Cross-lingual English Sentence Retrieval Considering Syntactic Structures of Japanese Queries

Kazuya Kuzuhara, Yoshihide Kato, and Shigeki Matsubara

Abstract—**This paper proposes a cross-lingual English sentence retrieval system for Japanese queries. The system retrieves English sentences exactly by considering dependency structures in both Japanese and English. It enables the system to reflect the user's intention. We utilize a constraint on dependency structure of Japanese, i.e, any dependency is directed from left to right. By considering the constraint, our system can eliminate irrelevant sentences from search results. We conducted an experiment to evaluate the system. The results demonstrated that the rate of improvement of retrieval precision reached to 38.0% and the precision of filtering was 95.3%. Thus, we confirmed the effectiveness of our proposed system.**

I. Introduction

WITH the progress of globalization, the opportunities of writing English sentences are increasing in the world. However, it is difficult for nonnative speakers to write natural English sentences. For this reason, a support environment to write English sentences efficiently is desired.

Generally, when Japanese people write English sentences, the contents to be written have been already prepared in Japanese. Nevertheless, it is difficult for Japanese people to write English sentences correctly because they do not know the English expression corresponding to Japanese well. For this problem, it is effective to provide an environment which can present English expressions corresponding to Japanese expressions as examples in order to write English sentences.

In this paper, we propose a method of retrieving English sentences with Japanese queries. Our method converts Japanese keywords into English words, and seeks the sentences which include the English words. To reflect the user's intention, our method filters out the irrelevant sentences by considering dependency structures of Japanese queries and English sentences. Considering dependency structures, the system can show sentences corresponding with user's intention.

We have developed a cross-lingual English sentence retrieval system *EscortCross*. The system uses a bilingual computerized dictionary to translate Japanese keywords into English, and a dependency parser to parse English sentences. The system introduces an example sentence retrieval method [1] in order to seek relevant sentences. We conducted an experiment to evaluate the system. The results of the experiment demonstrated that our method achieved high search accuracy in comparison with a simple sentence search. That is, we confirmed the effectiveness of our method.

This paper is organized as follows: Section II describes an outline of our cross-lingual English sentence retrieval. Section III describes our method in detail. Section IV reports the evaluation by a search experiment.

II. English Sentence Retrieval by Japanese Queries

A. Simple Method Using Keywords

Before explaining our method, we first describe a simple method to retrieve English sentences by using Japanese queries. This method consists of the following two steps:

1. Translate a Japanese query into English keywords using a bilingual dictionary.
2. Retrieve sentences by using the English keywords

As an example, let us consider a Japanese query "システムを開発する (develop a system)". The query consists of three words "システム (a system)", "を (case particle)" and "開発する (develop)". For the query, the system translates "システム" and "開発する" into "system" and "develop", respectively, and seeks sentences based on these English keywords[1]. The system returns, e.g, the following sentences.

(1) We also *develop* a pilot *system.*
(2) The technique *developed* here improved the *system.*
(3) The *developed system* is based on a Bayesian network framework.

The sentence (1) is probably relevant to the input query. On the other hand, the sentence (2) is irrelevant. There is no relation between the words "develop" and "system". The sentence (3) is also irrelevant although there exists a semantic relation between "システム" and "開発する".

The simple method, which uses only a bilingual dictionary, returns irrelevant sentences.

Manuscript received June 15, 2009. This research was partially supported by the Grant-in-Aid for Scientific Research (B)(No. 20300092) of JSPS.

Kazuya Kuzuhara is with the Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, 464-8601, Japan (corresponding author to provide phone: +81-52-789-4387; fax: +81-52-789-4385; e-mail: kuzuhara@el.itc.nagoya-u.ac.jp).

Yoshihide Kato is with the Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, 464-8601, Japan (e-mail: yosihide@el.itc.nagoya-u.ac.jp).

Shigeki Matsubara is with the Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, 464-8601, Japan (e-mail: matubara@nagoya-u.jp).

[1] There exists no English word corresponding to the case particle "を".

978-1-4244-4139-6/09/$25.00 ©2009 IEEE

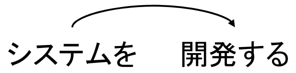

Fig. 1. Dependency structure for Japanese query “システムを開発する”

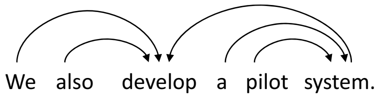

Fig. 2. Dependency structure for “We also develop a pilot system.”

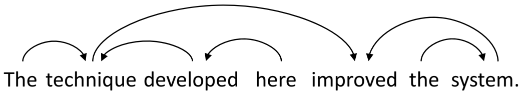

Fig. 3. Dependency structure for “The technique developed here improved the system.”

B. English Sentence Retrieval Method Based on Syntactic Information

In order to eliminate irrelevant sentences from the search results, we adopt an approach of utilizing syntactic structures. We expect that there exists a correspondence relation between the syntactic structure of a Japanese query and relevant English sentences. As an example, let us consider again a Japanese query “システムを開発する”. It is presumed that the user’s intention is to retrieve sentences in which a word corresponding to “開発する” is a verb and a word corresponding to “システム” is its object. This is possible by parsing the Japanese query.

Our method reflects user’s intention by considering the dependency structures of Japanese queries. A dependency structure is a set of dependency relations between words. The dependency structure characterizes syntactic structures by the dependency relations between words which constitute a query. For example, in a query “システムを開発する”, the dependency relation that “システム” modifies “開発する” exists.

In English, a syntactic structure can be similarly characterized by dependency relations. The structure is common to Japanese and English. For example, Figure 1 shows the dependency structure of a Japanese query “システムを開発する” and the dependency structure of the sentence (1) is shown in Figure 2. In Figure 1, “システム” modifies “ 開発する”, and in Figure 2, “system” modifies “develop”. Although the order of the corresponding words is not same by the difference of the word order between the Japanese and English, there are the same dependency relations between corresponding words.

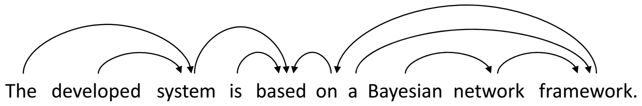

Fig. 4. Dependency structure for “The developed system is based on a Bayesian network framework.”

On the other hand, the dependency structures of the sentences (2) and (3), which are shown in Figure 3 and 4 respectively, do not have the same dependency relation. In Figure 3, no dependency relation exists between “system” and “develop”. In Figure 4, the direction of the dependency relation is different from the dependency relation of the Japanese query. The dependency relations of these sentences do not correspond with the Japanese query. If the system retrieves sentences based on correspondence of dependency relations, the system does not display incorrect sentences such as the sentences (2) and (3). Thus, the system can display sentences matched user’s intention.

III. Cross-lingual English Sentence Retrieval System

This section describes a cross-lingual English sentence retrieval system named EscortCross. This system utilizes dependency structures of Japanese queries.

A. An Overview of EscortCross

Figure 5 shows the configuration of the system. The system consists of three components: translation, search, and filtering.

The translation module receives a Japanese query which is a morphological sequence and returns sequences of English keywords. In order to do this, this module divides a morphological sequence into sequences which are entries of a Japanese-English dictionary. Each subsequence is converted into the corresponding English. In addition, this module analyze each subsequence morphologically. These processes are transforming a verb into its root form, deleting a particle and so on. If there exists a entry of the dictionary for each subsequence, this module returns the sequence of English words.

For example, the Japanese query “システムを開発する” is divided into “システムを” and “開発する”. Since “を” is a particle, the translation module deletes it. Each word is translated into “system” and “develop”. Therefore, the English word sequence of “system” and “develop” is returned. If more than one English word sequences are obtained, this module returns all of them.

The search module is based on the method in the literature [1]. It receives the resulting keyword sequences

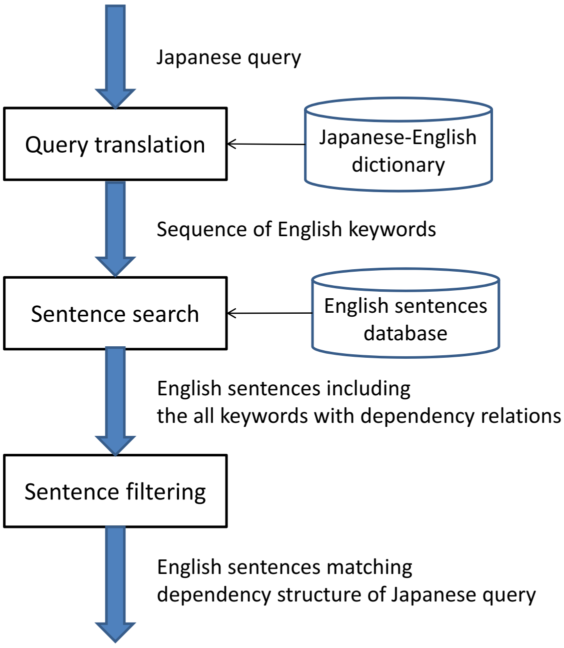
Japanese query
Query translation
Japanese-English dictionary
Sequence of English keywords
Sentence search
English sentences database
English sentences including the all keywords with dependency relations
Sentence filtering
English sentences matching dependency structure of Japanese query

Fig. 5. Configuration of EscortCross

of the translation module, and returns English sentences, which are stored in a database. These sentences include the all keywords and there exist dependency relations between occurrences of the keywords. In the database, dependency structures are given to all the English sentences in advance. For example, we consider the example in section II-A. If this module receives keywords “system” and “develop”, it returns the sentences (1) and (3).

The filtering module receives sentences of the result of search module and eliminates irrelevant sentences by considering the dependency structure of a Japanese query. Finally, the system displays the resulting sentences of filtering as search results. In section III-B, filtering is described in detail.

B. Filtering using Dependency Structure

In order to filter out irrelevant English sentences by using the dependency structure of a Japanese query, it is required to identify dependency structure of the query. One way to identify dependency structure is to use a dependency parser. However, the parser may assign an incorrect structure to the query. Moreover, if the Japanese query has syntactic ambiguity, the resulting dependency structure may not correspond with a user’s intention.

For example, there are two dependency structures of a Japanese phrase “従来のシステムの評価” as shown in Figure 6 and Figure 7. If the parser returns the structure which is different from user’s intention, the system can not return correct sentences.

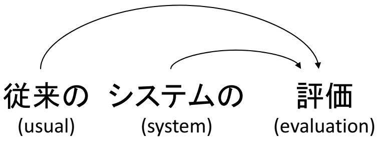

Fig. 6. Dependency structure of “Usual evaluation of the system”

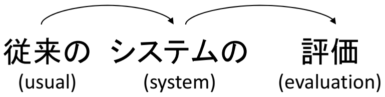

Fig. 7. Dependency structure of “Evaluation of usual system”

In order to avoid this problem, we use a constraint on the dependency structure in Japanese. The constraint is that the dependent word always appears in the left of its head word. The system assumes that any Japanese query satisfies this constraint. That is, it does not display the sentences which have dependency relations violating this constraint. We formalize this constraint as follows:

Japanese dependency constraint: Let E be a set of English keywords and D be a set of dependencies between keywords in E. We write $J(e)$ for the position of the Japanese keyword corresponding to e. For all e_i and e_j in E, if e_i modifies e_j in D then $J(e_i) < J(e_j)$ must hold.

IV. Evaluation Experiment

To evaluate the effectiveness of our proposed filtering, we implemented a cross-lingual sentence retrieval system EscortCross based on the proposed method. This system uses dictionary data of “Kenkyusha’s New Japanese-English Dictionary 5th Edition” [3] as bilingual-dictionary and annotates sentences in database with dependency structures by using the dependency parser RASP [2]. The system was implemented in Perl.

A. Outline of Experiment

In the experiment, we evaluate the effectiveness of filtering by observing whether the proposal system outputs correct sentences for Japanese queries or not. To compare with the proposal system, we developed a sentence retrieval system which does not have the filtering function (conventional system) and evaluated the search performance. We had subjects to retrieve sentences by the conventional system in order to construct a collection of Japanese queries and correct answer data. The subjects judge correction of sentences displayed by the conventional system. We evaluated the filtering function by comparing those data with the result of the proposal system. We measured precision, recall and F-measure by using the correct answers.

TABLE I
EXPERIMENTAL RESULT

	Conventional system		Our system	
Precision(%)	15.3	(117/764)	21.3	(104/488)
Recall(%)	100.0	(117/117)	88.9	(104/117)
F-measure	27.2		34.3	

TABLE II
THE BREAKDOWN IN PROPOSED SYSTEM FOR OUTPUTS OF CONVENTIONAL SYSTEM

Filtered sentences		Not filtered sentences		total
incorrect	correct	incorrect	correct	
263	13	384	104	764

The precision and recall are defined as follows:

$$Precision = \frac{\#\ of\ correct\ sentences\ which\ system\ returned}{\#\ of\ sentences\ which\ system\ returned}$$

$$Recall = \frac{\#\ of\ correct\ sentences\ which\ system\ returned}{\#\ of\ correct\ sentences}$$

The subjects are five graduate students. To motivate the subjects to make queries, we gave questions about English blank supplementary and English composition to the subjects. The subjects used the system in order to accomplish these questions.

B. Experimental Results

Table I shows the precisions, the recalls and the F-measures of the conventional system and the proposal system. Although the decreasing rate of recall was only $11.1\%((100.0 - 88.9)/100.0)$, the rate of improvement of precision reached to $38.0\%((21.3 - 15.8)/15.8)$. As a result, F-measure has improved 7.1 point.

Table II shows the breakdown in the proposal system for 764 sentences displayed by the conventional system. There existed 263 sentences which were not actually a correct answer, and these sentences occupied 95.3% of 276 filtered sentences. On the other hand, these 263 sentences were filtered correctly and occupy 40.6% of 647 sentences which were not correct answers in the conventional system. Thus, the effect of filtering was confirmed.

V. CONCLUSION

This paper has presented a cross-lingual English sentence retrieval system considering syntactic structures of Japanese queries. To achieve the retrieval of English sentences corresponding user's intention, the system filters out English sentences which do not correspond with Japanese by using dependency structures of Japanese and English. In the search experiment, the correct answer data were constructed by five subjects. We compared the performance of the system by the presence of filtering using these data. As a result, the retrieval precision improved 5.5% and F-measure improved 7.1% by filtering. Thus, the effectiveness of filtering was confirmed. In addition, as a result of evaluation of filtering, the precision was 95.3%.

The system uses all of the corresponding English words written in Japanese-English dictionary for Japanese queries. Therefore, the system displays much sentences including the words with wrong translation. In future work, it is necessary to reduce the mistakes of the translation.

REFERENCES

[1] Y. Kato, S. Egawa, S. Matsubara, and Y. Inagaki: English Sentence Retrieval System Based on Dependency Structure and its Evaluation, *Proceedings of IEEE 3rd International Conference on Information Digital Management (ICDIM-2008)*, pp.279-285 (2008).

[2] T. Briscoe, J. Carroll, and R. Watson: The Second Release of The RASP System, *Proceedings of 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-2006) Interactive Presentations Sessions*, pp.77-80 (2006).

[3] T. Watanabe, E. R. Skrzypczak, P. Snowden: Kenkyusha's New Japanese-English Dictionary 5th Edition, *Kenkyusha*, (2004).