Creating Corpora for Speech-to-Speech Translation

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, Seiichi Yamamoto

ATR Spoken Language Translation Research Laboratories, Kyoto, Japan
{genichiro.kikui, eiichiro.sumita, toshiyuki.takezawa, seiichi.yamamoto}@atr.co.jp

Abstract

This paper presents three approaches to creating corpora that we are working on for speech-to-speech translation in the travel conversation task. The first approach is to collect sentences that bilingual travel experts consider useful for people going-to/coming-from another country. The resulting English-Japanese aligned corpora are collectively called the basic travel expression corpus (BTEC), which is now being translated into several other languages. The second approach tries to expand this corpus by generating many "synonymous" expressions for each sentence. Although we can create large corpora by the above two approaches relatively cheaply, they may be different from utterances in actual conversation. Thus, as the third approach, we are collecting dialogue corpora by letting two people talk, each in his/her native language, through a speech-to-speech translation system. To concentrate on translation modules, we have replaced speech recognition modules with human typists. We will report some of the characteristics of these corpora as well.

1. Introduction

Corpora play crucial roles in developing speech-to-speech translation (S2ST) technologies. First, a fair amount of corpora are necessary for training parameters of corpus-based modules, which are now employed in almost every part of S2ST systems. Second, we require a corpus that correctly reflects utterances to be spoken to the system for evaluating the performance of the entire system and component modules, as well as for training statistical parameters. Thus, we should create corpora that maximally cover and correctly sample the set of utterances in the target task.

Since the ultimate goal of S2ST is to simulate human interpreters, it is quite natural to collect data from human-interpreted dialogues between two people who speak different languages. Actually, several researchers collected samples for the corpus by simulated dialogues with human interpreters [1][2]. In our case, Japanese and English native speakers performed simulated dialogues through a professional interpreter and we recorded utterances of these three participants. This corpus, called "SLDB" (Spoken Language Database)[1] has been extensively used for developing/evaluating S2ST technologies but has at least two problems. First, it is too time-consuming and expensive to enlarge the corpus. Next, utterances spoken to human interpreters may differ from those given to machines. A similar problem is found when collecting data for human-computer dialogue systems.

Since we have not found a method that solves the above problems at the same time, we have been trying to tackle them with three approaches. The first two aim mainly at solving the first problem, namely extending the coverage. The third approach relates to the second problem, namely creating a good sample of dialogues uttered to an S2ST system.

In the following sections, we introduce these three approaches and then compare the collected corpora.

2. Basic Travel Expressions Corpus (BTEC)

The Basic Travel Expression Corpus (BTEC) [3] was planned to cover utterances for every potential subject in travel conversations, together with their translations. Since it is almost infeasible to collect them through transcribing actual conversations or simulated dialogues, we decided to use sentences from the memories of bilingual travel experts. We started by investigating "phrasebooks" that contain Japanese/English sentence pairs that those experts consider useful for tourists traveling abroad. We collected these sentence pairs and rewrote them to make translations as context-independent as possible and to comply with our transcription style[1]. Sentences out of the travel domain or containing very special meaning were removed. The overall statistics of the first collection, called BTEC1, are shown in Table 1.

Table 1: Overall statistics of BTEC1.

	Japanese	English
# of utterances	200,241	200,241
# of word tokens	1,689,442	1,230,650
# of word types	21,329	17,076
# of words per sentence (average)	7.67	5.51

In addition, we categorized them into 20 topical classes such as accommodation, restaurant, airport, etc. as shown in Figure 1.

Basic, Shopping, Transportation, Staying Sightseeing, Restaurant, Airport, Business, Contact, Airplane, Home-stay, Study overseas, Drink, Exchange, Snack, Beauty, Going home

Figure 1: Topical classes.

BTEC sentences, as described above, did not come from speech conversation but were generated by experts as reference materials. This approach enabled us to efficiently create a broad coverage corpus, but it may have two problems. First, this corpus may lack utterances that appear in real conversation. For example, when people ask the way to a bus stop, they often use a sentence like (1). However, BTEC1 contains (2) instead of (1).

I'd like to go downtown. Where can I catch a bus? (1)
Where is a bus stop (to go downtown)? (2)

In order to cover these expressions, we are collecting the paraphrased corpora introduced in Section 3.

The second problem is that the frequency distribution of this corpus may be different from the "actual" one. In this

[1] For example, numbers were spelled out.

corpus, the frequency of an utterance (indirectly) corresponds to how many travel experts come up with this sentence and in how many situations they think the sentence will appear. Therefore, it is possible to think of this frequency distribution as a first approximation of reality, but this should be validated. Up to now, we need to record actual conversations to closely approximate the distribution as introduced in Section 4.

Despite these problems, BTEC serves as the primary source for developing broad-coverage speech-to-speech translation. Actually, BTEC1 is now being translated into several languages including Chinese, French, German, Italian, and Korean by members of C-STAR (International Consortium for Speech Translation Advanced Research)[1]. Using the resulting multi-lingual parallel corpora of travel expressions, we are working on a comparative study of corpus-based translation methods applied to various pairs of languages [4].

3. Paraphrased Corpus

Language is redundant, in the sense that different expressions have almost the same information (in a particular context). A paraphrased corpus tries to cover as many of these semantically equivalent expressions as possible. We asked bilingual subjects to write all of the possible translations for a given sentence, called *a seed sentence*, instead of asking them to produce monolingual paraphrases [5]. Although this approach requires bilingual subjects, we can replace difficult "semantic equivalence definitions" with the somewhat easier "translation equivalences". Since a seed sentence may have different interpretations, we presented one representative translation to restrict the meaning[2].

To systematically list up possible paraphrases, we introduced what we call a *cell-form,* a subset of the regular expression, shown in Figure 2. In this form, a sentence is represented with a horizontal sequence of cells, where each cell contains alternative phrases in that context. For example, the second row in Figure 2 represents four sentences with four different verbs (in the third cell from the left). Frequently used phrase alternatives (e.g., "May I ask", "Can I ask") are represented with a special symbol like Suffix-C.

Tea with lemon, please. (Eng)			
Lemon ti (lemon tea)	wo (case-marker)	itadaki (take) nomi (drink) tanomi (order) morai (take)	masu (polite-aux)
Lemon ti (lemon tea)	wo (case-marker) de (case-marker)		Suffix-C (request)

Figure 2: Paraphrase in the cell-form.

We collected paraphrased Japanese sentences for about 8,500 English seed sentences randomly selected from BTEC1. We presented English sentences and their translations to Japanese natives with good proficiency in English and asked them to list possible Japanese translations for each seed sentence (pair). Abrupt expressions, which are not used in normal travel situations, and sentences that differ only in phrase order[3] were not included.

We obtained 4.61 cell sequences on average for a seed sentence. Since paraphrases that were made by simple lexical or phrasal substitution are packed into the same cell-sequence, we can say more than 4 major paraphrases are obtained from each seed sentence. This seems small at a glance, but each cell contains 1.62 phrases on average, so the total number of paraphrased sentences becomes huge (more than 100 paraphrases per one seed sentence, depending on the length of the seed sentence).

Considering time and volume of data, this method has roughly a three-digit level of efficiency compared with the existing scheme by transcribing simulated dialogues. The collected sentences contain many similar expressions, but they will all be natural sentences.

4. MT Aided Dialogue Corpus (MAD)

The two approaches we have introduced so far focus on maximizing the coverage of the corpus rather than creating a quantitatively correct sample of reality. In this sense, the resulting corpus can be regarded as a *dictionary of utterances*.

The third approach, introduced in this section, is intended to collect representative utterances that people will input to S2ST systems.

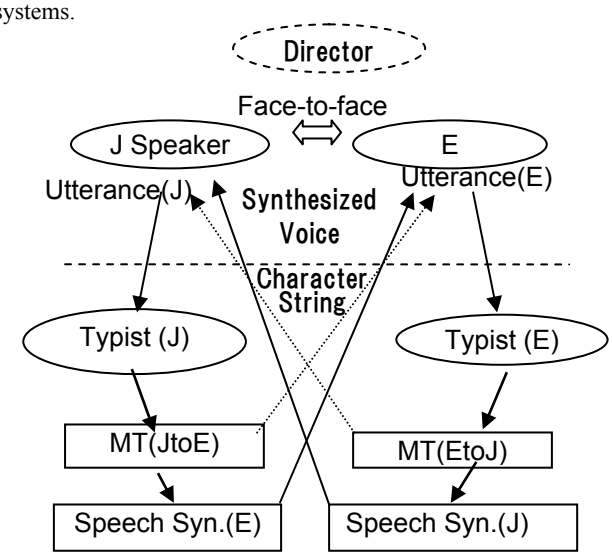

Figure 3: Data collection environment for MAD.

For this purpose, we carried out simulated (i.e., role play) dialogues between two native speakers of different mother tongues with Japanese/English bi-directional S2ST system, instead of using human interpreters. In order to concentrate on effects of MT by circumventing communication problems caused by speech recognition errors, we replaced the speech recognition modules with human typists. The resulting system is, thus, considered equivalent to be using an S2ST system whose speech recognition part is almost perfect. An overview of the data collection environment is shown in Figure 3. We employed a combined version of TDMT [6] and D3 [7] for the MT module (J-to-E and E-to-J) and CHATR [8] for the speech synthesizer.

[1] Main website http://www.c-star.org/, corpus website http://cstar.atr.co.jp/cstar-corpus/.

[2] First, we presented only the source (i.e., English) sentences to reduce the influence of translation samples. This, however, allowed paraphrasers too free a range of interpretation to keep paraphrases consistent.

[3] We are planning to generate them automatically.

This environment is somewhere between the "Wizard-of-Oz" (WOZ) approach in Verbmobil [2], which replaced the entire S2ST process with human, and an approach that relies only on an S2ST system [9][1].

We have carried out three sets of simulated dialogues so far. The first set (MAD1) is to see whether this approach is feasible with rather simple tasks such as "asking an unknown foreigner where a bus stop is". The second set (MAD2) focused on task achievement with slightly complicated tasks, such as planning a guided tour with travel agents. The third set contains carefully recorded speech data (MAD3). Table 2 shows an overview of MAD1 and MAD2.

Table 2: Overview of simulated dialogues.

	MAD1	MAD2
Task settings	49 patterns	8 patterns
# of dialogues	445	69
# of utterances	3,568	3,404
Average # of turns (per dalogue)	8	49
English speakers	12 people	11 people
Japanese speakers	24 people	11 people

4.1. Effect of MT in a dialogue

Results of subjective evaluation show that 61% of translated utterances are judged to be correct, 28% are partially correct, and the remaining 11% are non sense or no-output. We randomly picked up 300 translations from those judged to be partially correct and non sense in MAD1 and classified them by the following discourse flow as shown in Table 3.

Table 3: Classification of discourse flow after errors.

Type	Freq.
Requesting repetition. e.g.) Sorry, can you repeat that?	51 (32)+
Asking back or rephrasing the previous question. e.g.) Did you say you'd like to find a Chinese restaurant here?"	60 (30)+
No observable actions.	135
Director's intervention.	37
Stopping the dialogue	17

(+:the number of recovered translation errors.)

About 20% ((32+30)/300) of errors were recovered in the discourse by the dialogue participants. A typical recovery process begins with a confirmation utterance including a repetition request. In response to the confirmation utterance, the other speaker rephrases the original utterance, which was difficult for the MT system, into a simpler expression (e.g., by separating a complex sentence or coordination into a juxtaposition of simpler sentences, substituting "frequent" expressions for "sophisticated" ones, etc.)[2]. Furthermore, 135 errors did not have any effect on the dialogue, which means that the listener could guess the content of original utterance from partially correct translation and/or from contexts, or simply ignore the utterance. Further investigation of the effect of MT on task achievement is described elsewhere [11].

[1] Another difference is that NESPOLE! dialogues were taken place over the Internet.

[2] Our subjects could easily find that the system seldom makes speech recognition errors. However, when speech recognizers are included in future, they may have problems in "identifying where the problem lay with the system". [10]

5. Comparison of Corpora

This section describes some statistical data on the relationships among the different corpora introduced above. Since we have not finished pre-processing, e.g., tokenizing and tagging, the paraphrased corpora, we concentrate on SLDB (dialogues with human interpreter), BTEC (travel expressions), and MAD (Dialogues with S2ST system and typists).

5.1. Sentence length

Table 4: Basic statistics of each corpus.

	MAD1	MAD2	BTEC1	SLDB
# of utterances	3568	3404	200k(pairs)	32k
# of dialogues	445	69	—	618
Ave. # of words per utterance (J/E)	10.0/ 10.3	12.6/ 11.1	6.9/ 5.9	13.3/ 11.3
Simple Sentences (J)	68.3%	72.0%	82.8%	65.9%

The average length of utterances for each corpus is given in Table 4. This table shows that the MAD corpus is similar to SLDB rather than to BTEC in terms of utterance length. This is quite natural because MAD and SLDB are both from simulated dialogues that contain complex or compound sentences as well as many modifiers (e.g., adverbs and adjectives) to refer to actual situations.

5.2. N-gram coverage

Since we assume MAD corpora are quite expensive but most similar to the target conversation, it is important to know how much other corpora cover MAD. Figure 4 shows how much SLDB and BTEC1 cover MAD N-grams. At a glance, BTEC tri-grams cover 63.1% of MAD tri-gram tokens. Although we should consider that BTEC contains 10 times the number of sentences of SLDB, BTEC well covers local word sequences in MAD corpus.

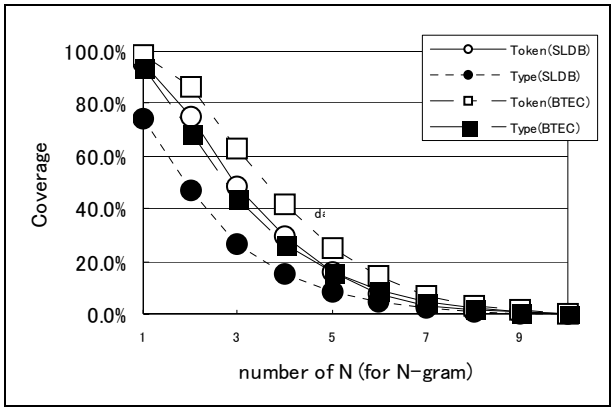

Figure 4: N-gram Coverage (Japanese).

5.3. Sentence level coverage

To investigate the coverage at "sentence construction level", we counted how many sentences of MAD are covered with those of BTEC and SLDB focusing on either content words or functional words. To put it concrete, we converted every sentence into the sequence of content words (or functional words) by removing all other words before counting overlaps. Interestingly, the result shows that SLDB covers MAD better than BTEC as shown in Table 565.

Table 5:Coverage of content/functional word sequence (Japanese).

	BTEC (token/type)	SLDB (token/type)
Content words	21.0%/18.6%	28.2%/20.2%
Functional words	46.9%/44.1%	52.6%/50.8%

5.4. Cross-perplexity

From the above two data results on coverage, we can hypothesize that BTEC and SLDB cover MAD complementarily. This hypothesis is partly validated by the next set of data on perplexity shown in Table 676. In this table, BTEC1+SLDB combines two language models trained on BTEC1 and SLDB with linear interpolation. Similarly, BTEC1+E combines BTEC1 and a corpus E, a sample of BTEC-type extra corpus whose size is the same as SLDB. This clearly shows that BTEC1 and SLDB are both required for handling MAD-type tasks.

Table 6: Cross-perplexities for MAD (Japanese).

	Training Corpus			
	BTEC1	SLDB	BTEC1+ SLDB	BTEC1+E
Size (# of utt.)	162k	12k	174k	174k
Cross-Perplexity	38.2	94.9	30.7	35.7

6. Concluding Remarks

In this paper, we introduced three approaches we are working on to create corpora for developing speech-to-speech translation. They are 1) collecting sentences from reference books compiled by bilingual experts (BTEC), 2) expanding corpus by paraphrasing, and 3) dialogues with an S2ST system (MAD). The first two were intended to cover wider subjects and expressions, while the last approach focused on collecting actual utterances. We did some quantitative comparison and found out that BTEC and SLDB (previous corpus collected through human interpreted dialogues) complement each other in handling broad-coverage real utterances. Based on these results, we are investigating corpus-based translation technology for real use, including objective evaluation of its quality [12].

Future directions include expanding the size of these corpora, trying simulated dialogues with full S2ST systems, and extending them to other domains. Another challenging topic is to investigate how to create a broad coverage corpus with characteristics of real utterances.

7. Acknowledgements

We express our thanks to Fumiaki Sugaya and Yumiko Kinjo for developing paraphrased corpora. This research was supported in part by a contract with the Telecommunications Advancement Organization of Japan (TAO).

8. References

[1] Morimoto, T., Uratani, T., Takezawa, T., Furuse, O., Sobashima, Y., ida, H., Nakamura, A., Sagisaka, Y., Higuchi, N. and Yamazaki, Y.: "Speech and language database for speech translation research", Proc. ICSLP-94, pp. 1791-1794 (1994).

[2] Jekat, S. J. and v. Hahn, W.: "Multilingual Verbmobil-Dialogs: Experiments, Data Collection and Data Analysis", in Verbmobil: Foundations of Speech-to-Speech Translation, Wahlster (Eds), Springer, pp.576-582 (2000).

[3] 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. LREC-2002, Vol. I, pp.147-152 (2002).

[4] Federico, M.,: "Evaluation Frameworks for Speech Translation Technologies", Proc. Eurospeech 2003, this volume.

[5] Sugaya, F., Takezawa, T., Kikui, G. and Yamamoto, S.: "Proposal for a very-large-corpus acquisition method by cell-formed registration", Proc. LREC-2002, Vol. I, pp. 326-328 (2002).

[6] 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).

[7] 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).

[8] Campbell, N.,: "CHATR: A high-definitionspeech re-seqencing system," Proc. ASA/ASJ Joint Meeting, pp. 1223-1228 (1996).

[9] Costantini, E., Burger S., and Pianesi, F.: NESPOLE!'s Multilingual and Multimodal Corpus, Proc. LREC-2002, pp. 165-170 (2002).

[10] Frederking, R. E., Black, A. W., Brown, R. D., Moody, J., and Steinbrecher, E.:"Field Testing the Tongues Speech-to-Speech Machine Translation System", Proc. LREC-2002, pp. 160-164 (2002).

[11] Takezawa, T., and Kikui, G.,: "Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus-Based Speech Translation, submitted to Eurospeech 2003.

[12] Sumita, E.,: "Corpus-Centered Computation", Proc. ACL-2002 Workshop on Speech-to-Speech Translation, pp. 1-8, (2002).