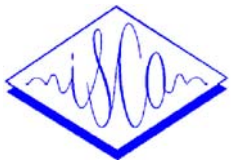

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Multilingual Corpora for Speech-to-Speech Translation Research

Genichiro KIKUI, Toshiyuki TAKEZAWA, Seiichi YAMAMOTO

ATR Spoken Language Translation Research Labs.
genichiro.kikui, toshiyuki.takezawa, seiichi.yamamoto@atr.jp

Abstract

Multilingual spoken language corpora are indispensable for developing new speech-to-speech machine translation (S2SMT) technologies. This paper first discusses characteristics that corpora for S2SMT should have, then surveys existing corpora. Finally, it compares these corpora.

1. Introduction

Multilingual spoken language corpora play important roles in developing speech-to-speech machine translation (S2SMT) technologies. Considering that data-driven or corpus-based empirical approaches have become the trend, corpora are necessary for designing, training and evaluating entire systems as well as their component modules.

S2SMT consists of three components: speech recognition (SR), machine translation (MT), and text-to-speech (TTS). Each component has long constituted by itself an independent research topic for which many corpora are available. These corpora are definitely useful in S2SMT research as well. But, do we need nothing else?

In the dawn of S2SMT research, people needed to focus mainly on making component technologies more robust by using corpora for each component. Thus, corpora that were specialized for collective S2SMT use were almost outside of their scope or, at least, were considered to be resources for the future.

After the mid 90s, however, people gave more attention to issues inherent to overall S2SMT, such as evaluating the entire system (e.g., [1]), developing S2SMT models that integrate SR and MT (e.g., [2], [3]), usability issues (e.g., [12]), etc. To address these issues, we need data obtained from the S2SMT task in question. For example, if we are going to train a combined model of SR and MT, we need a corpus of input speech annotated with the resulting translations as well as transcriptions. In line with this trend, researchers are beginning to spend considerable effort to create corpora specialized for S2SMT research under domestic/international research funds.

In the following sections, we first discuss the properties that corpora for S2SMT research should have, then survey how corpora are created and used both inside and outside our research institute. Note that we focus on SR and MT and ignore TTS in this paper since TTS is currently independent of S2SMT.

2. What Is Required for S2SMT Corpora?

2.1. Within-Task Speech Corpora

First of all, we need speech data that will be spoken to the target S2SMT system. This data is indispensable for evaluating the entire system.

To make discussions clearer, we classify S2SMT tasks into two categories: *dialog translation (DT)* and *lecture (or monolog) translation (LT)*. The former aims at mediating communication between people who speak different languages, while the latter translates a lecturer's speech into the language that is understood by the audience, simultaneously or after finishing the lecture.

"Real utterances" in DT need to be collected by recording dialogs between two persons speaking different languages through an S2SMT system. Ideally, the S2SMT system has characteristics that are similar to the one that we are going to develop. Obviously, this is a chicken-and-egg situation. To get around this problem, researchers employed a professional interpreter as "an ideal S2SMT system" visibly (e.g., SLDB in 3.3) or invisibly (i.e., Wizard-of-OZ)[6]. Recent progress in S2SMT technologies allows us to use an actual S2SMT system for this data collection as introduced later.

When we collect data by letting people use an S2SMT system, we should be careful about the instructions that we give them. Our experimental results suggest that instructions affect the accoustic and linguistic characteristics of utterances[11], which may result in altering system performance. This further affects the frequency of particular utterances, for example, clarification utterances.

As compared with DT, it may be easier to collect data for LT [1]. We can collect lecture speech by simply recording lectures independent of a particular S2SMT system[2].

Recorded speech data should have multiple reference translations for each sentence/utterance in order to ap-

[1] This does not mean that *developing* LT is easier than DT. Actually, DT may become easier if we ask users to speak in a machine-friendly way.

[2] Another advantage of collecting LT data is that lectures are, in many cases, directed to the public. They are even recorded or videotaped. These help us to avoid privacy and copyright problems.

ply automatic evaluation measures (e.g., BLEU, NIST, mWER).

2.2. Parallel Corpora of Spoken Language

If we could extend a within-task corpus to be large enough, we could use it for training as well as for evaluation. However, since this hardly ever happens due to cost, researchers employ this kind of corpus only for training integration parameters[3] or for *adaptation*.

Consequently, component modules are trained on separate large corpora that sufficiently cover the target tasks. Considering that speech and language corpora for speech recognition are already available, we particularly need multilingual parallel corpora of spoken language for improving MT modules.

If we want to use these corpora for evaluating MT, each sentence/utterance should have multiple reference translations, as described above.

3. Corpora for S2SMT

This section surveys two kinds of corpora required for S2SMT as discussed in the previous section:

1. Within-task Speech Corpora, and
2. Parallel Corpora of Spoken Language.

3.1. Within-Task Speech Corpora

Several research institutes have collected corpora by letting people speak using an S2SMT system or a professional interpreter. These corpora are summarized in Table 1.

3.1.1. Verbmobil

The Verbmobil project[4] researchers collected bilingual dialogs mediated by human interpreters and their S2SMT system[6]. They tried two settings when using human interpreters: interpreters present in front of the dialog participants, and interpreters sitting behind computers (Wizard-of-Oz).

3.1.2. NESPOLE!

NESPOLE![7] researchers conducted dialog experiments using their S2SMT system[12]. In this experiment, an English- or German-speaking tourist located in an arbitrary location conversed with an agent in an Italian travel agency through the Internet mediated by their S2SMT system.

3.1.3. ATR/SLDB

SLDB (Spoken Language Data Base)[9] is a corpus of bilingual simulated dialogs between Japanese and English speakers mediated by professional interpreters. The task domain was restricted to hotel reservation conversations. Each speaker was given the following instructions to limit the complexity;

(1) Speak loudly and clearly.
(2) Each utterance must be made within ten seconds.

The database contains original speech, transcriptions in the source language, and reference translations, to allow for end-to-end evaluation[1] (and adaptation), Reference translations were prepared for a test set consisting of 330 sentences.

3.1.4. ATR/MAD

MAD (Machine-translation Aided Dialogues)[10],[11] is a database of bilingual dialogs mediated by our S2SMT system. In order to focus on the MT part, we replaced the SR module with a human typist, which means that the S2SMT has perfect SR.

MAD differs from SLDB in the following two points: 1) MAD does not use human interpreters but our S2SMT system instead, and 2) MAD includes dialogs in various tourism domains (e.g., meal ordering at a restaurant, and lost-baggage trouble at an airport), while SLDB concentrates on the hotel reservation task.

In addition to the instructions used for SLDB, we added the following statements:

(3) In case an error occurs, try to continue the dialog by confirming (to the other speaker) or repeating unclear parts.
(4) The system sometimes needs time. Please wait for it.

After conducting three experimental collections (MAD1-3), we added the following instructions to see how they affect utterances.

(5) Speak briefly, stating one idea or event in one utterance.
(6) Japanese side: Try to fill zero-pronouns.
(7) Use a monotone voice.
(8) Speak at a fixed rate.

Table 2 shows conditions for each (sub)set of the data. Besides these datasets, we collected data, called MAD4/CR, by using our full S2SMT system (i.e., without using human typists). In this case, we gave all the above instructions. The characteristics of these subcorpora are discussed in Section 4.

Table 2: *Dialog conditions.*

instructions	(1)-(4)	(1)-(6)	(1)-(8)
MAD4 subset name	AT	BT	CT

Table 1: *Within-task speech corpora.*

name	languages	interpreter	domain	size	# of sp.
Verbmobil	de ↔ en /de ↔ ja	Human and MT	scheduling	9K turns	-
NESPOLE!	de ↔ it/en ↔ it	MT	tourist inquiry	7.2K/6.7K tokens	28/28
ATR/SLDB	ja ↔ en	Human	hotel	16K utterances	71
ATR/MAD1-4	ja ↔ en	MT	travel	11K utterances	45

3.2. Multilingual Parallel Corpora

There are few reports that include concrete figures of parallel corpora of spoken language. Table 3 shows two examples.

Table 3: *Parallel corpora of spoken language.*

corpus	languages	domain	size
Verbmobil	de/en	scheduling	58K turns
ATR/BTEC (set1-3)	ja/en/ (it,zh,kr)	travel	417K utterances

3.2.1. Verbmobil

Many spontaneous monolingual dialogs in German, English and Japanese were collected in the Verbmobill project. Out of these dialogs, 58K "turns" of German and English transcriptions were translated into English and German, respectively, by professional translators. These data were used for training statistical translation models[5].

3.2.2. ATR/BTEC

BTEC (Basic Travel Expression Corpus)[10] was planned to extend linguistic coverage to various domains. Although colloquial expressions can be collected by simulated dialogs to a certain extent, an enormous cost is required to cover various domains. So, we decided to ask bilingual travel experts to "write down" sentences/expressions that they expect to be used in various travel situations (including visiting abroad and entertaining foreign guests/customers). They are now being translated into many languages, including Chinese, French, Italian, and Korean, by C-STAR partners[3].

BTEC expressions are not the same as actual utterances but they cover local word sequences fairly well as discussed later.

4. Discussions

This section compares the above corpora collected by ATR.

[3]http://www.c-star.org/ or http://cstar.atr.jp/cstar-corpus/

4.1. Utterance Length

Shorter utterances usually attain better results in SR/MT. The same thing holds for simple sentences [4]. Table 4 shows the average utterance length and ratio of simple sentences for each MAD sub-corpus, SLDB and BTEC. When no extra instructions are given, MAD utterances are much longer than BTEC utterances and close to SLDB utterances. This tendency is more significant when dialog participants try to achieve complex tasks as in MAD2. When extra instructions are given to "simplify" utterances, the average length becomes shorter. Moreover, when a speech recognizer is included, the length becomes much shorter. These results are intuitively acceptable.

4.2. N-Gram Coverage

It is important to know how much a within-task corpus is covered by existing or less expensive corpora.

Table 4: Basic Statisitcs

corpus	#-of-words /utterance	simple sent. /utterance
SLDB	13.3	65.9 %
MAD1	10.0	68.3 %
MAD2	12.57	72.0 %
MAD4/AT	11.1	69.5 %
MAD4/BT	9.8	79.8 %
MAD4/CT	9.0	79.9 %
MAD4/CR	8.0	83.5 %
BTEC1	5.9	82.8 %

In a previous paper[10], we showed that BTEC1 covers about 60% of the trigrams of MAD1, counted by tokens. We calculated this score for variations of MAD subsets as shown in Table 5. We find that MAD corpora are roughly placed between SLDB and BTEC. When instructions become tighter, utterances tend more to be covered by both BTEC and SLDB. One possible explanation for this is that as the restrictions become tighter, the people choose more commonly used expressions. But, this needs further investigation.

[4]A simple sentence is defined to be a sentence that contains only one finite verb.

Table 5: Trigram Coverage

corpus	SLDB(%)	BTEC(%)	# of Words
SLDB*	72.0	55.1	206K
MAD4/AT	46.8	53.2	4.6K
MAD4/BT	48,4	57.4	4.5K
MAD4/CT	51.3	58.2	4.0K
MAD4/CR	52.2	57.7	0.9K
BTEC1*	37.1	72.1	1182K

SLDB*/BTEC*=Separate testset of SLDB/BTEC

4.3. Instructions

From the above discussion, we find that tighter instructions make utterances shorter, and simpler. At the same time, extra instructions make utterances closer to common corpora.

In other words, the tighter the instructions becomes, the easier the S2SMT task gets. This allows us to control the difficulty of dialog S2SMT tasks in order to solve S2SMT problems step-by-step.

5. Conclusion

This paper discussed essential properties of S2SMT corpora, then surveyed existing corpora. Collecting colloquial expressions without conducting/recording dialogs, as in BTEC, is a promissing approach to develop a broad-coverage corpus. By analyzing existing corpora, we confirmed that we can make the utterances input to S2SMT closer to those of BTEC by the instructions given to users. This will help us to create corpora for a "bootstrap" style step-by-step development of S2SMT.

6. Acknowledgements

The authors would like to thank Prof. W. Wahlster (DFKI) for answering our questions on Verbmobil corpora. The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology.

7. References

[1] Sugaya, F., Takezawa, T., Yokoo, A., Sagisaka, Y., and Yamamoto, S., "Evaluation of the ATR-MATRIX Speech Translation System with a Pair Comparison Method between the System and Humans", ICSLP 2000, pp. 1105-1108, 2000.

[2] Ney, H., "Speech translation: Coupling of recognition and translation", ICASSP 1999, pp. 517–520, 1999.

[3] Zhang, R., Kikui, G., Yamamoto, H., Soong, F., Lo, W-K., Watanabe, T., and Sumita, E., "Improved Spoken Language Translation Using N-best Speech Recognition Hypotheses", submitted to ICSLP 2004,

[4] Wahlster, W., (ed), "Verbmobil: Foundations of Speech-to-Speech Translation", Springer, 2000.

[5] Vogel, S., Och, F.J., Tillmann, C., Niessen, S., Sawaf, H., and Ney, H., "Statistical Methods for Machine Translation", In "Wahlster(ed), Verbmobil: Foundations of Speech-to-Speech Translation", Springer, 2000.

[6] Jekat S. J., and Hahn W.v., "Multilingual Verbmobil-Dialogs: Experiments, Data Collection", In "Wahlster(ed), Verbmobil: Foundations of Speech-to-Speech Translation", Springer, pp. 575–582, 2000.

[7] Lazzari, G., "Spoken translation challenges and opportunities", Proc. ICSLP 2000, 2000.

[8] Nakamura, A., Matsunaga, S., Shimizu, T., Tonomura, M., and Sagisaka, Y., "Japanese speech databases for robust speech recognition", Proc. ICSLP 1996, pp. 2199–2202, 1996.

[9] Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N., and Yamazaki, Y., "A speech and language database for speech translation research", Proc. ICSLP 1994, pp. 1791–1794, 1994.

[10] Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S., "Creating Corpora for Speech-to-speech Translation", Proc. Eurospeech 2003, pp. 381–384, 2003.

[11] Takezawa, T., and Kikui, G., "A Comparative Study on Human Communication Behaviors and Linguistic Characteristics for Speech-to-Speech Translation", Proc. LREC 2004, to appear, 2004.

[12] Constantini, E., Burger, S., and Pianesi, F., "NESPOLE!'s Multilingual and Multimodal Corpus", Proc. LREC 2002, pp. 167–170, 2002.