Building a Conversation Corpus by Text Derivation from "Germ Dialogs"

Naoki Asanoma, Setsuo Yamada, Osamu Furuse, and Masahiro Oku

NTT Cyber Solutions Laboratories, NTT Corporation
{asanoma.naoki, yamada.setsuo, furuse.osamu, oku.masahiro}@lab.ntt.co.jp

Abstract. We propose a method for building a spoken-language text corpus for a spoken-language system. Conventional methods to build a new corpus include transcribing recorded conversations, collecting text from existing documents, or writing original texts. However, these often have difficulties, such as insufficient corpus size and low cost effectiveness, when preparing the text data in the applied system's domain. To address these issues, we have developed a method that uses "germ dialogs," which are short-scripted dialogs that enable writers to continue or replace them in a logical sequence that sounds natural. This enables the corpus size to be proliferated in a cost-effective manner. Our results show that the proposed method can be used to create higher degree of adequateness for the system's domain than conventional methods. The text data collected for the proposed method are used to generate a language model for our speech translation system between English and Japanese.

1. Introduction

A text corpus plays a crucial role in many spoken-language applications, such as speech translation and statistical natural language processing. The system's accuracy often depends on whether we can accumulate a large amount and wide variety of text data containing frequent or domain-specific linguistic expressions. However, there are fewer existing spoken-language corpora than there are written-language corpora. To make matters much more difficult, spoken-language corpora specific to the systems' domain is often unlikely to even exist. For these reasons, we must make an effort to build a spoken-language corpus in the system's domain. Conventionally, a spoken-language corpus has been built using the following four methods:

(a) Collecting text from existing documents :
The text related to the system's domain is copied from existing documents. Electronic data can be also used in some cases.

(b) Transcribing recorded dialogs (Hirschman, 1992; Heeman & Allen, 1995; Takezawa, 1999; Allwood et al., 2000)**:**
A scripted, situational dialog is recorded. Then, the recorded dialogs are transcribed.

(c) Storing keyboard chats (Kikui et al., 2003):
Two participants chat through their keyboard terminals according to preferences or interests. The chat logs are stored as text data.

(d) Writing imaginatively (Hirasawa et al., 2004):
Given specific conversational scenarios, writers imagine the following scenes and then create sentences that are likely to be uttered.

If we can find a lot of text related to the system's domain, (a) is the most suitable method. However, most of the time very little text exists. Additionally, copyright problems can arise. To avoid these problems, method (b) or (c) is usually used. Method (b) approximates the scenes to which the system will be actually applied and produces good quality text. For example, the CALLHOME corpus from the Linguistic Data Consortium was constructed using this approach (CALLHOME, 1996). However, using method (b) the quantity is apt to be small because it requires at least two people, and it takes a large amount of labour to build a large corpus. Method (c) has the same problem.

In contrast to these methods, method (d) reduces the cost of construction. We can create bigger volumes of text using method (d) or a compromise between (c) and (d), in which just

one person imaginatively writes chat texts. However, it is difficult to persistently create the variety of expressions available in either method because only one person has a limited imagination.

Although combining paraphrases of fragmentary linguistic expressions can create a lot of example sentences in a single sitting, such texts do not accurately reflect the statistics of linguistic phenomena. Moreover, in natural conversation we can not prepare all scenes in advance.

To overcome the problems of conventional methods, we propose a method for easily proliferating conversation texts that can reduce costs by providing writers with "germ dialogs". The germ dialogs are short scripted dialogs that enable the writers to easily image a follow-up dialog. This method is an improvement over the creative writing method (d).

The remaining part of the paper is organized as follows. Section 2 explains the method of deriving text from germ dialogs. Section 3 describes the corpus built by the proposed method. Section 4 presents evaluations of the proposed method, based on language models made from prepared corpora. Section 5 will describe our conclusions.

2. Deriving Text from Germ Dialogs

In our approach, new dialog texts are derived from the germ dialogs, which can be prepared by creative writing or spoken dialog transcription in conventional methods and by taking excerpts from published conversation books. Because the size of derived texts is assumed to be much larger than that of germ dialogs, we can exclude germ dialogs from the corpus, thus avoiding the copyright issue, even if the excerpt is taken from copyrighted materials. The germ dialogs set the conversation scenes and make it easier to create sentences than to imagine them. Moreover, they control the consistency of the topics and style set by each germ dialog.

We derive text from germ dialogs by applying such techniques as "retrace", "fill-in", "replacement", and "follow-on". Figure 1 illustrates the text derivation from a germ dialog between two speakers, A and B. "A1" means the first utterance of speaker A, and the number denotes the utterance order.

Retrace:

Writers create all the possible dialogs arising from the germ dialog.

Fill-in:

Some utterances in a germ dialog are left blank intentionally. The writer fills in the blanks with as many reasonable expressions as they can imagine in a given context.

Replacement:

Some utterances in a germ dialog are replaced with new utterances produced by writers. The new utterances may have the same meaning as the originals, or they may be paraphrased or even have a different meaning than the original.

Follow-on:

This technique is an easier way to expand the dialog corpora without losing naturalness, as compared with above three techniques. Additionally, it can be positioned as the main part of the proposed text derivation method.

Writers are asked to create two or more possible utterances that might reasonably follow the germ dialogs. Then, two or more responses are also created for each possible utterance. For instance, one utterance may yield two utterances in response: these two draw two responses each, or four sentences in all, and so on. In this way the dialog grows exponentially.

In addition to the possible utterances, the number of dialog turns is a direct factor in determining how much the corpus size can be expanded. However, we consider that the follow-on dialogs should be limited to three or four turns, because more turns might dilute the germ dialog.

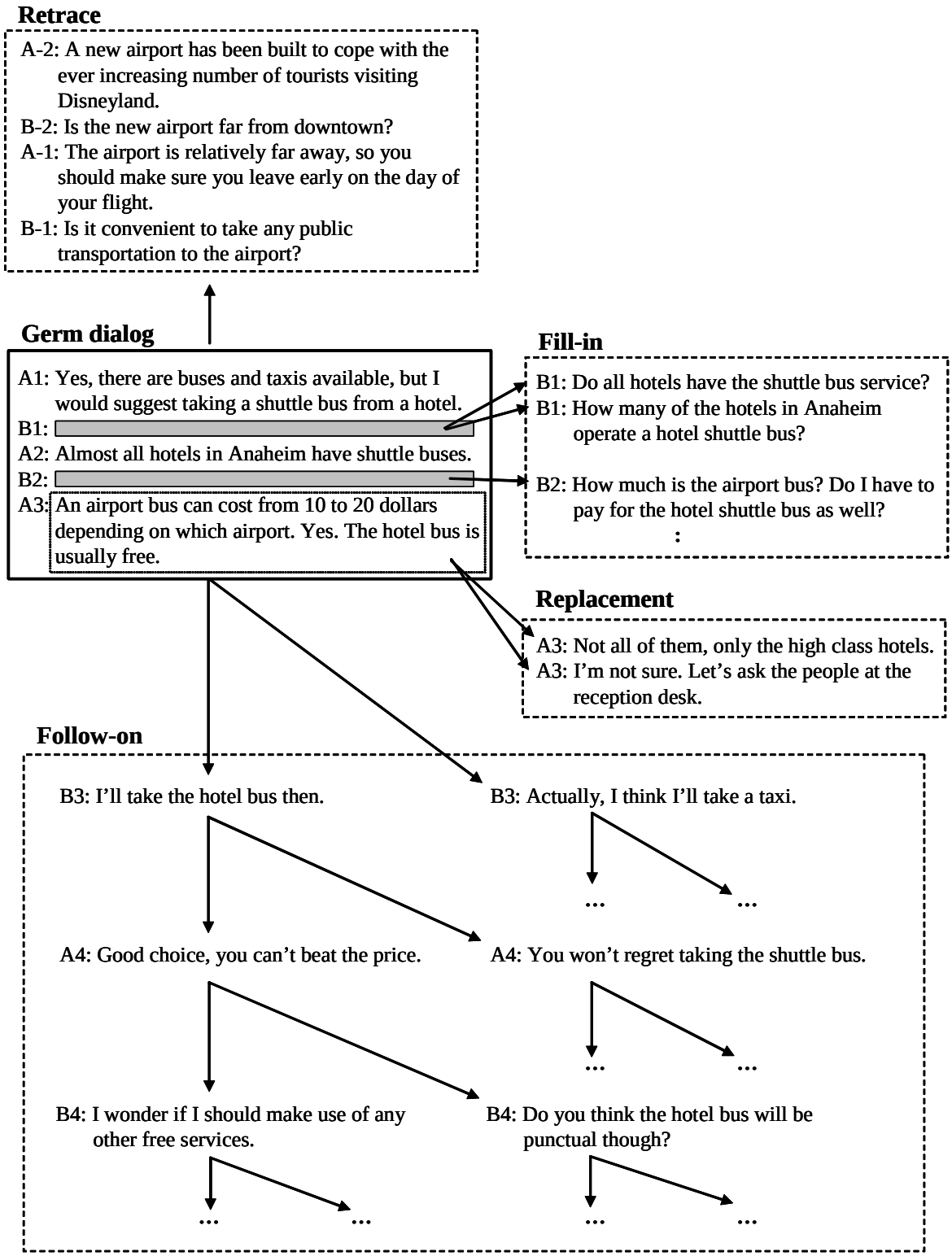

Figure 1: Text Derivation from a Germ Dialog

3. Building a Conversation Corpus

We used the proposed method to build a conversation corpus in English and Japanese. We selected fifty germ dialogs from each language from published books, including daily conversations about various topics, such as weather, health, travel, movies, and so on. The length of the germ dialogs was two or three dialog turns, which was enough to imagine the conversational situation clearly. Ten writers were then asked to create utterances from each of the fifty germ dialogs in their native language through the "retrace", "fill-in", "replacement", and "follow-on" techniques as follows. The numbers below are the same in English and Japanese.

Retrace: One dialog consisting of two turns (four utterances) was created for each germ dialog by each writer. The total number of utterances was 2,000; (50 germs ∗ 1 dialog ∗ 4 utterances ∗ 10 people).

Fill-in: Two utterance blanks for each germ dialog were filled in with two possible utterances, selected by each writer. The total number of utterances was 2,000; (50 germs ∗ 2 blanks ∗ 2 utterances ∗ 10 people).

Replacement: One utterance was replaced with three possible utterances for each germ dialog by each writer. The total number of utterances was 1,500; (50 germs * 1 utterance * 3 replacements * 10 people).

Follow-on: Each writer followed through two turns, using binary branching for each germ dialog. The total number of utterances was 15,000; (50 germs * (2 + 4 + 8 + 16) utterances * 10 people).

Thus, 20,500 utterances (2,000 + 2,000 + 1,500 + 15,000) were created from 50 germ dialogs, and this was about one hundred times the total utterances of the 50 germ dialogs, each of which consisted of about four utterances.

Some utterances created consisted of two or more sentences. For example, 15,000 Japanese utterances created using "follow-on" resulted in 18,059 Japanese sentences.

4. Evaluation

We applied the corpus built using our method to provide a speech translation system using the domain's statistical language model, where the possibility of word transition was statistically calculated for speech recognition, parsing and so on. To verify the cost effectiveness of our method, we evaluated a Japanese language model generated from the corpus by comparing the proposed method with a conventional method (c). We also compared the quality of the corpus built by the proposed method with that of an existing corpus, CALLHOME, created using method (b).

During our evaluation, we compared the costs according to the work load differences required to build the same amount of text, using the same number of skilled staff. We measured the adequateness to the domain by comparing the coverage and perplexity of the language model. These were based on the idea that the wider coverage is and the lower perplexity is, the better the quality of the corpus becomes. Our statistical language models were generated from word trigrams calculated from the corpora. Coverage and perplexity are often used as indices of statistical language models in research areas, such as speech recognition.

4.1. Comparison with Keyboard Chats

As a conventional method, we built the corpus from keyboard chats; method (c) in Section 1. Two people were asked to have a keyboard chat about a specific topic. They were also asked to type texts as if they were actually speaking. Using this method, we collected 1,464 Japanese utterances or 2,753 sentences. To evaluate the proposed and conventional methods under the same conditions, we randomly extracted 2,753 sentences from the corpus built using our new "follow-on" method.

As an evaluation corpus for actual spoken dialogs, we chose CALLHOME Japanese transcripts (LDC96T18) that are available for general use; that is, the comparisons were between the proposed method & CALLHOME and keyboard chat storage & CALLHOME. We extracted 25,822 sentences from CALLHOME corpus, which were then 'pruned' by removing their tagging information. The coverage and perplexity of the two language models was calculated for this evaluation corpus. The perplexity values of the two methods were calculated for a 5,893-word vocabulary set, which was constructed by merging the text vocabularies built from the two methods.

Table 1: Comparison between the Proposed Method and the Keyboard Chat Storage

	Adequateness to CALLHOME		Relative Cost	# of Sentences
	Coverage	Perplexity		
Proposed Method	84.0%	513.9	1	2,753
Keyboard Chat Storage	82.8%	667.0	Between 4 and 5	2,753

Table 2: Comparison between the Proposed Method and CALLHOME

	Adequateness to Keyboard Chats		# of Sentences
	Coverage	Perplexity	
Proposed Method	90.9%	208.5	18,059
CALLHOME	87.3%	654.8	25,822

Our experimental results are shown in Table 1. The relative costs are defined here as the work-load ratio to the workload of collecting 2,753 sentences using the proposed method. (So the relative cost, or the work-load ratio of the proposed method is 1.) In Table 1, our method clearly shows a higher degree of adequateness for the system's domain, with higher coverage and lower perplexity than conventional methods. The keyboard chat storage cost about four or five times more than the proposed because method (c) requires at least two people at the same time, while just one person can work on the proposed method. This result means that the proposed method can create a better quality conversation corpus faster than conventional methods using the same amount of money. That is why the proposed method of using germ dialogs is cost-effective for collecting dialog texts. In addition, the dialog texts derived from the germ dialogs were subjectively evaluated as being as natural as spoken dialogs.

The CALLHOME corpus includes many coarse or slang expressions while the two evaluated corpora include many well-formed or polite expressions. These differences in expression styles seem to have caused relatively high perplexity values.

4.2. Comparison with an Existing Corpus

We also evaluated the proposed "follow-on" method by comparing it with conventional method (a) in Section 1, which uses CALLHOME Japanese transcripts. In this evaluation, 400 sentences, which were randomly extracted from keyboard chat texts, were used as an evaluation corpus. A vocabulary of 18,607 words was created by merging vocabularies taken from both methods in the texts. As shown in Table 2, both the perplexity and the coverage of the proposed method are much better than those of the existing CALLHOME corpus, although method (a) costs less money.

5. Concluding Remarks

This paper describes a method for effectively building a conversation corpus using text from dialogs based on germs and developed by writers. To improve the cost effectiveness, the initial dialogs were provided to the writers as "germ dialogs." Therefore, they could easily imagine the dialog that follows logically from the germ dialogs. Our method will enable us to build the corpus cost-effectively. Our results showed that the proposed method can produce a high degree of adequateness for the system's domain more cost-effectively than conventional methods.

The text data collected using the proposed method can now be used to generate a language model for speech translation systems for goal-oriented communication between English and Japanese (Asanoma et al., 2004; Kataoka et al., 2004). We are planning to apply our speech translation system to daily conversation using the corpus built using our method. We also believe that the data can be used for statistical parsing

or machine translation.

The quality of the corpus built by the proposed method depends on the quality of the germ dialogs. Therefore, one of our future projects will be to decide how features in the germ dialogs can be used to create a better corpus.

6. References

ALLWOOD, J., BJÖRNBERG, M., GRÖNQVIST, L., AHLSÉN, E. & OTTESJÖ, C. (2000). 'The Spoken Language Corpus at the Linguistics Department, Göteborg University'. In Forum Qualitative Social Research, Vol. 1, No. 3, December 2000.

ASANOMA, N., YAMADA, S., FURUSE, O., OKU, M., KATAOKA, A. & TAKAHASHI, Y. (2004). 'A Cross-Lingual Communication System with Agent-Mediated Architecture'. NTT Technical Review. Vol. 2, No. 12, 56-61, December 2004.

CALLHOME Japanese Transcripts, (LDC96T18). http://www.ldc.upenn.edu/Catalog/LDC96T18.html, Linguistic Data Consortium (LDC).

HEEMAN, P.A. & ALLEN J. (1995). 'The TRAINS 93 Dialogues'. Trains Technical Note 94-2, March 1995.

HIRASAWA, J., AMAKASU, T., YAMAMOTO, S., YAMAGUCHI, Y. & IMAMURA, A. (2004). 'Cyber Attendant System with Spontaneous Speech Interface'. NTT Technical Review. Vol. 2, No. 3, 64-69, March 2004.

HIRSCHMAN, L. (1992). 'Multi-Site Data Collection for a Spoken Language Corpus'. In Proceedings of the 'Fifth DARPA Speech and Natural Language Workshop', Morgan Kaufmann, February 1992, NY.

KATAOKA, A., ASANOMA, N., TAKAHASHI, Y., YAMADA, S. & FURUSE, O. (2004). 'Agent-Mediated Architecture for Efficient Goal-Oriented Communication across Languages'. In Proceedings of 'Asian Symposium on Natural Language Processing to Overcome Language Barriers', 24-30, March 2004, Hainan Island, China.

KIKUI, G., SUMITA, E., TAKEZAWA, T. & YAMAMOTO, S. (2003). 'Creating Corpora for Speech-to-Speech Translation'. In Proceedings of 'Eurospeech-2003', 381-384, September 2003, Geneva, Switzerland.

TAKEZAWA, T. (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', 17-20, May 1999, Taipei, Taiwan.