Input Text Repairing for Multi-lingual Chat System

Kenichi Yoshida and Fumio Hattori

Graduate School of Science and Engineering, Ritsumeikan University, Japan
cc016049@is.ritsumei.ac.jp, fhattori@is.ritsumei.ac.jp

Abstract. Even though various communication tools have resulted in a remarkable increase of global communications, language barriers remain high and complicate communication across languages. Although the multi-lingual chat system allows users to chat with each other in different language using machine translation. the quality of translation is not so high when the input sentence reflects spoken language. In this paper we propose a method that repairs the input sentences in spoken language by retrieving similar sentences using keywords.

Keywords: Language Grid, multi-lingual chat, cross-cultural communication, machine translation.

1 Introduction

The advance of the Internet technology and various communication tools have resulted in a remarkable increase of global communications. However, language barriers remain high and complicate inter-cultural communication. The Language Grid Project [1], which is an infrastructure that makes it possible to combine various language resources on the Internet, started to solve this problem in 2006 by improving the understanding of the Internet contents written in different languages and by people from different countries.

The multi-lingual chat system[1] [2], which is one of the applications developed by the Language Grid Project, allows users to chat with each other in different languages using machine translation. Most sentences in multi-lingual chats are spoken words. However, almost all of the machine translation resources used in multi-lingual chat systems translates written words. Therefore, the translation quality in multi-lingual chat systems is not always ensured. Although a multi-lingual chat system provides such functions as back translation and auto completion, they are insufficient for practical use. To improve the quality, repairing input spoken sentences to written sentences that match the machine translation is expected to be effective. In this paper, a method is proposed that repairs input sentences in spoken language by retrieving similar sentences using keywords.

[1] Multi-lingual chat system is developed by College of Informatics, Kyoto University.

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 210–217, 2009.

In Section 2, the details of the multi-lingual chat system are introduced. An input repairing system is proposed in Section 3. In Section 4, experiments and discussion are described. Section 5 concludes our paper.

2 Multi-lingual Chat System

2.1 Overview of Multi-lingual Chat System

A multi-lingual chat system adds a translation function to traditional text chat systems. Users can send messages in their native languages and receive messages from partners. In other words, users can exchange messages with partners whose native languages are different.

A multi-lingual chat system utilizes the machine translation resources available on the Internet. However, current machine translation systems are designed to handle written documents: that is, well-formed sentences that are easily and correctly translated. However, the spoken words that often appear in chats are rarely translated correctly because the translation quality is not high in multi-lingual chat systems. For example, “Shukudai susunderu?” in Japanese, which means “How have you finished your homework?” might be translated as “Is homework developed?”

The characteristics of the Japanese spoken sentences used in chatting are different from written sentences. First, subjects are often omitted. In addition, people often answer with very simple predicates. Second, the end of a sentence can be spelled in several different ways. Third, several words have the same meaning but the expressions are different. For example, Japanese has at least three expressions that mean “dinner”: “yorugohan,” “yuhan,” and “banmeshi.” These factors prevent machine translation from translating correctly. To improve the quality of the translation of multi-lingual chats, repairing input sentences to adapt to machine translation is considered to be effective.

2.2 Back Translation

To repair input sentences, the multi-lingual chat system provides a back translation function. Fundamentally speaking, to confirm whether the input sentence is translated correctly, the target language must be understood. However, this is rare for the users of multi-lingual chat systems.

Back translation re-translates the translated sentence in the target language to the source language. Users can compare the original input and the back-translated sentences and confirm whether the meaning of both sentences is the same. If the meanings are identical, users can expect that the input sentence was translated correctly. Conversely, if the meanings are different, the translation might be incorrect. In the latter case, users need to repair the input sentence and translate again by repeating this process until the meanings become equivalent. Using the above example, when “Is homework developed?” is back-translated into “Naishoku ha hatten saserareruka?”, the user realizes that the input sentence was not translated correctly into his/her native language.

A preliminary experiment showed that 65% of sentences need to be repaired twice or more using back translation. This means that this function is very useful for improving translation quality. However, it only informs users that the input sentence is not good. Users still have to repair the input sentences by themselves. Learning how to repair input sentences is not easy. The fact that 65% of sentences need to be repaired twice or more impairs the immediacy of the chat system.

2.3 Auto Complete

Another function, auto complete, which is also equipped in the multi-lingual chat system of the Language Grid Playground, retrieves example sentences that match the input text. Since the example sentences have corresponding translated sentences, the selected example sentence is translated instantly and correctly. However, it can't handle the variety of spelling at the end of sentences or the variety of synonyms. In addition, the number of examples is limited, so the auto complete function is only useful in few cases.

3 Input Text Repairing by Retrieving Generalized Sentences

3.1 Method Overview

Given an input sentence, this function retrieves generalized sentences in written styles using keywords. A generalized sentence is one in which several words are replaced to generalized words. For example, "ongakushitsu wa doko desuka?" ("Where is the music room?") can be generalized as "«basho» wa doko desuka?" ("Where is «the place»?"). In this example, «basho» («the place») is the generalized word. The keywords used for retrieval are extracted from the input sentence and generalized.

From input sentence "Ongakushitsu, doko?" ("Music room, where?"), "ongakushitsu" ("music room"), "doko" ("where"), and "?" are extracted. Then "ongakushitsu" ("music room") is generalized as «basho» («the place»), so that the keywords used for retrieval are «basho» («the place»), "doko" ("where"), and "?". The grammatical structure of the input sentence is disregarded in this method. Retrieving a generalized sentence database might return the following sentence: "«basho» wa doko desuka?" ("Where is «the place»?"). Finally, the generalized words are specialized as they appear in the input sentence, resulting in the following repaired input sentence: "ongakushitsu wa doko desuka?" ("Where is the music room?").

The generalized sentence database includes sentences whose use can be anticipated based on each domain in which the system is used. For example, if the system is used in an elementary school, sentences about schoolwork or classrooms must be registered. It also includes sentences commonly used in daily life such as asking about places or exchanging information.

3.2 Process Flow

Fig. 1 shows the process flow of the input repairing method. The example of input repairing for a dialogue between a teacher and a foreign pupil is shown in Fig. 2.

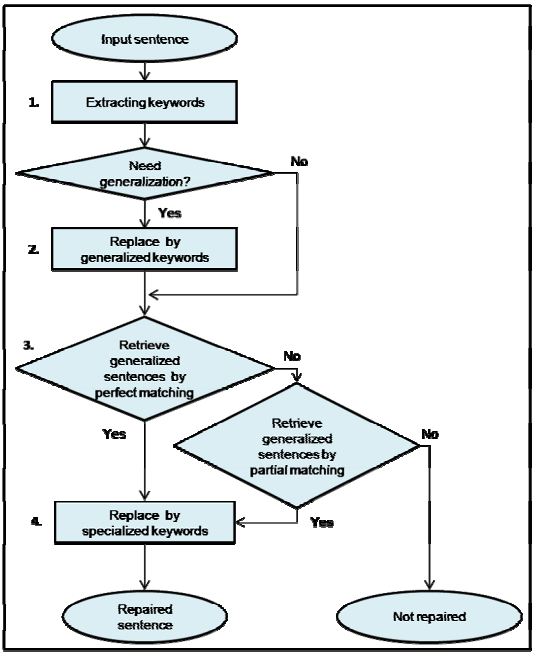

Fig. 1. Process flow

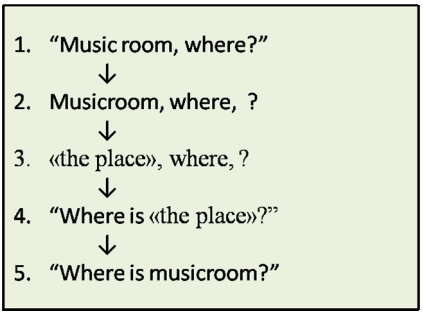

Fig. 2. Example of input repairing

1. Extracting keywords:
   First, by applying morphological analysis, nouns, adjectives, independent verbs, and question or exclamation mark are extracted. For example, from the sentence "Ongakushitsu, doko?" ("Music room, where?"), "ongaku" (noun-generality), "shitsu" (noun-suffix-generality), "doko" (noun-pronoun-generality), and "?" (mark-generality) are extracted. Suffixes are combined with the preceding noun. That is, "noun-suffix-generality" is always combined with the preceding "noun-generality" to form one word. So "ongaku" (noun-generality)" and "shitsu" (noun-suffix-generality)" are combined into "ongakushitsu" ("music room").

2. Keyword generalization:
   In this step, the extracted words are replaced by generalized words using domain ontology, which is a formal representation of a set of concepts within a domain and the relationships among those concepts. Section 3.3 explains domain ontology in detail.
3. In this paper, only noun-generality is generalized. In the preceding example, "ongakushitsu" ("music room") is generalized as "«basho»" ("«the place»"). After the generalization, the three keywords, "«basho»" ("«the place»"), "doko" ("where"), and "?", are used for retrieving a generalized sentence database.
4. Retrieving the generalized sentence database:
   In the beginning, the generalized sentences match all three keywords perfectly. If no sentence is matched, partial matching is done using one or two keywords.
5. Word specialization:
   Generalized words in the retrieved generalized sentence are replaced backwards with specialized words, as in step 2. In the preceding example, "«basho»" ("«the place»") is replaced again with "ongakushitsu" ("music room"), and the sentence is shown to the user.

3.3 Keywords Generalization Using Domain Ontology

Sentence variety appears as input of the chat system. Since preparing all example sentences is impractical, we focus on the fact that many sentences have the same structure, but only the noun in the sentence differs. These sentences can be generalized to one sentence that includes generalized words instead of the original nouns. Therefore, keywords extracted from input sentences have to be generalized to retrieve the generalized example sentence. This generalization is done using domain ontology. Fig. 3 shows a sample domain ontology for places in a school. Keywords are generalized by ascending the concept hierarchy in the domain ontology.

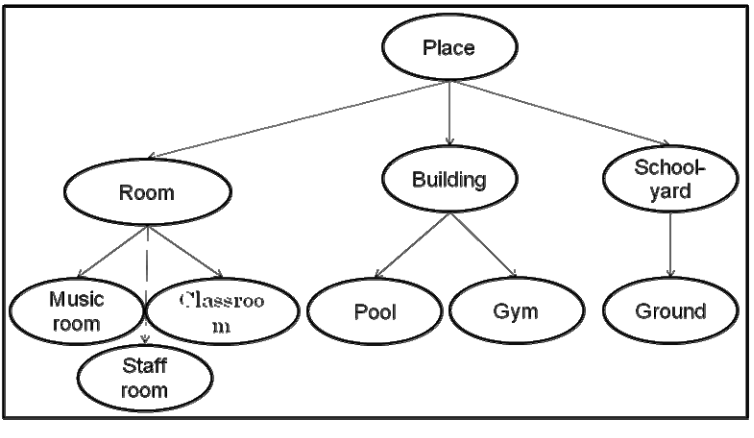

Fig. 3. Sample of domain ontology

3.4 Retrieval of Generalized Sentences

A generalized sentence database is composed of pairs of generalized sentences and corresponding keywords. Fig. 4 shows the structure of the database. Generalized sentences are developed by generalizing examples in the domain using the same method for generalizing input sentences as described above.

Key Word Group	Generalized Sentence
«the place», where, ?	Where is «the place»?
today, «lesson», «the place», give	Today's «lesson» is given in «the place».
⋮	

Fig. 4. Concept chart of generalized database

Retrieval is done by comparing the keywords extracted from the input sentence with the keywords in the generalized sentence database. If all keywords match, a corresponding generalized sentence is returned. When no sentences match, two options are provided: A and B. Option A retrieves sentences that include more than four words that match three or more keywords. In the example of Section 3.1, "«basho» wa doko ni aruka shitte imasuka?" ("Do you know where «the place» is?" is retrieved using keywords "«basho»" ("«the place»"), "doko" ("where"), and "?". Option B retrieves generalized sentences from the database with one keyword missing, but the combination of keyword sentences partially matches the keywords, including the generalized keywords. For example, the used keywords are the generalized keyword "«basho»" ("«the place»") and another keyword, such as "doko" ("where"). The result might be "«basho» ha doko desuka? ("Where is «the place»?").

3.5 System Architecture

Fig. 5 shows the system architecture of the input repairing system.

1. The input sentence is passed through the morphological analysis module to extract keywords from it.
2. Keywords are generalized using the domain ontology.
3. Generalized sentence database are retrieved using generalized keywords.
4. The generalized word is returned to the original one.
5. The repaired sentence is presented to the user.

A sample snapshot of the multi-lingual chat system with an input repairing facility is shown in Fig. 6. When a sentence is input into the input field and the translation button is pushed, the repaired candidate sentences appear below.

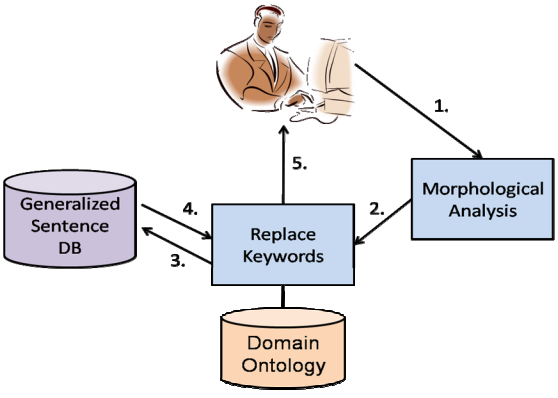

Fig. 5. System configuration

Fig. 6. Sample snapshot of input repairing

4 Experiment

4.1 Experimental Method

An experiment was conducted supposing a chatting situation between a Chinese pupil and a Japanese teacher at an elementary school. The translation quality, which was compared with and without input repairing, was measured by the number of back translations needed until an acceptable translation result was obtained. The translation of Japanese into Chinese was assumed.

4.2 Experimental Result and Discussion

Using multi-lingual chat without input repairing, 10 of 23 sentences were translated correctly, a success rate of 43%. On the other hand, using the system with the

proposed input repairing, 19 of 23 sentences were translated correctly, a success rate of 83%.

Table 1 shows several examples of input repairing. In the first, the back translation result implies that the input sentence was mistranslated because an article was inadequately supplemented by the machine translation. In this case, input repairing worked well. The second example shows that input repairing caused a mistranslation. Since the repaired sentence was complex, perhaps the machine translation could not understand it well. The last one was caused by incorrect morphological analysis.

Table 1. Examples of input reparing

Input sentence	Back translation	Judgment	Extracted key words	Repaired Input sentence	Back translation	Judgment
Doko iku no? (Where, go?)	Doko ga ikimasu ka? (Does where go?)	NG	doko, iku, ? (where, go, ?)	Doko ni iku no desuka? (Where do you go?)	Doko ni iki masuka? (Where do you go?)	OK
Sensei ni taiiku yasumutte tsutaete. (Tell teacher, absent from gym)	Sensei no taiikukan ni yasun de tsutaeru. (I' ll absent and tell to teacher' s gymnasium.)	NG	«jyugyou», yasumu, tsutacru («class», absent, tell)	Taiiku wo yasumu to tsutae te kudasai. (Please tell that I will absent from gym.)	Yasumi no taiikukan wo dentatsu shite kudasai. (Please inform absent gym.)	NG
Marathon rashii yo. (Appears marathon)	Marathon no youdesu. (It a;;ears to be marathon.)	OK	marathon ra, shiiru (marathons, force)	Failed	–	–

5 Conclusion

A method was proposed to repair input sentences for multi-lingual chat systems. The method retrieves similar generalized sentences using keywords extracted from input sentences. The experiment shows that the successful translation rate was improved from 43% to 83%. The idea of retrieving generalized sentences using keywords might be applicable for young children or people with such language difficulties as aphasia. Input repairing must still be improved. For example, complex sentences must be decomposed into simple sentences. Flexibility that can satisfy the cases of morphological analysis error is also desired.

Acknowledgement

This research was supported by the Strategic Information and Communications R&D Promotion Programme of the Ministry of Internal Affairs and Communications, Japan.

References

1. Ishida, T., Grid, L.: An Infrastructure for Intercultural Collaboration. In: IEEE/IPSJ Symposium on Applications and the Internet (SAINT 2006), keynote address, pp. 96–100 (2006), http://langrid.org/
2. http://langrid.org/playground/chat/ChattingMain.html