.

Masaki Naito, Hirofumi Yamamoto, Harald Singer, Hideharu Nakajima, Atsushi Nakamura, and Yoshinori Sagisaka. (2001). A continuous speech recognition system for conversational speech. *The IEICE Transactions on Information and Systems, PT. 2 (Japanese Edition)*, J84-D-II(1):31–40.

Fumiaki Sugaya, Toshiyuki Takezawa, Akio Yokoo, and Seiichi Yamamoto. (1999). End-to-end evaluation in ATR-MATRIX: Speech translation system between English and Japanese. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 2431–2434.

Eiichiro Sumita, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kahioka, Kai Ishikawa, and Satoshi Shirai. (1999). Solutions to problems inherent in spoken-language translation: the ATR-MATRIX approach. In *Proceedings of the Machine Translation Summit VII*, pages 229–235.

Eiichiro Sumita. (2001). Example-based machine translation using DP-matching between word sequences. In *Proceedings of the ACL-2001 Workshop on Data-Driven Methods in Machine Translation*, pages 1–8.

Toshiyuki Takezawa and Genichiro Kikui. (2003). Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, pages 2757–2760.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 147–152.

Toshiyuki Takezawa, Genichiro Kikui, Atsushi Nakamura, Yoshinori Sagisaka, and Seiichi Yamamoto. (2004). Spoken language corpora development at ATR. In *Proceedings of the 18th International Congress on Acoustics*.