Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus-Based Speech Translation

Toshiyuki Takezawa and Genichiro Kikui

ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan

{toshiyuki.takezawa, genichiro.kikui}@atr.co.jp

Abstract

A huge bilingual corpus of English and Japanese is being built at ATR Spoken Language Translation Research Laboratories in order to enhance speech translation technology, so that people can use a portable translation system for traveling abroad, dining and shopping, as well as hotel situations. As a part of these corpus construction activities, 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 front of such systems. 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. In this paper, we present an overview of our activities and discussions based on the basic characteristics.

1. Introduction

The current tasks of speech translation are goal-oriented cooperative dialogues such as travel conversations. Such dialogues usually consist of short utterances. A huge corpus covers the variations in the expression of short utterances. Therefore, corpus-based technologies are promising for speech translation.

There are three important points to consider in designing and constructing a corpus for future speech translation research. The first is to have a variety of speech samples, with a wide range of pronunciations, speaking styles, and speakers. The second is to have data for a variety of situations. A "situation" means one of various limited circumstances in which the system's user finds himself/herself, such as airport, hotel, restaurant, shopping, and travel circumstances; it also involves various speakers' roles, such as communication with a middle-aged stranger, a stranger wearing jeans, a waiter/waitress, or a hotel clerk. The third is to have a variety of expressions.

According to our previous study [1], human-to-machine conversational speech data had similar characteristics to human-to-human indirect communication speech data, such as spoken dialogues between Japanese and English speakers through human interpreters. Moreover, the human-to-human indirect communication data had an intermediate characteristic, i.e., it was positioned somewhere between direct communication data, that is, Japanese monolingual conversations, and read speech data from conversational text. If we assume that a speaker would accept a machine-friendly speaking style, we could take a great step forward: a clear separation of speech data collection and bilingual data collection. In the following, we focus on bilingual language data collection and discuss solutions to the second and third points, i.e., varieties of situations and expressions.

In order to cover a variety of situations, ATR Laboratories have collected a Basic Travel Expression Corpus (BTEC) [2]. To cover a variety of expressions, ATR Laboratories have also tried to collect paraphrases of many basic expressions [3]. These two trials succeeded in collecting a large data size, however, the characteristics might be different from the targets with which speech translation systems must deal, because neither of them are transcriptions of spoken utterances.

In order to study the communication behaviors and linguistic expressions preferred in front of speech translation systems, we started to collect Machine-translation-Aided bilingual Dialogues (MAD) using an experimental translation system between English and Japanese. 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 component technology: Machine Translation (MT).

Section 2 describes the necessity of MT-aided dialogues. Section 3 describes an experimental system construction. Section 4 presents dialogue experiments. Section 5 offers discussions. Finally, section 6 gives our conclusion.

2. Necessity of MT-Aided Dialogues

ATR Laboratories have collected a Basic Travel Expression Corpus (BTEC) [2] and a bilingual travel conversation corpus of Spoken Language (SLDB) [4]. BTEC and SLDB are designed to be complementary. BTEC is a collection of Japanese sentences and their English translations, while SLDB is a collection of transcriptions of bilingual spoken dialogues. Whereas BTEC covers a wide variety of travel domains, SLDB covers a limited domain, i.e., hotel situations. The size of BTEC is more than 200,000 sentences, while the size of SLDB is more than 20,000 sentences.

According to our previous study [2], there might be some differences between the target with which speech translation systems must deal and the bilingual corpora built at ATR Laboratories.

BTEC contains edited colloquial travel expressions, which are not transcriptions, so some people might not use similar expressions and its frequency distribution might be different from actual dialogues.

SLDB does not contain any recognition/translation errors because professional human interpreters help communication between people speaking different languages. However, even a state-of-the-art speech translation system cannot avoid recognition/translation errors.

Machine-translation-Aided bilingual Dialogues (MAD) are necessary for a discussion on the quantitative analysis of the difference between the target of speech translation and these corpora. Of course, this kind of dialogue data is also used for

Table 1: Overview of dialogue experiments

	MAD1	MAD2
Purpose	To verify feasibility	To study the relationship between task achievement and the number of utterances
Task	Short dialogues such as asking an unknown foreigner whether there is a taxi stand nearby (8 turns per dialogue on average)	Slightly complicated dialogues, such as planning a local tour and making reservations for it (49 turns per dialogue on average)
Period	12 days from May to June 2002	11 days from August to September 2002
Japanese speakers	Two people per day, total 24 people (19 participants)	One person per day, total 11 people (11 participants)
English speakers	One person per day, total 12 people (12 participants)	One person per day, total 11 people (11 participants)
User interface	Mobile phones and a PDA	Headset microphones with headphones and small-sized portable PCs
Task settings	49 patterns	8 patterns
The number of utterances	Total 3,568 utterances	Total 3,404 utterances
The number of dialogues	Total 445 dialogues	Total 69 dialogues

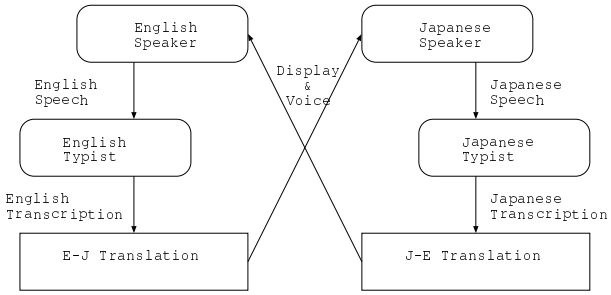

Figure 1: Experimental system configuration

performance evaluation tests of basic component technologies, such as speech recognition and machine translation.

NESPOLE! assumes the use of e-commerce so its corpus contains dialogues through the Internet [5]. We assume a portable translation system for traveling abroad so our corpus contains face-to-face dialogues. Verbmobil [6] also assumed face-to-face dialogues but its corpus did not contain MT-aided bilingual dialogues.

3. Experimental System Construction

Figure 1 shows an experimental system configuration. An English typist transcribes an English utterance and inputs it into a machine translation system 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 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 TDMT (Transfer Driven Machine Translation) system [7] and a D3 (DP-matching Driven transDucer) system [8] was used for a Japanese-to-English translation system. If the value of a distance measure between input and 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. An extended TDMT system was used for an English-to-Japanese translation system. CHATR [9] was used as a Japanese speech synthesis system. As for an English speech synthesis system, CHATR was used for the first series of dialogue experiments, and AT&T Labs' Natural Voices^{TM} was used for the second series of dialogue experiments.

Table 2: Average number of words per utterance

	BTEC	SLDB	MAD1	MAD2
Japanese	6.87	13.30	10.00	12.57
English	5.87	11.27	10.25	11.06

4. Dialogue Experiments

Two series of dialogue experiments were carried out: the first series of dialogue experiments (MAD1); and the second series of dialogue experiments (MAD2). Table 1 shows an overview. Sufficient typists were able to carry out their work accurately and quickly enough for these experiments.

In the first series of dialogue experiments, mobile phones and a PDA were used as user interfaces. A Japanese speaker used a mobile phone and an English speaker used another mobile phone. Both speakers shared a PDA as a display device. However, users found it difficult to understand the state-of-the-art synthesized speech through a mobile phone, so they almost always looked at the display device to confirm the output. Thus, this user interface was not so good for the experiment.

In the second series of dialogue experiments, headset microphones with headphones and small-sized portable PCs were used as user interfaces. The headset microphone quality was more adequate for current dialogue speech recognition research than the mobile phone quality.

After explaining the purpose of the experiment, instructions were given to dialogue participants such as (1) Speak loudly and clearly; (2) One utterance must be done within ten seconds; and (3) Errors sometimes occur. In such cases, try to continue the dialogue by uttering a confirmation or re-speaking. The leader of the experiment sometimes advised the participants during the experiments.

5. Discussions

5.1. Basic Characteristics

As for basic characteristics, Table 2 shows the average number of words per utterance, Table 3 shows the average number of

Table 3: Average number of sentences per utterance

	BTEC	SLDB	MAD1	MAD2
Japanese	1.07	1.35	1.29	1.44
English	1.08	1.38	1.61	1.54

Table 4: Simple and complex sentences in Japanese

	BTEC	SLDB	MAD1	MAD2
Simple sentences	82.8%	65.9%	68.3%	72.0%
Complex sentences	17.2%	34.1%	31.7%	28.0%

sentences per utterance, and Table 4 shows the percentages of simple and complex sentences in Japanese. All of these tables include the values of BTEC and SLDB as well as MAD1 and MAD2.

According to these tables, the basic characteristics of MAD are similar to those of SLDB, that is, transcriptions of bilingual conversations through human interpreters. The basic characteristics of BTEC are different from those of MAD and SLDB. The sentences in BTEC are very short, and more than 80% of BTEC sentences are simple sentences.

5.2. Confirmation Utterances Caused by System Errors

The basic characteristics of MAD are similar to those of SLDB, however, SLDB does not contain any recognition/translation errors. Therefore, we carried out an analysis of confirmation utterances caused by system errors. Utterances in MAD were classified into the following three categories, which are similar to speech act types.

(a) Initiate actions and usual confirmations

(c) Confirmation utterances caused by system errors

(e) Replies, follow-ups, and others

If one utterance had some meaningful units, such as sentences, the utterances were given more than one category such as "ae." There were a few utterances which had no categories because of mis-operation. Table 5 shows the result of this classification of utterances.

The experimental system was a kind of speech translation system which had an extremely accurate speech recognition system. Confirmation utterances caused by system errors occurred approximately 5% of the time with the state-of-the-art translation system for spoken dialogues under such conditions. Confirmation utterances caused by system errors could be classified into two categories: (1) Fixed form of confirmations; (2) Confirmations that contain previous utterances of the other party. Typical examples for those are as follows.

(1) Fixed form of confirmations
"Could you repeat what you said? I couldn't understand."
"Sorry, can you repeat that?"

(2) Confirmations that contain previous utterances of the other party
"Are you saying that I can have the single room for tonight, and then exchange it for the double room tomorrow?"
"Did you say you'd like to find a Chinese restaurant here?"

The fixed form of confirmations tended to be used when the translation output of the previous utterance could not convey any information. Confirmations that contain previous utterances of the other party tended to be used when the translation output of the previous utterance could convey some information but they were grammatically broken.

BTEC contains basic expressions, such as "Could you say that again?", so that BTEC and its paraphrases are expected to cover the fixed form of confirmations.

Table 5: Classification of utterances

	MAD1		MAD2	
	Japanese	English	Japanese	English
a	518 (29%)	440 (24%)	359 (22%)	292 (17%)
c	74 (4%)	48 (3%)	75 (5%)	93 (5%)
e	877 (50%)	878 (49%)	910 (55%)	940 (54%)
ac	0 (0%)	0 (0%)	11 (1%)	1 (0%)
ae	302 (17%)	428 (24%)	265 (16%)	422 (24%)
ce	0 (0%)	2 (0%)	23 (1%)	6 (0%)
ace	0 (0%)	1 (0%)	3 (0%)	1 (0%)
no	0 (0%)	0 (0%)	1 (0%)	2 (0%)
Total	1771 (100%)	1797 (100%)	1647 (100%)	1757 (100%)

5.3. Task Achievement and the Number of Utterances

In the second series of experiments, we studied the relationship between task achievement and the number of utterances, in particular, the influences of confirmation utterances caused by system errors upon task achievements. We selected twelve subtasks, such as deciding on a local tour, for this analysis. Table 6 shows the result of the relationship between task achievement and the number of utterances. Six pairs are selected and shown in the table from eleven pairs of the second series. Pairs 1, 2, and 3 gave joyful impressions to three dialogue observers and pairs 4, 5, and 6 gave boring impressions to them. The number in parentheses indicates the number of utterances caused by system errors. The "—" symbol indicates that the subtask was not carried out by the pair.

According to Table 6, the number of utterances for subtasks tended to depend upon the personalities of the dialogue participants even if the confirmation utterances by system errors might have some influence.

5.4. Rich Communications and Efficient Communications

As shown in Table 6, joyful pairs, such as 2 and 3, tended to speak many utterances. They tended to say not only the required items but also optional related topics. They also tended to use complex sentences and to say more than one sentence per utterance. Some of the other speakers tended to use simple sentences and to say one sentence per utterance. Here, we call these two groups: (1) Rich communications; and (2) Efficient communications. Typical examples are as follows:

(1) Rich communications
"I need to go downtown. Could you tell me where I get on the bus?"
"I'm looking for a spring jacket. Is there anything in blue?"

Table 6: Relationship between task achievement and the number of utterances

Subtask	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
1: Deciding a local tour	9	15	14	9	9	4
2: Asking about transportation	5	7	12	—	7	6 (1)
3: Confirming sightseeing spots	11	38 (10)	27	18	24 (5)	14
4: Asking about payment	5	12	7 (1)	11 (2)	5	5
5: Telling one's name	2	2	3	4	2	—
6: Telling one's address	2	3	9 (4)	4	3	4
7: Asking for non-smoking seats	2	2	4	3	2	8
8: Ordering a drink	3	7 (2)	6	5	4	3
9: Ordering a main dish	9	14 (7)	8	9	4	9
10: Ordering another drink	3	3	3	5	2	4
11: Explaining Japanese cuisine	20 (4)	3	4	3	5	5
12: Altering raw fish if necessary	3	8 (2)	5	4	4	6
Total	74 (4)	114 (21)	102 (5)	75 (2)	71 (5)	68 (1)

(2) Efficient communications
"Where is a taxi stand?"
" I'm looking for a black jacket."

BTEC contains basic travel expressions, such as "Where is a taxi stand?" However, it is too rude to just say "Where is a taxi stand?" to a stranger in either English or Japanese, so such a basic expression is rarely used in human communications even when using a speech translation system. This suggests that good communications may have many utterances and words for achieving a task.

5.5. Future Research Directions

Humans tend to communicate as part of their daily social life, so they prefer to use complex sentences and to say more than one sentence per utterance. There are two directions for future research. One is to collect such preferred expressions from basic travel expressions in BTEC by making paraphrases. The other is to instruct users of speech translation systems to use simple sentences and to say one sentence per utterance. It may be too hard for speakers of rich communications to change their communication style rapidly. After conducting some experiments, such as speaker and language model adaptation, we will examine user instruction methods based on expected improvements of speech recognition accuracy. We will also conduct some analysis of the relationship between translation quality and the communication behaviors of speakers.

6. Conclusion

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 front of 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, we plan to collect field data after introducing a speech recognition system instead of human typists.

7. Acknowledgments

The authors wish to thank Ms. Yayoi Suzuki, Mr. Atsushi Nishino, Mr. Kouji Takashima, Mr. Takanori Matsui and Ms. Tomoko Somekawa 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."

8. References

[1] Takezawa, T., Sugaya, F., Naito, M. and Yamamoto, S., "A comparative study on acoustic and linguistic characteristics using speech from human-to-human and human-to-machine conversations," *Proc. 6th International Conference on Spoken Language Processing*, Vol. III, pp. 522–525, 2000.

[2] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S., "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," *Proc. 3rd International Conference on Language Resources and Evaluation*, Vol. I, pp. 147–152, 2002.

[3] Sugaya, F., Takezawa, T., Kikui, G. and Yamamoto, S., "Proposal for a very-large-corpus acquisition method by cell-formed registration," *Proc. 3rd International Conference on Language Resources and Evaluation*, Vol. I, pp. 326–328, 2002.

[4] Takezawa, T., "Building a bilingual travel conversation database for speech translation research," *Proc. 2nd International Workshop on East-Asian Language Resources and Evaluation — Oriental COCOSDA Workshop '99 —*, pp. 17–20, 1999.

[5] Costantini, E., Burger, S. and Pianesi, F., "NESPOLE!'s multilingual and multimodal corpus," *Proc. 3rd International Conference on Language Resources and Evaluation*, Vol. I, pp. 165–170, 2002.

[6] Wahlster, W., *Verbmobil: Foundations of speech-to-speech translation,* Springer, 2000.

[7] Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K., and Shirai, S., "Solutions to problems inherent in spoken-language translation: the ATR-MATRIX approach," *Proc. Machine Translation Summit*, pp. 229–235, 1999.

[8] Sumita, E., "Example-based machine translation using DP-matching between word sequences," *Proc. ACL-2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 1–8, 2001.

[9] Campbell, N., "CHATR: A high-definition speech re-sequencing system," *Proc. ASA/ASJ Joint Meeting*, pp. 1223–1228, 1996.