Initiatives of the Japan Patent Office

on Machine Translation

Kei Kato

Patent Information Policy Planning Office, Policy Planning and Coordination Department, Japan Patent Office

3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan

kato-kei@jpo.go.jp

Abstract

Recently, each Intellectual Property (IP) Office has faced common big challenges: promoting the international work sharing in the examination process and developing an environment to access foreign patent documents written in languages other than its native language. As a means of resolving such challenges, the Japan Patent Office (JPO) has actively utilized machine translation. Currently, the JPO has widely provided the foreign general users with the Japanese-English machine translated information on its patent examination results through the "One Portal Dossier" allowing them to retrieve dossier information for applications filed with the IP5 Offices (Japan, the U.S., Europe, China, and the Republic of Korea). Also, the JPO launched a new system in January 2015, utilizing the Chinese-Japanese machine translation dictionary including more than 2 million words. This system enables users to search in the Japanese language more than 12 million machine translated patent and utility model documents of the Chinese and Korean languages.

1. Introduction

The number of patent applications filed worldwide greatly increased to 2.57 million in 2013, compared with 1.58 million in 2004. Especially, the number of patent applications filed in China has dramatically increased and grew to 32.1% of the total applications in 2013 from 8.3% in 2004. From a viewpoint of searching patent information, this situation shows that the needs for searching foreign documents including those of the Chinese languages have increased. That is, as the examiners judge the novelty of inventions, inventive step, etc. based on the result of prior art search in the patent examination, it is necessary to search prior art documents including these foreign patent documents precisely as well as efficiently in order to grant a stable right recognized in the world. Actually, as the rate at which patent documents of the Chinese languages are cited as prior art in the examination has gradually accelerated at each Intellectual Property (IP) Office, it has become a challenge common to all IP Offices to improve an access environment such as understanding and searching foreign patent documents written in languages other than its native language.

On the other hand, it is not easy to understand and to exhaustively search foreign patent documents around the world. Needless to say, while translation enables us to understand them in our native languages, translating quite a large number of documents only by human has some limitations in terms of costs and resources, as the number of patent applications filed worldwide has increased year by year. Under this situation, the expectation of im-

provement of the enviorment for searching foreign documents by utilizing machine translation is getting higher.

Furthermore, it is hoped that machine translation will be utilized in order to disseminate patent information to other IP Offices. Currently, while intellectual property related activities are getting globalized, applications for a patent of one invention are more often filed in several countries or regions of the world. In case applications for a patent of the same invention are filed in several countries or regions, examiners in each IP Office need to share examination information (dossier information) owned by each IP Office in order to eliminate inefficiencies of conducting duplicate examinations or searches. But if each IP Office provides its examination information in its native language, it is difficult for other IP Offices to utilize such examination information. In order to promote the international work sharing, it is important to disseminate not only information in its native language but also one translated into other languages such as English.

In view of this situation where attention is getting focused on utilizing machine translation in the area of patent information, the Japan Patent Office (JPO) has taken measures to actually utilize machine translation. In this paper, I will introduce a search system for the Chinese and Korean documents utilizing machine translation and a service to provide examination information which the JPO has provided.

2. Improvement of the Environment for Searching Patent Documents of the Chinese and Korean languages

2.1. The Chinese and Korean Gazette Translation and Search System (CKGS)

As shown in Figure 1, in 1998 the documents to be understood and searched in the Japanese language accounted for 55% of the total patent documents issued in the world. But recently as the number of patent documents of the Chinese language has rapidly increased, documents to be understood only in the Chinese and Korean languages have accounted for 65% of the total. In January 2015, the JPO launched a new service of the "Chinese and Korean Gazette Translation and Search System" (CKGS) which enables users to search in the Japanese language patent and utility model documents of the Chinese and Korean languages in order to enhance the convenience of the environment for searching them.

A rule-based translation method is adopted for this system so that important technical terms can be translated and searched properly. The system stores patent and utility model documents machine translated from Chinese and Korean to Japanese, so that Japanese key words can be used for the full-text search of these documents. As of the end of March 2015, the system can be used to search about 12 million Chinese and Korean patent and utility model documents and it is planned to increase the number of such searchable documents gradually in the future.

The quality of texts machine translated by the system plays an important role if the CKGS is to be used for practical applications. To this end, the JPO has been conducting projects since Fiscal 2012 to develop a specialized dictionary for patent terms which will be used for the system's machine translation, as its initiatives to improve machine translation quality (JPO, 2013, 2014a, 2015). Specifically, corresponding Japanese and Chinese sentences are extracted from the patent family of a same invention applied to Japan and China in order to prepare Chinese-Japanese translation corpuses as well as to develop a specialized dictionary. Currently, about 153 million Chinese-Japanese translation corpuses and the specialized user dictionary containing more than 2.2 million words have been developed. The user dictionary is installed to the CKGS to improve the system's translation quality.
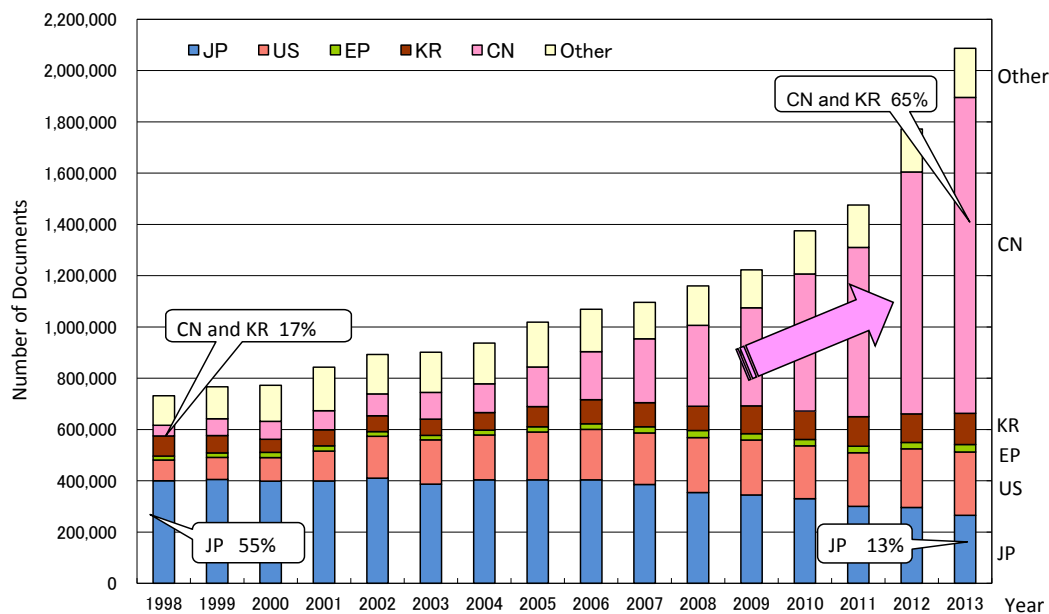
Figure 1. Rapid increase in Chinese patent documents. Note: Patent (incl. utility model) documents issued worldwide are categorized by language and duplicated data are eliminated. Regarding patent documents for the same invention which have been filed and published at multiple Offices, those documents published in Japanese are counted as JP. In case of no Japanese publications, such documents are counted first as US (English) , secondly as EP (English, French, German) ,thirdly as KR (Korean) and last as CN (Chinese), if each language is applicable.

Furthermore, measures are taken to enable the system to handle new technical terms as they appear. Specifically, based on information concerning unknown words which are detected as those not registered in the machine translation dictionary during the system's machine translation and that concerning incorrect translations and other errors reported by the system's users, efforts are made to update the dictionary, make additional registrations to the system's translation memory, and tune the translation engine parameters. With these efforts, the dictionary and the translation engine will be improved so that new technical terms can be handled as they appear.

In addition, through the Internet, the CKGS has become available not only to examiners but also to general users. As described above, continued efforts are being made to improve the system's machine translations so that practical application environment for uses to search Chinese and Korean documents can be provided, and the system's users appreciate it very much.
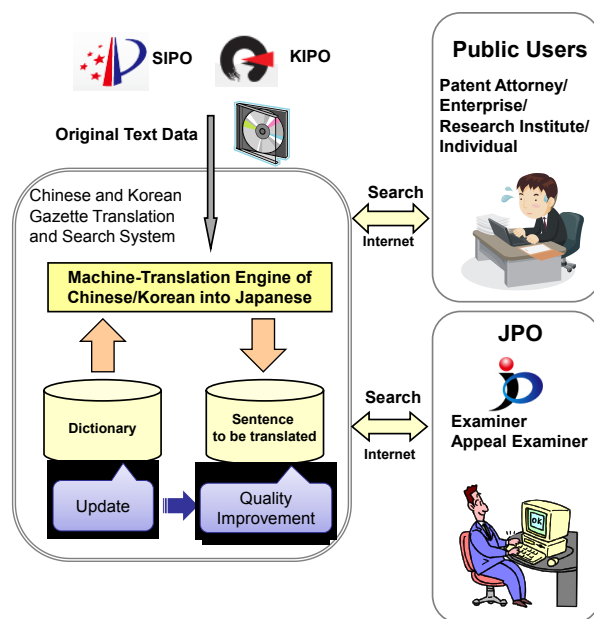
Figure 2. Conceptual diagram of Chinese and Korean Gazette Translation and Search System.

2.2. Quality Evaluation of Machine Translations from Chinese to Japanese and Those from Korean to Japanese

For realizing machine translations to be widely accepted and used, the quality of such translations plays a very important role. As a prerequisite to deciding whether to introduce such machine translations, the quality of such translations must be evaluated properly. A "Survey on How to Evaluate Quality of Patent Document Machine Translations" was conducted in FY 2013 to investigate how the quality evaluation method of machine translated patent documents should be (JPO, 2014b). Based on the survey's results, "Quality Evaluation Procedures for Patent Document Machine Translations" was developed in Fiscal 2014 as a guideline to evaluate the quality of machine translation results properly (JPO, 2014c).

In order to verify the results of the JPO's efforts to improve the system's machine translation quality in accordance with the "Quality Evaluation Procedures for Patent Document Machine Translations," machine translations provided by the Chinese and Korean Gazette Translation and Search Systemwere evaluated in the end of 2014 with regard to how accurately the system could translate technical terms (or the system's translation accuracy of technical terms) and how well the system could communicate original contents (or the system's original contents communication level).

With regard to the system's translation accuracy of technical terms, the system's machine translations of Chinese documents into Japanese were surveyed. A total of 196 technical terms were selected from various fields, and evaluated and classified into the following 4 grades by human evaluators.

- A (Properly Translated Word): Compared with one translated by a human, it is a word translated to a technically same or similar meaning and generally used.
- B (Acceptably Translated Word): It is not a translation word generally used as a technical word, but its meaning is almost correct.

- C (Mistranslated Word): It is a mistranslated word.
- D (Untranslated Word): It is a word not translated or whose translation was erroneously omitted.

With regard to the system's original contents communication level, the system's machine translations from Chinese and Korean to Japanese were surveyed. 100 machine translated sentences of each of the languages were selected from various fields, and evaluated and classified into the following 5 grades by human evaluators.
- 5: All contents of its important information are communicated correctly (100%).
- 4: Almost all contents of its important information are communicated correctly (80% or more).
- 3: No less than half of the contents of its important information are communicated correctly (50% or more).
- 2: Several contents of its important information are communicated correctly (20% or more).
- 1: Its translations cannot be understood, or almost no contents of its important information are communicated correctly (less than 20%).

In order to verify how effective it was to install a Chinese-Japanese dictionary developed by JPO to the system, the system's machine translations from Chinese to Japanese were evaluated before and after that dictionary with approximately 1 million words was installed to it for both surveys on the system's translation accuracy of technical terms and on the system's original contents communication level. In addition, for the purpose of comparison, the translated sentence using statistic machine translation (SMT) with 100 million Chinese-Japanese translation corpuses were also evaluated in the same manner.

Figure 3 shows the evaluation results of the system's original contents communication level through machine translations from Chinese and Korean to Japanese. While the system's average original contents communication level for Korean-Japanese translations is above 4 points, one for Chinese-Japanese translations is merely above 2 points. In addition, in all the technical fields, the system's original contents communication levels for Korean-Japanese translations are higher than those for Chinese-Japanese translations. It was concluded that this was due to the fact that the grammatical characteristics of Korean were closer to those of Japanese than those of Chinese, and the results indicate that continued efforts must be made to improve the system's translation accuracy of Chinese documents. Furthermore, the system's translation accuracy of either Chinese or Korean documents varies from one field to another, and measures must also be taken for individual fields. With regard to this, by the end of FY 2015, the JPO plans to measure the system's translation accuracy of Chinese documents in individual fields, and focus on developing the Chinese-Japanese dictionary for the technical fields whose translation accuracy is low.

Figures 4 and 5 show how accurately the system can translate Chinese documents' technical terms and how well the system can communicate their original contents, before and after the Chinese-Japanese dictionary developed by JPO was installed, respectively. That dictionary being added to the system's translation engine, the system's translation accuracy of technical terms increased by about 4%, and the system's original contents communication level by about 13%. Specifically, the translation accuracy of technical terms by the system's machine translation based on a rule-based method is higher than one by the SMT.
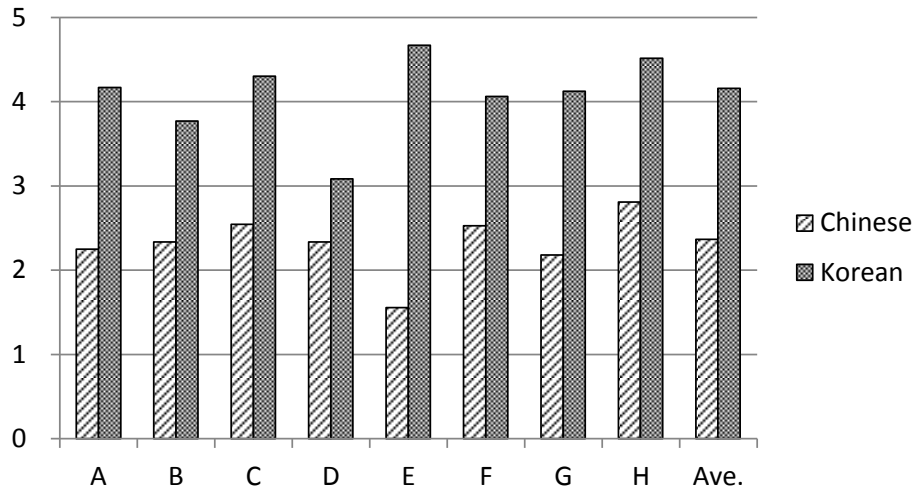
Figure 3. Original contents communication levels through machine translations from Chinese and Korean to Japanese in various fields. The figures show average evaluation points for respective fields averaging points of individual machine translations in the respective fields which are evaluated according to the 5 ranks of 1 to 5 points. "A" represents human necessities, "B" performing operations/transporting, "C" chemistry/metallurgy, "D" textiles/paper, "E" fixed constructions, "F" mechanical engineering/lighting/heating/weapons/blasting, "G" physics and "H" electricity.

On the other hand, the original contents communication level by the SMT is evaluated to be higher. Specialized technical terms are frequently used as key words to search patent documents, and the translation accuracy of such technical words plays an important role for
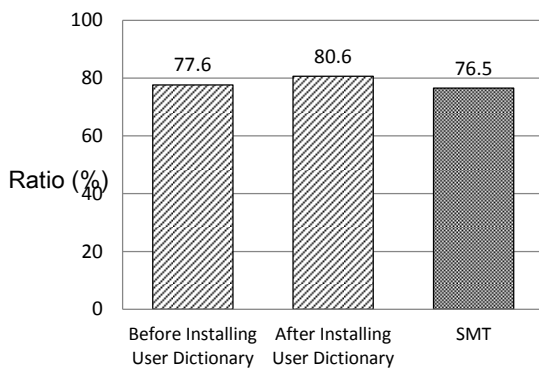

Figure 4. Evaluation of translation accuracy of technical terms in Chinese-Japanese machine translations: The above figures represent the sums of evaluation A (Properly Translated Word) and B (Acceptably Translated Word) ratios averaged over the 196 words.
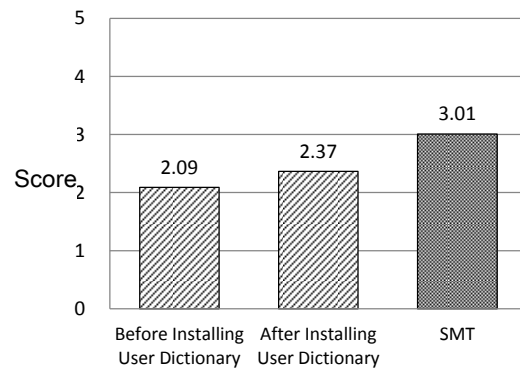

Figure 5. Evaluation of original contents communication Level: The above figures represent 5 level evaluations of 1 to 5 points on machine translations averaged over the 100 sentences.

such searches. However, original contents communication level is also important to understand documents' contents, and thus efforts must be made to improve the system's translation quality.

3. Providing Information on Examination in English

3.1. Advanced Industrial Property Network (AIPN)/One Portal Dossier (OPD)

The number of yearly patent applications in the world increased 1.6 times in 10 years from 2004 to 2013, the 80%, which showed 2.08 million applications in 2013, were filed with five intellectual property offices in Japan, U.S., Europe, China and the Republic of Korea (or the IP5 Offices). Out of these, not a few applications for single inventions were filed in multiple countries and regions.

In order to facilitate examination work sharing among examiners in respective IP offices, the JPO has been providing foreign IP offices with its information on patent applications and examinations (dossier information) translated from Japanese to English by machine translation since October 2004 through a network called the Advanced Industrial Property Network (AIPN). The AIPN enables individual IP offices' examiners to cross check the individual IP offices' information on patent applications and examinations (dossier information) mutually so that if a same invention is filed to multiple countries and regions, duplicate searches and examinations on the same invention can be excluded. The AIPN enables Japan's dossier information to be understood in English, and individual IP offices' examination results to be shared among the individual IP offices.

Furthermore, an IT service called the "One Portal Dossier"(OPD) was started in 2013
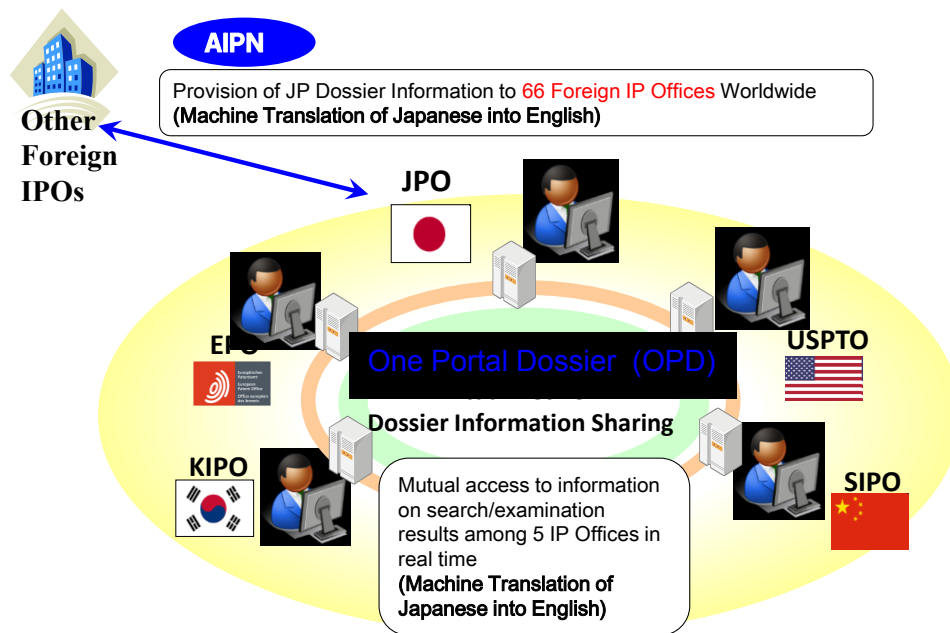

Figure 6. Conceptual diagram of AIPN/OPD.

for the IP5 offices' examiners, which collects the IP5 offices' information on examinations together, including the JPO's information on examinations machine translated by the AIPN, and provides such information in an easy-to-see format. Currently, thisOPD's service is enhanced, and the JPO's information on examinations is machine translated and provided to general users together with other IP5 offices' dossier information. In 2014 alone, through this service, the JPO's examiners browsed about 160,000 cases in the dossier information of the EPO, the USTPO, the SIPO and the KIPO, and the examiners of the EPO, the USTPO, the SIPO and the KIPO browsed about 210,000 cases in the dossier information of the JPO. Such dossier information helps to make examinations to be more efficient.

The JPO's information on examinations, which is provided through the AIPN/OPD, contains texts machine translated by a translation engine based on a rule based method. In general, machine translations from Japanese to English tend to show low translation accuracy when compared with those from English to Japanese. Notwithstanding, the JPO is constantly enhancing its machine translation dictionary to improve its service's translation accuracy. Specifically, the JPO collects untranslated words (unknown words) and registers their translations to its dictionary, as well as analyze feedbacks about mistranslations in the AIPN from foreign IP offices and incorporate their corrections into its dictionary. Thereby, the JPO is making efforts to enhance its dictionary.

3.2. Quality Evaluation of Japanese-English Machine Translations

In order to verify the translation accuracy of machine translations from Japanese to English which the JPO provides, in 2011 the JPO conducted a "Survey on Translation Accuracy Evaluation for Providing English Texts of Machine Translations of Information Related to Patent Examinations" (JPO, 2011). This survey evaluated 511 sentences extracted from patent gazettes and reasons for refusal in the fields of electricity, physics and chemistry. Two types of translation methods, that is, the JPO's machine translation system (AIPN) and commercially available Japanese-English translation software for patent, were applied to the sentences to obtain their machine translations, and the 2 types of machine translation methods were evaluated automatically as well as by human evaluators. The human evaluators evaluated the individual machine translations from the view point of whether the machine translations correctly reflect their originals' meanings, and assigned 0 to 4 points to each of them according to a 5 grade evaluation scheme. Three kinds of evaluation methods, that is, BLEU, NIS and IMPACT were used for the automatic evaluation.

Figure 7 shows the results of the human evaluators' evaluation. While the AIPN's average point is 2.5 points, the commercial product's is 2.0 points. In addition, the results indicate that the AIPN's points are higher than the commercial product's in all the technical fields. Results obtained by the automatic evaluation methods (BLEU, NIS and IMPACT) show trends almost similar to one obtained by the human evaluator's evaluation (Figures 8 to 10). Because of the results mentioned above, it can be concluded that the JPO's efforts to improve the accuracy of its Japanese-English machine translations have made some progress. However, it is noticed that the translation accuracy varies from one technical field to another. In addition, different evaluation methods produced different results, and a method to evaluate translation accuracy must be reviewed and examined continually. The JPO is still continuing to make its efforts to improve the quality of its Japanese-English machine translations even after the 2011 survey, and also plans to conduct another survey on the quality evaluation of its Japanese-English machine translations including the analysis of evaluation methods in FY 2015 in order to verify the results of this effort.
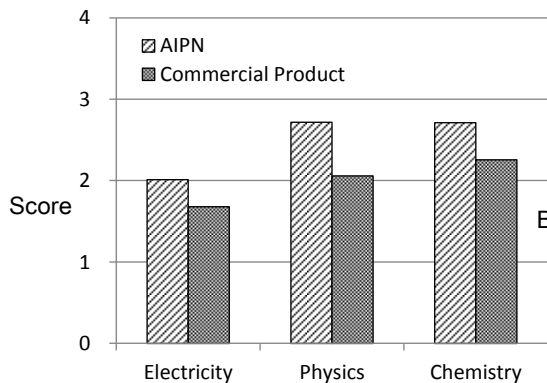
Figure 7. Evaluation of Japanese-English machine translations' accuracy by human evaluators.
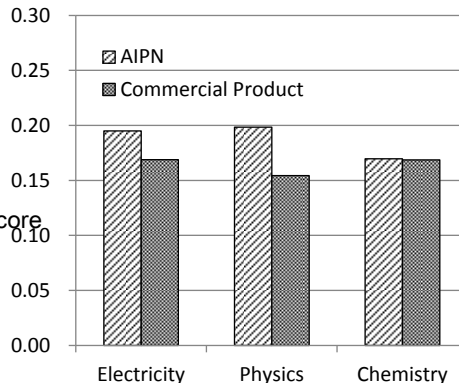

Figure 8. Evaluation of Japanese-English machine Translations' accuracy by BLEU.
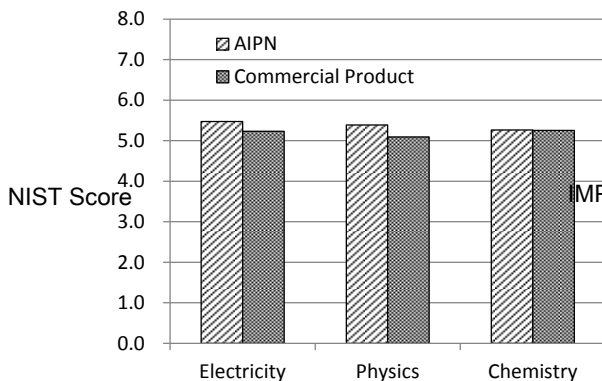

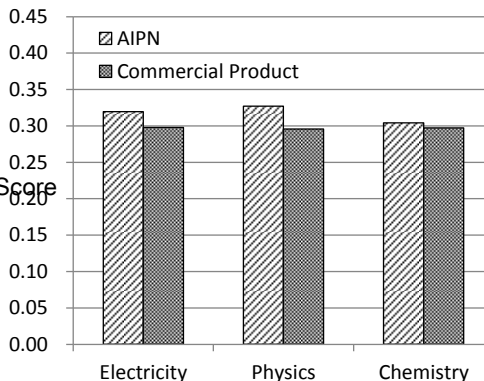Figure 9. Evaluation of Japanese-English machine translations' Accuracy by NIST.

Figure 10. Evaluation of Japanese-English machine translations' Accuracy by IMPACT.

4. Conclusion

Demands for machine translations of patent information are increasing year by year from the view point of patent information searches and dissemination. While machine translation is a developing technology and is being improved constantly, its accuracy has not become satisfactory yet, and one of its important issues is how to improve its accuracy. To improve machine translation quality, a solution by translation logic may be effective to solve difficulties in analysis due to languages' grammatical characteristics. The JPO is evaluating the quality of Chinese-Japanese and English-Japanese machine translations in Fiscal 2015, and henceforth it may be necessary in the future, for example, to consider whether to adopt a statistic machine translation engine instead of a rule-based machine translation method based on quality evaluation results and the fact that the translation accuracy of statistic machine

translation is improving in recent years. In addition, to improve translation quality, methods and criteria to evaluate individual machine translation schemes' qualities are also important because they are preconditions to evaluating translation quality. With regard to this, it is a common issue among the IP5 offices how to improve machine translation quality, and the offices plan to have further discussions on making methods and criteria for machine translation quality evaluations common among the IP5 offices.

References

JPO (2013). Survey on Chinese-Japanese Dictionary Development and Machine Translation Performance Improvement for Machine Translations of Chinese Patent Documents.

JPO (2014a). Survey on New Words for Machine Translations of Chinese Patent Documents .

JPO (2015). Survey on Dictionary Development and Machine Translation Quality Evaluation for Machine Translations of Chinese Patent Documents.

JPO (2014b). Survey on How to Evaluate Quality of Patent Document Machine Translations .

JPO (2014c). Quality Evaluation Procedures for Patent Document Machine Translations.

JPO (2011). Survey Report on Translation Accuracy Evaluation for Providing English Texts of Machine Translations of Information Related to Patent Examinations.