DEVELOPING AN ONLINE CORPUS AND TEACHING MATERIALS FOR WRITING TECHNICAL PAPERS IN ENGLISH

S. Matsuda[*, a], T. Beppu[a], S. Takahashi[a], H. Inaba[a], T. Horiuchi[a], T. Hashimoto[a], M. Hattori[a], M. Omori[a], M. Higa[a] and M. Takao[a]

[a] National Institute of Technology, Matsue College, Matsue, Japan

*matsuda@matsue-ct.jp

Abstract

In order to nourish engineers who can work internationally, what kinds of competences are required? From my experiences in engineering fields, they would be skills to read and write technical papers in English and to present their researches in international conferences. Then, I wonder if our curriculum is enough to let our students acquire such skills. Perhaps, some extra support would be necessary for them. As the first reason for that, our school is a college of technology and cannot provide more English classes than usual high schools. Second, there are differences in vocabulary, expressions and constructions between general English and English for Science and Technology (EST). Even the seventh graders in our school have difficulties writing the abstract section in a technical paper. Furthermore, in ordinary dictionaries, we hardly find example sentences for scientific verbs such as "converge" and "quantify." For technical writing, we need to learn the differences mentioned above and to get used to many expressions in science. In particular, beginners should write English sentences by replacing some parts in them with words they want to use. In this study, we attempt to develop an online system to support beginners' technical writing based on an EST corpus. Users of our system can find example sentences by choosing verbs in Japanese. This means that Japanese translation must be added to each of the example sentences. We are currently collecting more English sentences from scientific journals and translating them into Japanese. In the translation process, we have introduced the methodology called "Semantic Order Filing System (SOFS)." The SOFS classifies words/phrases in a sentence into seven categories called "Semantic Order Folder." For example, the first and second folders correspond to the subject and predicate verb of a sentence. That is, each folder represents a semantic function. The SOFS would help beginners understand the word order specific for English sentences.

Keywords: *technical writing, corpus, English for science and technology, semantic order filing system, college of technology*

Introduction

Recent years Japanese students at colleges of technology have had more opportunities to make a presentation in international conferences. This means that they are required to write an English abstract in 2-3 pages. However, it is truly a troublesome job for teachers to correct their abstracts. English sentences they write often do not have any subject. For example, if they translate a Japanese sentence "測定システムを図 *X* に示す。", their English translation may be "Show the measurement system in Fig. *X*." A correct translation is "Figure *X* shows the measurement system." The error may be the influence of their mother tongue. In Japanese language, subjects tend to be omitted in the case where they are obvious. Not only native English teachers but also Japanese teachers have difficulties in correcting their erroneous sentences. In this work, we attempt to develop a corpus-based support system for technical English writing, which makes it easier for native English teachers to correct English sentences Japanese students wrote.

Materials and Methods

The most important thing to construct a corpus-based system is to collect quality English sentences. For this purpose, we made use of a corpus of English for Science and Technology, the PERC Corpus (PERC, 2008). We can access the Web site for free. All the sentences in the PERC Corpus were collected from academic journals in 22 scientific fields. Because our system targets students at colleges of technology, we selected as a sub-corpus six fields: civil engineering, computer science, construction and building technology,

Table 1 An Example of Japanese Translation Given in Our Corpus-based System

Semantic function	who	does/is	whom/what	where	when	how	why
English sentence	We	propose	a new method	in this paper.	—	—	—
Japanese translation	我々は	提案する	新しい手法	本論文で	—	—	—

electrical and electronic engineering, engineering and telecommunications. The English sentences we obtained are used for the purpose of academic research only.

The process of collecting English sentences consists of three steps. First, we extracted 105 Japanese verbs from Japanese abstracts, which the seventh graders in our school had written for their dissertation. Second, we translated the Japanese verbs into English ones. After that, other teachers in our school examined both the Japanese and English verbs, and added more verbs that were considered indispensable for writing technical papers. As a result, we obtained 200 English verbs with no duplication. These verbs are used as base form in collecting English sentences by the POS search in the PERC Corpus, which enables us to specify the "part of speech" of a searched word.

Our corpus-based system allows the users to search English sentences by choosing Japanese verbs. Also, they can see Japanese translation attached to English sentences obtained. In other words, our system is a parallel corpus between Japanese and English.

However, the Japanese translation is shown in a different manner from ordinary Japanese sentences. It is actually shown in a semantic order. Table 1 (see the previous page) shows an example of Japanese translation given in our system. Tajino (2011) named such a way to show an English sentence and its Japanese translation "Semantic Order Filing System (SOFS)." The SOFS would implicitly make the users aware of the word order specific for English sentences. In Japan, most of English teachers teach their students the constructions of English sentences by means of five elements: subject (S), verb (V), object (O), complement (C) and modifier (M). But this way of teaching seems not to be effective for those who are really poor at English. Therefore, we have introduced the SOFS as an alternative way.

We intend the system to be used online. Thus, it has been programmed with "Ruby on Rails." The programming language, Ruby is an object-oriented language developed by Yukihiro Matsumoto and is suitable for making Web applications. Figure 1 shows a snapshot of our system. It was named "the Corpus-based support System for Technical English Writing (C-STEW)."

The users first choose a Japanese verb from the upper pull-down menu. Next, they choose an agent for the verb in the lower pull-down menu. After a pair of verb and agent was chosen, an English sentence including those words is shown in the field below. Predicate verb is in red both in English and Japanese sentences. The future version of the system will show a section tag indicating which section of an article the sentence appeared in. Furthermore, a supplementary explanation on the verb usage might be added to each sentence. Note that a series of procedures shows only one English sentence. This is not preferable for learning how to write technical English. Therefore, if the users do not choose any agent, they can see all the sentences with the verb they chose.

Results and Discussion

We tried to use the system by ourselves and it works well as a prototype of support system for technical writing. But we have found two problems. One is that there are too many verbs and agents in each pull-down menu. In the present version, each verb has only five example sentences. However, as the number of sentences increases, it becomes more difficult to find the usage of a verb which the users want to know. The other is that one Japanese verb does not always

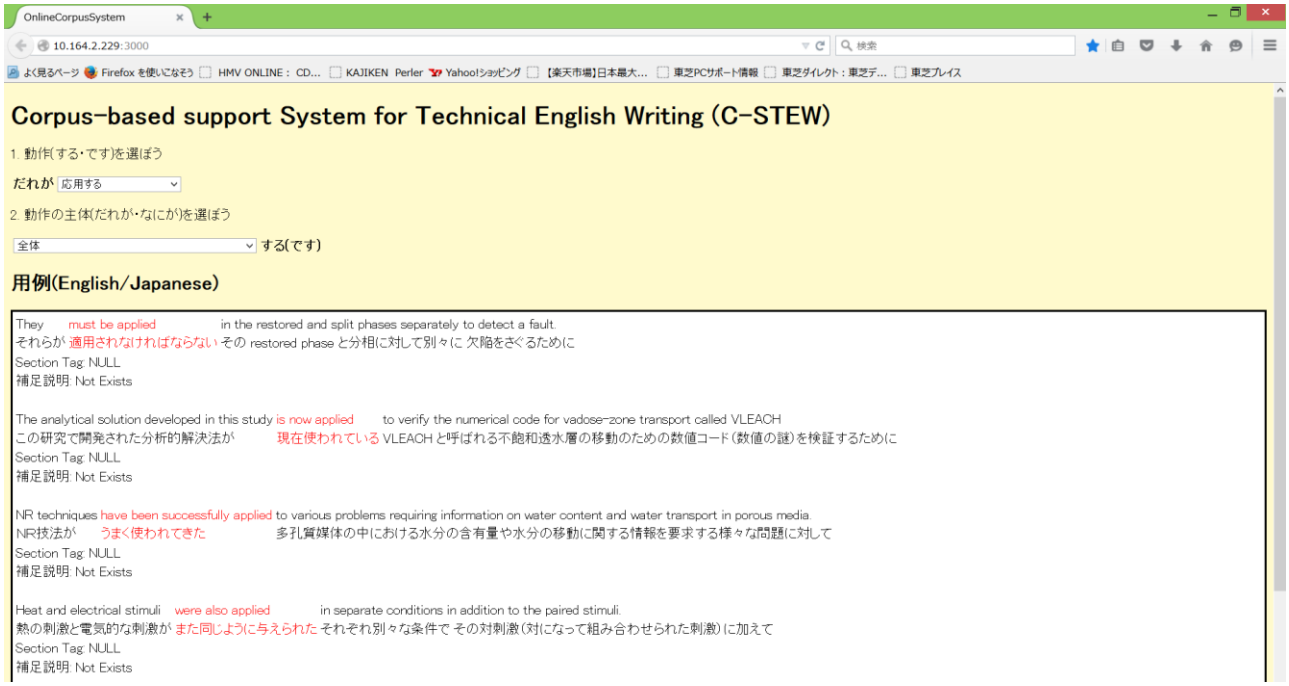

Fig. 1 A snapshot of our corpus-based system.

correspond to one English verb. That is, the relationship between Japanese and English verbs is not one-to-one correspondence. In order to solve the two problems, some method for grouping the verbs and agents would be required.

As a first step for grouping, we attempted to analyze the collocational relationship between words in the Japanese verbs and agents. For this analysis, we employed free software for text mining, "KH Coder" (Higuchi, 2014). This is because the software includes a morphological analyzer for Japanese language, "ChaSen," which was developed by Nara Institute of Science and Technology. In addition, KH Coder provides us with a large variety of statistical tools. Co-occurrence network is one of such tools and enables us to know which word pair frequently occurs in the same sentence.

Figure 2 shows a co-occurrence network of words in the verbs and agents. Each node represents a word and an edge is drawn between two words if they co-occur frequently. We evaluated the co-occurrence frequency by Jaccard coefficient. This is defined as follows:

$$J(A,B) = \frac{P(A \cap B)}{P(A \cup B)}, \quad (1)$$

where $P(A \cap B)$ is the co-occurrence probability of words A and B in the same sentence. $P(A \cup B)$ is the occurrence probability of word A or B. The value of the coefficient ranges from 0 to 1. As the value increases, the intensity of co-occurrence becomes larger. Actually, the thickness of each edge corresponds to this intensity. For all the edges in Fig. 2, Jaccard coefficient is greater than 0.2.

The color of each node represents the betweenness centrality. The degree of the centrality is determined by whether a node is located in the shortest path between two nodes except that node. Pink color indicates the highest centrality. White and aqua follow the pink in this order.

In the upper right of Fig. 2, we can see a bright pink node surrounded by eight nodes. The node corresponds to a Japanese verb, 行う(conduct). In Japanese language, this verb is often combined with a noun implying some behavior. For instance, four nodes around the pink one show 分解(decomposition), 実験(experiment), 測定(measurement) and 抽出(extraction), respectively and they mean actions taken in a laboratory.

A hub node like 行う(conduct) seems to give a key for grouping the verbs and the agents in each pull-down menu. Namely, the users first choose an action from categories such as (a) do it inside a laboratory and (b) do it outside a laboratory. Then, they choose an agent of

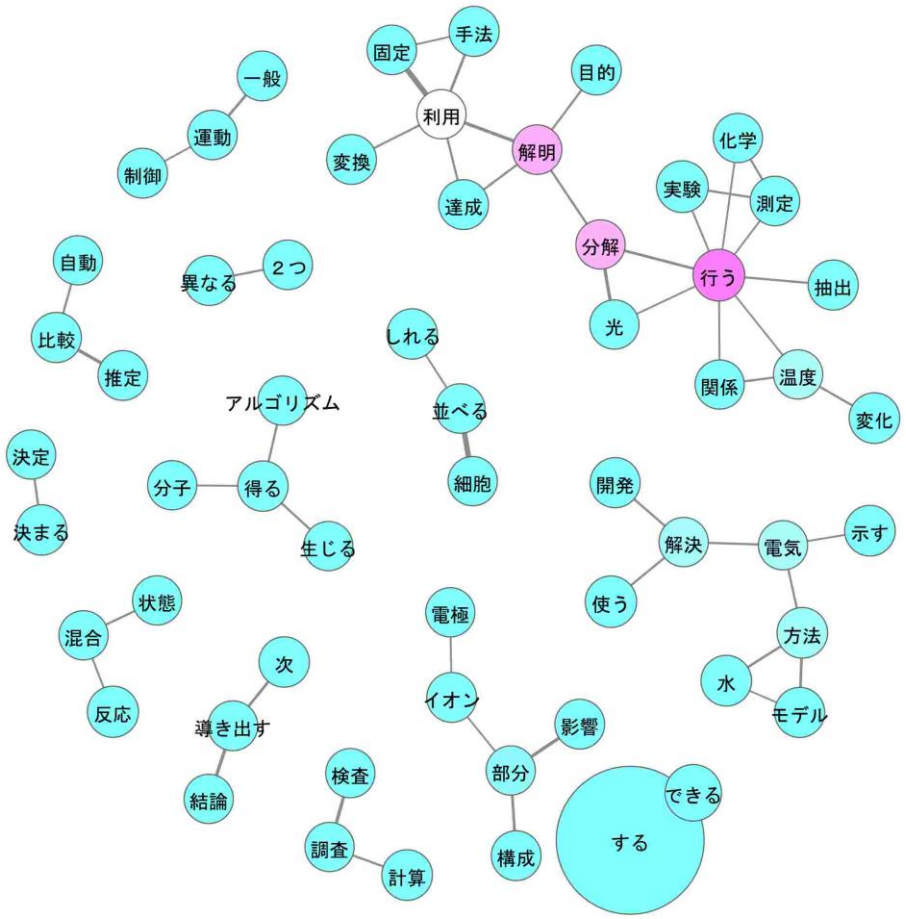

Fig. 2 A co-occurrence network of words in the Japanese verbs and agents.

the action from categories such as (c) human and (d) machine. However, there is a problem in making such categories. We cannot know which part of speech, verb or noun, each word in the network is attributed to. For example, 測定(measurement) is definitely a noun but this could be separated from a verb, 測定する(measure) by the Japanese morphological analyzer. In other words, lots of Japanese nouns can be verbified by being followed by a verb, する(do).

Considering the above findings, co-occurrence network should be analyzed for English words, not Japanese words. Then, we should try to categorize the words and give each category a title in Japanese. This would make it easier for the users to choose a pair of verb and agent suitable for what they want to express.

Conclusions

We have developed a prototype of support system for technical English writing. Throughout the development, it was clarified that categorizing the verbs and the agents of example sentences was necessary and this could be attained by analyzing the co-occurrence network of English words in our corpus. At the same time, we need to translate more English sentences into Japanese using the SOFS. Furthermore, we are going to have the sixth graders use our corpus-based system to write the summary of their research plans in English.

Acknowledgements

We thank Mr. Kazuhiro Fujihara for programming the corpus-based system in Ruby. This work has been supported by the project to improve KOSEN students' English skills promoted by National Institute of Technology.

References

Higuchi, K. (2014). *Quantitative text analysis for social survey—aiming at the inheritance and development of content analysis*. Kyoto: Nakanishiya Publishing.

PERC (2008). *The Professional English Research Consortium Corpus*. Retrieved from https://scn.jkn21.com/~percinfo/index.html.

Tajino, A. (2011). *"Semantic order," English learning method for those who do not know where they should start again*. Tokyo: Discover 21.