APPLICATION OF WORD ALIGNMENT FOR SUPPORTING ENGLISH TRANSLATION OF JAPANESE STATUTES

TOYAMA Katsuhiko, IMAI Kazuhiro, OGAWA Yasuhiro

Department of Information Engineering, Graduate School of Information Science, Nagoya University

ABSTRACT

Legal information is one of the fundamental information in society, and a statute database serves as a social information infrastructure. Recently, there has been increasing social demands for compiling a database which consists of not only Japanese statutes but also their English translations. In this paper, we describe the problem of translating Japanese statues and show how to solve it by utilizing technologies developed for natural language processing. In particular, we show how to support both the compilation of a standard bilingual dictionary and the unification of translation equivalents of legal technical terms in compliance with the dictionary by using word alignment.

1. INTRODUCTION

Legal information is one of the fundamental information in society, and a statute database serves as a social information infrastructure. Recently, there has been increasing social demands for compiling a database which consists of not only Japanese statutes (acts, cabinet orders, ordinances of ministries, etc.) but also their English translations. This demand has arisen from various motivations related to social and economic globalization to conduct international transactions more smoothly and to promote more international investment in Japan. Another major motivation is the desire to provide technical assistance to legal reform in developing countries and former socialist countries by the Japanese government.

In this paper, we describe the problem of how to translate Japanese statutes and show how to solve it by utilizing technologies developed for natural language processing. In particular, we show that the word alignment technique for automatic extraction of bilingual lexicons can be applied to support both the compilation of a standard Japanese-English bilingual dictionary and the unification of translation equivalents for legal technical terms.

2. PROBLEM IN TRANSLATION OF JAPANESE STATUTES

Various kinds of problems arise in translating Japanese statutes into English [10]. Among them, unifying the translation equivalents of legal techinical terms in statutes is a fundamental problem. Since translations of Japanese statutes have been made so far individually, fragmentarily, sometimes privately, and manually by government ministries and agencies with jurisdictions, affiliated organizations, and private publishing companies of law books, several kinds of translation equivalents for the same term in the same field of the legal domain may be used, which can cause misunderstanding. For example, as a translation equivalent of the Japanese legal term "善意の (*zen'i no*)", we can find at least "*without notice*", "*without knowledge*", "*innocent*", "*in good faith*", and "*bona fide*" in various translations; here, the meanings are clearly not the same, and in fact some of them are incorrect. Although it is desirable to make a one-to-one correspondence between the terms of two languages as much as possible, in practice the meanings of matched terms are not always exactly the same, and the delicate differences in meaning are not always clear. Furthermore, there may be cases where translation equivalents have to be properly selected according to the context. However, the criteria for proper selection are also not always clear.

One solution to overcome this problem is to determine *standard translation equivalents* and to compile *a standard bilingual dictionary* for legal technical terms in statutes, where the dictionary is open to the public and strongly promoted to translators and lawyers. The differences in meanings among several translation equivalents and the criteria for the proper selection of them should also be described by some kind of notation in the dictionary. This solution will partly contribute to maintaining the quality of translations.

However, since it is widely recognized that the compilation of such a dictionary would be very expensive, we must find a method to accomplish this as efficiently as possible, even if the final decisions in editing dictionary lexicons are made by human experts on the legal domain. In the next section, we show how to automatically extract bilingual lexicons.

3. COMPILATION OF STANDARD BILINGUAL DICTIONARY FOR JAPANESE STATUTES

In this section, we show how to utilize the word alignment technique to support the compilation of a standard Japanese-English bilingual dictionary for Japanese statutes. This key task is clarified by briefly describing the progress made in the

compilation.

3.1. Construction of Parallel Corpus

As mentioned in the previous section, translations of Japanese statutes have been made so far. Therefore, We first collected existing English translations and their Japanese source texts. Then, we manually constructed a parallel corpus, where corresponding Japanese and English sentences are aligned.

Usually, it is a troublesome problem to determine which English sentences correspond to which Japanese ones in constructing a parallel corpus, since we have to consider the cases where one source sentence may correspond to several target sentences and vice versa. However, we noted that there are units such as articles and paragraphs in statutes, and these were maintained even in the translated statutes. Accordingly, we made correspondences by utilizing these structural units in order to reduce the problem.

As a result, we constructed a parallel corpus that consists of 153 statutes (39,560 Japanese-English sentences),

3.2. Automatic Extraction of Bilingual Lexicons

There have been several studies on automatic extraction of bilingual lexicons from a parallel corpus, and several kinds of methods have been proposed for calculating similarity between terms in two languages [6]. Among them, we adopted *the Dice coefficient* as a measure of similarity since it is shown to be more effective than mutual information [5] information [5] and it is simple. The coefficient is described as follows:

$$Dice(x, y) = \frac{2 \cdot freq(x, y)}{freq(x) + freq(y)} \quad (0 \leq Dice(x, y) \leq 1),$$

where $freq(x)$ and $freq(y)$ denote the numbers of occurences of a term x in the source sentences and a term y in the target ones, respectively, and $freq(x, y)$ denotes the number of cooccurences of x and y in the aligned sentences.

When using the Dice coefficient, only the term with the maximum similarity value is usually extracted as an equivalent y to a given term x. In this task, however, we extracted all terms whose similarity values are in top three and more than 0.7, since the coverage for extracting bilingual terms is considered important.

Furthermore, in this task, we need to extract not only English translation equivalents y to a given Japanese entry x but also Japanese entries themselves. Therefore, we used character N-grams including *hiragana* characters in Japanese sentences, since the words in them are not separated by spaces, while we used word N-grams in English sentences, and we set $N \leq 25$. That is, for each sequence of N characters in Japanese sentences, we calculated the Dice coefficient for each sequence of N words in the English sentence.

As a result, we succeeded in extracting not only bilingual words such as nouns and verbs but also bilingual wordings such as "なお従前の例による (*nao juzen no rei ni yoru*) / *it shall be handled as heretofore*" and "のいずれかに該当する (*no izureka ni gaito suru*) / *fall under any of*". This result is considered effective for supporting translation of statutes, since such "boiler-plate" expressions are so often used in statute sentences.

3.3. Selection of Bilingual Lexicons by Legal Experts

The automatically extracted bilingual lexicons may include inappropriate ones in morphological and syntactical sense, since they were automatically cut off from the texts in the corpus. Also, the correspondence of the meanings of the bilingual lexicons may be incorrect.

Therefore, human experts in the legal domain examined the automatically extracted lexicons twice. In the first step, they deleted or edited the bilingual lexicons that were invalid as expressions in each language and that were unnatural as translations. In the second step, legal experts selected appropriate bilingual lexicons for registration in the dictionary from the viewpoint of standardness.

3.4. Compilation of Dictionary Supported by Bilingual KWIC

After then, legal experts examined each bilingual lexicon again by referring to the parallel corpus of statutes, and they attached to them criteria of usage when it was necessary to make a proper selection from among multiple translation equivalents according to the context. Example sentences in which the lexicons appear and other comments from the legal viewpoint were also attached at the same time if necessary. To support this task, we developed a GUI tool *Bilingual KWIC* [8], which is illustrated in Figure 1.

Given a parallel corpus, Bilingual KWIC not only automatically extracts bilingual lexicons by using the word alignment technique described in Section 3.2 but also displays them within their contexts, i.e., in KeyWord In Context (KWIC) form. Concretely put, if a term in the source language is input in the keyword field, every source sentence that includes the term is retrieved and displayed in the left pane, where every occurrence of the term in the source sentences is colored. At the same time, in the right pane, their corresponding target sentences are displayed on the same lines as their source sentences, where occurrences of automatically calculated translation equivalents to the input term are centered and colored in blue. The bilingual sentences can be sorted just before or after the input term and the calculated equivalents. Therefore, users can easily select appropriate bilingual lexicons by comparing several calculated equivalents and referring to their contexts in the sentences. Users can also easily correct the errors made by the automatic extraction, find derivational patterns of bilingual lexicons, and acquire other contextual information for usage.

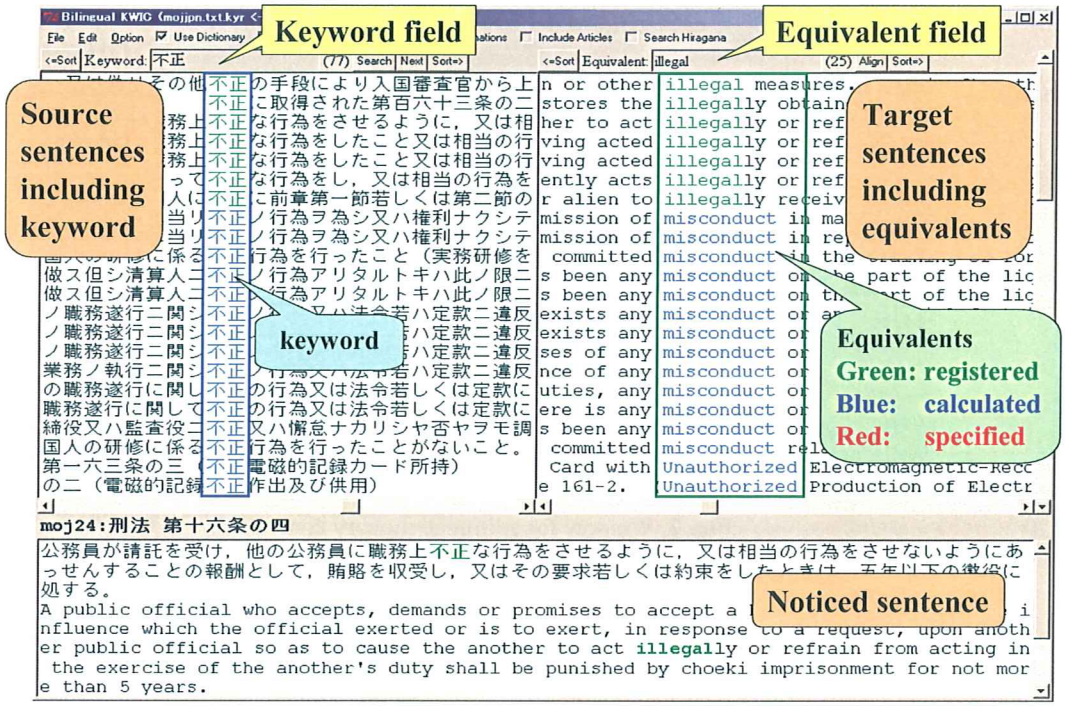

Fig. 1. Overview of Bilingual KWIC.

Bilingual KWIC does not need any bilingual dictionary, since it automatically calculates translation equivalents. However, it is possible to embed a bilingual dictionary in it. In this case, the target sentences containing the equivalents registered in the dictionary are retrieved and displayed first, and occurrences of the equivalents are displayed in green. Then, other translation equivalents are calculated and displayed in blue. Furthermore, if the users believe that both registered and calculated equivalents are inappropriate, they can specify another equivalent by inputting it in the equivalent field so that target sentences including it are shown, where the occurrences of the specified equivalent are displayed in red.

For example, as illustrated in Figure 1, Japanese sentences in the corpus that include the Japanese legal term "不正 (*fusei*)" are displayed in the left pane, where occurences of this term are centered and colored in blue or green. On the other hand, in the right pane, corresponding English sentences including the equivalent "*illegal*" registered in the dictionary are displayed first, where occurences of this equivalent are centered and colored in green. Below them in the pane, other English sentences are displayed, where automatically calculated equivalents "*misconduct*" and "*unauthorized*" are centered and displayed in blue. These equivalents are sorted according to their values of the Dice coefficients.

In fact, when compiling the standard dictionary by using Bilingual KWIC, we used a temporal bilingual dictionary to make the task efficient, where the bilingual lexicons that have the maximum number of occurrences of English equivalents in the corpus for each Japanese entry were registered.

Bilingual KWIC also has a function to support the registration of dictionary contents. Users can open another window illustrated in Figure 2 when the cursor is on a sentence and then fill out or edit such fields as for entry in the source language, part of speech and pronunciation, equivalent in the target language, usage for proper selection, example bilingual sentences, and other comments. The window also indicates other equivalents for the same entry if they exist.

3.5. Further Compilation of Bilingual Dictionary

After the step of compiling the dictionary by using Bilingual KWIC, 2,246 Japanese entries and 3,329 English translation equivalents were included, where each Japanese entry had 1.5 English equivalents on average. However, this number of Japanese entries is not necessarily sufficient as a standard bilingual dictionary.

Two reasons are assumed to account for this situation. As mentioned in subsection 3.2, automatically extracted bilingual lexicons were those whose similarity values exceed the threshold, and Japanese entries were selected strictly from those bilingual lexicons. Since making the threshold lower may necessitate much computation and much manual selection to avoid meaningless bilingual lexicons, as a method to overcome this, we can assign as many Japanese entries as possible before the calculation by collecting them from existing Japanese technical glossaries in the legal domain.

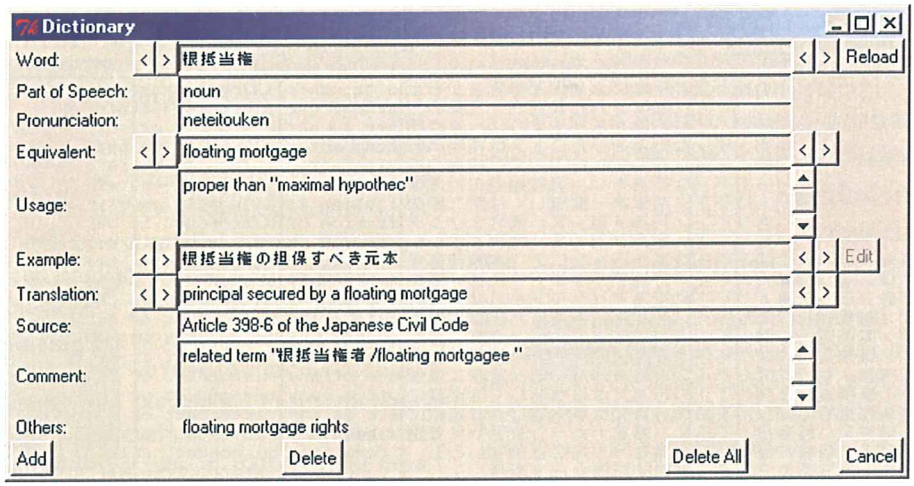

Fig. 2. Window for editing dictionary contents.

The other reason is deviations in terms within the fields of the legal domain in the parallel corpus. Although there are about 1,800 acts and 5,500 orders and ordinances that are currently effective in Japan, the number of the statutes that constitute the corpus is only 153, much less than the whole. Therefore, there are some legal terms that do not appear in the corpus even if they are important and often used.

Furthermore, "boiler-plate" expressions that include other terms in them such as "この法律は、... から施行する (*kono horitsu wa ... kara shikosuru*) / *This Act shall come into force as from ...*" essentially could not be extracted in this simple method even if they often appeared in the statutes.

Then, as the fourth step, legal experts added the missing Japanese entries to the dictionary by using existing glossaries and Bilingual KWIC.

After making some further corrections, the dictionary finally came to include 3,315 Japanese entries and 3,974 English translation equivalents, where the average number of translation equivalents is 1.2. This dictionary's first version was released in April 2006 [2, 3].

4. SUPPORT TOOL FOR UNIFYING TRANSLATIONS

Once the standard bilingual dictionary for statutes is compiled, use of it is strongly promoted when making English translations of Japanese statutes to unify translation equivalents. Therefore, it is desirable to support translators when selecting translation equivalents in compliance with the dictionary.

Here, it is noted that some statutes have already been translated into English as mentioned in Section 3. These should be reused and revised to be in compliance with the standard dictionary, since the cost required to correct different translation equivalents from the dictionary is smaller than that of retranslating the whole statute.

In this section, we describe a support tool for checking dictionary compliance of a given bilingual text [4]. This tool was developed by also applying the word alignment technique.

4.1. Design of Support Tool

This tool checks whether the translation is in compliance with the standard bilingual dictionary when aligned bilingual texts are given and supports translators to replace the inappropriate translation equivalents with the standard ones.

There are two key points in designing such a tool. One is how to find translation equivalents in target sentences that are not only in compliance with the dictionary but also those that are not. The other key point concerns the user interface, where translators can easily recognize the places that are not in compliance with the dictionary and correct them. Of course, a fast method to search the dictionary is also required, so we utilized TRIE [1] as a data structure of the dictionary in the tool.

The output of the tool is illustrated in Figure 3, where each Japanese source sentence and its English translation are displayed one after the other. The tool utilizes not only the standard dictionary but also *the inappropriate equivalents list*, which is mentioned in Section 4.3. The tool processes each Japanese-English bilingual sentence as follows:

1. In the Japanese sentence, find Japanese entries that are registered in the standard dictionary by using longest-first string matching.

2. If standard translation equivalents to the Japanese entries found in step 1 appear in the corresponding English sentence, display both the Japanese entries and their equivalents in boldface letters (Figure 3 (a)).

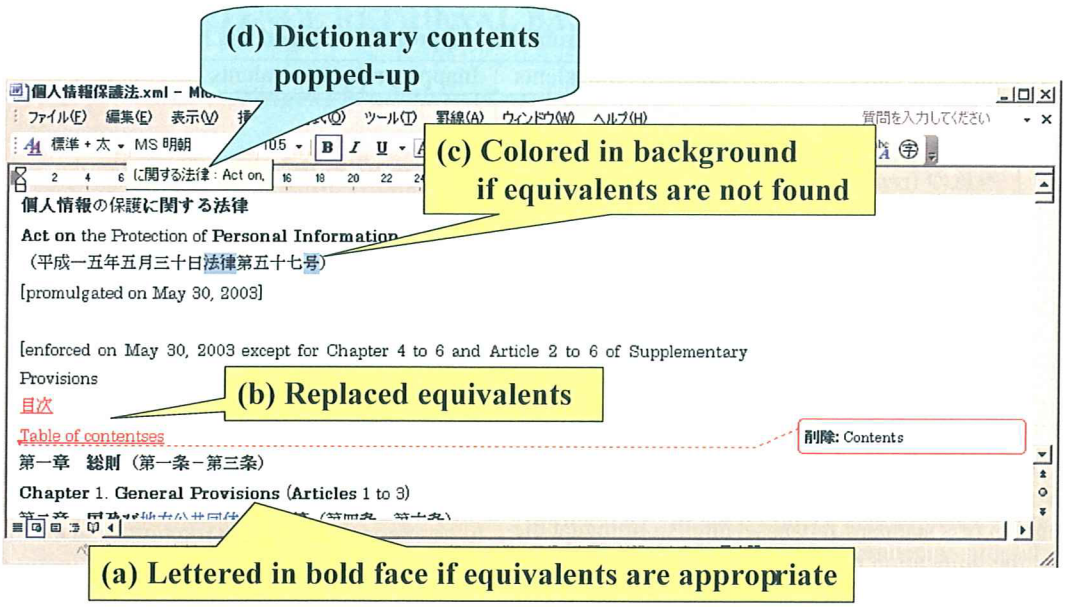

Fig. 3. Overview of support tool for unifying translation equivalents.

3. Otherwise, find translation equivalents to the Japanese entries in the English sentence by retrieving the inappropriate equivalents list. If found, replace them with the standard ones by using the dictionary (Figure 3 (b)).

4. If translation equivalents to the Japanese entries cannot be found at all, display the entries in the blue background (Figure 3 (c)).

In steps 2 and 3, word matching is executed by considering variants such as plural forms of nouns and third-person singular present forms of verbs. Moreover, if the cursor is placed on Japanese entries or English equivalents registered in the dictionary in each sentence, the contents of the dictionary related to them are displayed in a pop-up window (Figure 3 (d)).

4.2. Output by Utilizing wordML

We utilized MS-Word for the output interface of the tool, since we can assume that the users of the tool are used to this software and it has a function for supporting the correction of documents, allowing us to drastically reduce the cost of developing the interface. In MS-Word 2003, documents can be described in wordML [7], a scheme language based on XML. To realize the display of bold face letters or colored background, suitable values may be given as attributes of some structure element in wordML. Replacement of the translated terms can also be done by using attributes of mark-up tags to delete and insert characters in the sentences. Since wordML also has a function to display footnotes in pop-up windows, the contents of the dictionary are described as if they were footnotes attached to the suitable terms in the sentences.

4.3. Inappropriate Equivalents List

If translated equivalents in the translations are different from the standard ones and if we cannot recognize where they are in the translation, we cannot easily replace them with the standard ones. To effectively find the inappropriate equivalents, we developed *the inappropriate equivalents list*, which consists of bilingual lexicons whose English equivalents are either non-standard or incorrect.

The list is compiled for each Japanese entry in the standard dictionary beforehand by also utilizing the word alignment technique. That is, by calculating the Dice coefficients in the bilingual corpus as well as the method of Section 3 and removing standard English equivalents from the automatically extracted bilingual lexicons, inappropriate equivalents can be acquired.

In fact, we compiled the list by using 10 acts and their English translations (4,594 bilingual sentences) provided by the ministries, where we extracted only the bilingual lexicons whose values of the Dice coefficient were more than 0.8 and whose Japanese entries occur more than 18 times in the sentences. These values were determined according to the results of preliminary experiments. As a result, 152 English terms were automatically extracted as candidates of inappropriate translation equivalents to 86 of 1,579 Japanese entries appearing in the acts, where only 37 terms were strictly inappropriate. Although this result is not necessarily sufficient and it remains a problem how to efficiently make the list richer, we left this task to future work and manually edited the list so that it included 116 English terms as inappropriate equivalents to 79 Japanese entries. Some examples from the list are shown

Table 1. Examples from the inappropriate equivalents list.

Japanese entry	Standard equivalents	Inappropriate equivalents
法律 (*horitsu*)	act, code	law
条 (*jo*)	article	section
善意の (*zen'i no*)	without knowledge	without notice, innocent, in good faith, bona fide
個人情報 (*kojin joho*)	personal information	individual information

in Table 1.

5. CONCLUSION

We showed that the word alignment technique for automatic extraction of bilingual lexicons can be applied to support both the compilation of a standard Japanese-English bilingual dictionary and the unification of translation equivalents of legal terms in compliance with the dictionary. We have succeeded in compiling the first version of the dictionary and releasing it on web sites [2, 3] in only about one year. Several translations of major Japanese statutes have also been released on the web sites, and the number of them are increasing, where almost all of them were checked to unify the equivalents by using the tool we developed.

These results are supplied to the project of the Japanese government to establish an infrastructure for promoting translation of Japanese statutes into foreign languages [9].

The next task is to update the standard bilingual dictionary, which has already started. We are going to automatically extract bilingual lexicons not only from other bilingual corpora but also from the translations in compliance with the first version of the dictionary, translating them again so that the dictionary is compiled in a spiral manner.

Acknowledgements

The authors would like to thank Professor MATSUURA Yoshiharu, Associate Professors KAKUTA Tokuyasu and Frank BENNETT, and Mr. SANO Tomoya, Graduate School of Law, Nagoya University, for their discussions, suggestions, and helps.

6. REFERENCES

[1] Aoe, J., : Key Search Strategies –Trie and Its Applications–, *IPSJ Magazine*, **34** (2), 244–251, 1993.

[2] Cabinet Secretariat: Translations of Japanese Laws and Regulations, `http://www.cas.go.jp/jp/seisaku/hourei/data1.html`.

[3] English Translation Project of Japanese Statutes: `http://www.kl.i.is.nagoya-u.ac.jp/told/`.

[4] Imai, K., Ogawa, Y., Toyama, K.: A Support System for Unifying Terms used in Japanese-English Translation, *Proc. 5th Forum on Information Technology*, 215-218, 2006.

[5] Kitamura, M., Matsumoto, Y.: Automatic Extraction of Translation Patterns in Paralell Corpus, *IPSJ Magazine*, **38** (4), 727–736, 1997.

[6] Matsumoto, Y., Utusro, T.: Lexical Knowledge Acquisition, Dale, R., Moisl, H., Somers, H. (eds.), *Handbook of Natural Language Processing*, 563–610, Marcel Dekker, 2000.

[7] Microsoft Corporation: The XML Files: XML in Microsoft Office Word 2003, `http://msdn.microsoft.com/library/default.asp?url=/library/en-us/odc_2003_ta/html/odc_ancword.asp`, 2006.

[8] Ogawa, Y., Toyama, K.: Bilingual KWIC – GUI Support Tool for Bilingual Dictionary Compilation, *Proc. 6th Symp. of Natural Language Processing*, **2**, 77–84, 2005.

[9] Study Council for Promoting Translation of Japanese Laws and Regulations into Foreign Languages: Final Report, `http://www.cas.go.jp/jp/seisaku/hourei/report.pdf`, 2006.

[10] Toyama, K., Ogawa, Y., Matsuura, Y.: A Design of the Translation System for Japanese Statutes, *Jurist*, 1281, 2–5, 2004.