Proposal of a very-large-corpus acquisition method by cell-formed registration

Fumiaki Sugaya, Toshiyuki Takezawa, Genichiro Kikui and Seiichi Yamamoto

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto, Japan
{fumiaki.sugaya, toshiyuki.takezawa, genichiro.kikui, seiichi.yamamoto}@atr.co.jp

Abstract

One promising way to improve the performance of a speech translation system is to collect a large volume of data in the target tasks/domains. However, a naïve expansion of the traditional data collection scheme consumes valuable resources. Advanced speech recognition technology can provide a highly accurate recognizer if a machine-friendly speech is permitted. We propose a new data collection scheme that is supported by this speaking style. The preliminary results of data collection show that the proposed scheme has a three-digit efficiency.

1. Introduction

One promising way to improve the performance of a corpus-based speech translation system is to increase the volume of data used in the training set of each module, including those for speech recognition and language translation. However, conventional enlargement schemes using the current methods demand a linear expansion of resources, including person-hours and money. In our previous experience of data collection, we were forced to collect large amounts of data under various conditions to be observed in real-world situations. When we consider all of the various factors in data collection simultaneously, expansion of combinational conditions becomes unavoidable. Separation of combined factors is indispensable to reducing the expansion of conditions in the data collection.

At ATR, we have developed the ATRMATRIX speech translation system (Takezawa et al., 1998b). Through an end-to-end evaluation using this system, we showed that speaking styles become machine-friendly as speakers become familiar with the system (Sugaya et al., 1999). Although we certainly support the idea that the ability to handle a human-friendly speaking style would be ideal, state-of-the-art speech recognizers cannot proceed such a spontaneous style without serious degradation of system performance. However, the views gathered by a questionnaire given in the overall test revealed that the present system with a machine-friendly speaking style is effective and helpful. If we assume that a speaker would accept a machine-friendly speaking style, we could proceed by a great step. Such a speaking style leads to a positive separation of speech data collection and language data collection. This separation can reduce the combinational expansion in data collection. In the following, we focus on language data collection to improve the performance of the speech translation system.

The bottleneck in the present speech translation system comes from the system's coverage being limited to specific tasks/domains. To solve this bottleneck, we must collect a large language corpus covering many target tasks/domains. However, the traditional linear expansion of a corpus is not productive. We propose an alternative method using translation paraphrasing. In this method, we present a paraphraser with a translation seed sentence and ask him/her to output natural paraphrased sentences having the same meaning as the translation seed sentence. We studied two paraphrasing collection schemes. One is to ask a paraphraser to output complete paraphrased sentences. In the evaluation of translation from Japanese to English, the paraphrased sentences were diversified in the test condition: each of five native speakers of English produced three different English sentences from the seed Japanese sentence. This scheme was shown to be effective in our automatic evaluation scheme (Sugaya et al., 2001), but its linear expansion is not productive when we collect a large amount of data. Instead of increasing the number of English speakers and the allotment of three sentences each, we propose another scheme. A cell-formed registration scheme lets a paraphraser output as many sentences as he/she can. To efficiently proceed with the paraphrasing and to make it easy to check the results, we introduced a system for cell-formed registration. A paraphraser types the output and registers new sentences in a particular form. In this scheme, the high cost of the transcription process can be avoided. Since sentences typically have common parts, a paraphraser can efficiently focus on and specifically type new words, phrases and sentences into the form with an easy-to-use interface.

Section 2 briefly explains the ATR data collection as the typical traditional data collection as a conventional approach to data collection. Section 3 shows speech our recognizer's performance. These data lead us to focus on language data collection. The cell-formed registration method is explained in Section 4. Section 5 explains the test results. Section 6 presents our conclusions.

Conventional data collection at ATR

2. Conventional data collection at ATR

We would like to introduce the SLDB database collection (Morimoto et al., 1994; Takezawa et al., 1998a; Takezawa, 1999) at ATR which uses the conventional scheme. We believe that this DB is one of the largest databases collected specifically in the travel arrangement task/domain, and it contains both speech data and language data. The ATR-MATRIX system developed at ATR is based on this DB. The data size is shown in Table 1. The sentence size of this corpus is 16,725. Since this corpus took several years to collect including the DB cleaning process, it is clear that the collection of large volumes of data demands a more efficient scheme.

Table 1 Data collection at ATR

Language direction	Japanese-to-English	Monolingual in Japanese
Number of dialogues	618	892
Number of speakers	71	499
Number of utterances	16,107	22,874
Number of words	301,961	491,159
Test Set Perplexity	18.4	21.4

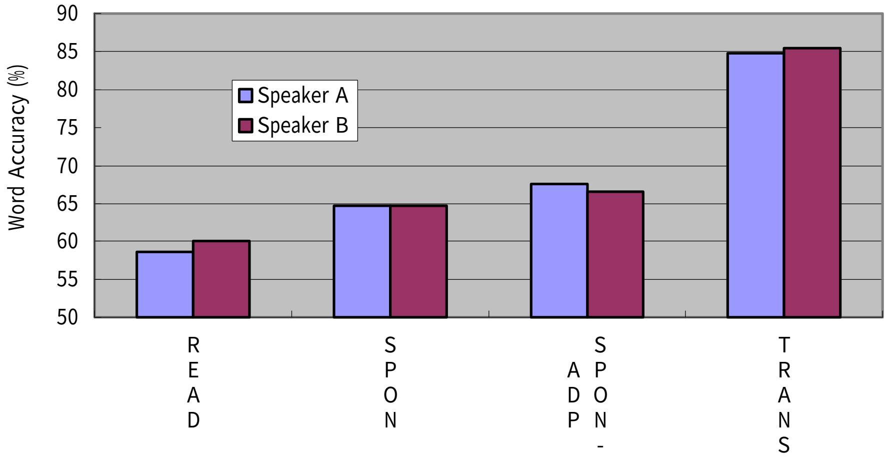

Figure 1 Speech recognizer ’s performance

3. Speech recognizer’s performance and data collection

Figure 1 shows the SPREC speech recognizer’s performance in a monolingual conversation through a tele-conferencing system between two Japanese persons in the travel arrangement task/domain. SPREC is a speech recognition subsystem used in the ATR-MATRIX speech translation system. Speakers are requested to hold a conversation in a spontaneous speaking style. The conversation is transcribed for the following experiment. In Figure 1, three acoustic models are used: a read acoustic model (READ), a spontaneous acoustic model (SPON), and a speaker adaptation model made from the spontaneous acoustic model (SPON-ADP). In Figure 1, the SPON acoustic model shows better performance than READ’s. However, the recognition rate is low because the speakers speak spontaneously. The speaker adaptation model can improve the recognition rate by just a few points. A drastic improvement is observed when speakers read transcription of the conversation data (TRANS) using the SPON acoustic model. These facts lead us to focus on language data collection that assumes a machine-friendly speaking style.

4. Cell-Formed Registration

To explain the proposed scheme, Japanese-to-English paraphrasing is used. However, the actual data collection is done in an English-to-Japanese direction. As shown below below, cell-formed registration efficiently registers the paraphrased sentences. In the conventional form, three sentences are input in the following form:

(1) Hi, I'd like to make a reservation.
(2) May I please make a reservation?
(3) Can I make a reservation?

In the cell-formed registration, we obtain a compact expression by deleting duplicate expressions:

I’d like to May I please Can I	make a reservation.

Since human paraphrasers are used, the expanded sentences using the cell-formed registration are all natural, just careless registration leading to overgeneration.

Our concern is English-to-Japanese data collection. We collected paraphrased Japanese sentences for 130 English seed sentences in the travel arrangement task/domain. The paraphrasers type the results by themselves. This input process can drastically reduce the cost-intensive transcription process. Table 2 shows one example of the Cell-formed registration scheme in the English-to-Japanese direction. Figure 2 shows statistics for 130 English seed sentences. Each plot is the average relative frequency for the expansion rate rounded by 2 to the power of N (N=1,2,…). An English sentence would expand to 423.5 Japanese sentences on average. The maximum expansion rate was 12,898. The minimum expansion rate was 2. We are now working to expand the number of English seed sentences to 10,000. If the same statistics hold, the expected number of produced sentences would be roughly 10,000 * 400, or 4 M sentences. In the proposed scheme, we are estimating that it will take six months to collect data of about 200 times the amount of conventional ATR's SLDB data. Considering time and volume of the data, it has roughly a three-digit level of efficiency compared with the conventional scheme. The collected sentences will be similar expressions, however they will all be natural sentences. It will be very interesting to use this huge set of data for evaluating the performance of a language model and language translation. The issue of uniformity in the collected corpus remains a topic for future work.

Table 2 Example of data collection

How many hours will you be late?			
Nanjikan	— kurai gurai hodo teido	okure osokunari	masuka soudesuka
dore	kurai gurai hodo	okure osokunari	masuka soudesuka
dono	kurai gurai teido	okure osokunari	masuka soudesuka

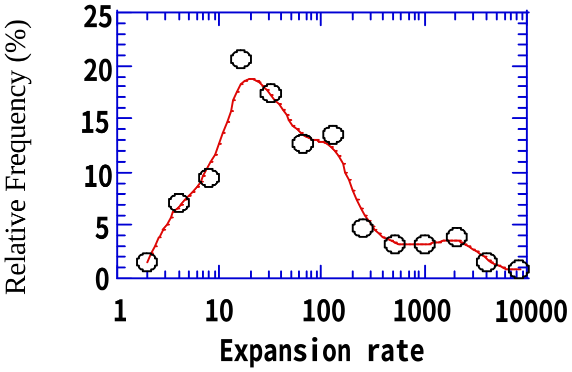

Figure 2 Distribution of expansion rate

5. Conclusion

The cell-formed registration scheme is proposed as a method to efficiently collect a large language DB. Its preliminary evaluation shows a three-digit efficiency in data collection. Application of this large DB to a language model and language translation remains an interesting issue for future work.

6. Acknowledgements

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

Fumiaki Sugaya, Keiji Yasuda, Toshiyuki Takezawa, and Seiichi Yamamoto. 2001. Precise measurement method of a speech translation system's capability with a paired comparison method between the system and human. In Proceedings of the Machine Translation Summit VIII, pages 345-350.

Fumiaki Sugaya, Toshiyuki Takezawa, Akio Yokoo, Seiichi. Yamamoto. 1999. End-to-end evaluation in ATR-MATRIX: speech translation system between English and Japanese. In Proceedings of Eurospeech'99, pages 2431-2434.

Tsuyoshi Morimoto, Noriyoshi Uratani, Toshiyuki Takezawa, Osamu Furuse, Yasuhiro Sobashima, Hitoshi Iida, Atsushi Nakamura, Yoshinori Sagisaka, Norio Higuchi, and Yasuhiro Yamazaki. 1994. A speech and language database for speech translation research. In Proceedings of the 3td International Conference on Spoken Language Processing, pages 1791-1794.

Toshiyuki Takezawa, Tsuyoshi Morimoto, and Yoshinori Sagisaka. 1998a. Speech and language databases for speech translation research in ATR. In Proceedings of the 1st International Workshop on East-Asian Language Resources and Evaluation - Oriental COCOSDA Workshop'98 -, pages 148-155.

Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. 1998b. A Japanese-to-English speech translation system: ATR-MATRIX. In Proceedings of the 5th International Conference on Spoken Language Processing, pages 2779-2782.

Toshiyuki Takezawa. 1999. Building a bilingual travel conversation database for speech translation research. In Proceedings of the 2nd International Workshop on East-Asian Language Resources and Evaluation – Oriental COCOSDA Workshop'99 -, pages 17-20.