Use of the Technical Field-Oriented User Dictionaries

Tadaaki Oshio, Tomoharu Mitsuhashi, Tsuyoshi Kakita
Japan Patent Information Organization

Abstract

Japio performs various patent-related translation businesses, and owns the original patent-document-derived bilingual technical term database (Japio Terminology Database) to be served for the translators. Currently the database contains more than 780,000 J-E technical terms.

To adapt the database to the Patent Translation Task, Japio compiled machine translation dictionaries from it. 34 technical field-oriented dictionaries were created based on Japio's original technical fields. Terms are evaluated and selected according to their actual frequency in the bilingual patent document corpus of concerned technical field.

1. Introduction

Japio (Japan Patent Information Organization) is a non commercial organization serving economy and society through provision of high-quality patent information. One of its main activities is translation of various patent-related documents. Currently, the number of documents to be translated per year is about 400,000. Consequently, documents to be translated vary among widely-ranged domains including chemical, electrical, mechanical and physical.

Japio has engaged in translation business since 1971. There had been some attempts to introduce machine translation into the translation process in 1980s and 1990s, including compilation and use of the original patent-originated technical term dictionary. Then in the late 1990s, the Japanese Patent Office launched the IPDL, while the European Patent Office started the esp@cenet. Those services provide the translated Japanese patent document databases searchable in English language. In this context, unification of technical terms in patent document translation has become a priority subject. To achieve this goal, Japio restarted enhancing its original technical term database in 2000. We call the dictionary 'Japio Terminology database' in this paper. Japio Terminology database has been serving for the translators by providing them translations of technical terms.

2. Motivations

Japio's main motivation for participating in the NTCIR Patent Translation Task [1, 2] is to evaluate the advantage/disadvantage of the Japio Terminology Database, the technical field-oriented machine translation dictionaries created from it, and the machine translation system based on the dictionaries. We also have interests in the correlation between BLEU value and human rating, as well as the capability of latest alternative translation methods such as SMT and EBMT. It is a good opportunity for us to know the actual level of our database, dictionaries and machine translation system, to considering how we should improve them.

3. Japio Terminology Database

Japio is not a developer of MT engine; the system we used for the Patent Translation Task is based on a commercial MT engine. Difference is that we applied our original patent-derived Japio Terminology Database to the system.

Each record of the Japio Terminology Database consists of Japanese term, English term, and the Japio technical field. Currently the dictionary contains more than 780,000 records. All records of the database are extracted from patent-related source documents. Japio technical field consists of 34 fields.

Actually, the Japio Terminology Database itself is not directly applied to the PAJ translation-aid system. The database is used as the source of technical field-oriented dictionaries. Only the necessary terms are extracted from the database according to each technical field to compile the technical field-oriented 34 dictionaries (Japio Technical Field Dictionaries). The dictionaries are switched over according to the source document when they are put into use in the PAJ translation-aid system. Note that it is not an automatic process: the source documents are also subdivided into 34 subsets according to Japio technical fields, and the operator selects the dictionary of the same technical field manually according to the document subset to be translated. Therefore, the records of the Japio Technical Field Dictionaries don't contain the Japio technical field information.

4. Measures for translational variations

One Japanese term may correspond to multiple English translations. Such overlapping of terms often happens in our technical field-oriented dictionaries. To cope with this problem, Japio takes two countermeasures. The first measure is associated with Japio's terminology collection policy. That is, we give priority in registering long compound words to the Japio Terminology Database as they are, rather than breaking them down to more short, versatile terms. It is because long compound words tend to have less translational variations, and thus have less possibility of overlapping each other. Currently, the average length of Japanese terms in the Japio Terminology Database is 7.24 characters.

As the second measure, Japio also introduce the concept of 'Frequency Index' to set priorities between the overlapping terms. The Frequency Index of each term is determined by the frequency in the patent document parallel corpus. Japio has the sentence-aligned parallel corpus of the Japanese patent documents and their English translations. The corpus is subdivided into 34 subsets based on Japio technical fields to be used for index calculation. Each subset is used for calculating the Frequency Index of a Japio Technical Field Dictionary of the same field. The index of a term is determined by the number of the English-translated documents which include the English term among the ones which have corresponding Japanese documents that contains the Japanese term. For example, if there were 100 Japanese documents which include the Japanese term "表示" in a certain subset and among them 25 corresponding English documents include the corresponding English term "indication," the Frequency Index of the term "表示/indication" is 0.25. We use this index to rank the overlapping terms in a Japio Technical Field Dictionary, as well as to exclude minor translations from the dictionary by setting threshold. Consequently, the translation most frequently appeared in the past bilingual patent documents is provided to the translator as the 'recommended' translation for each certain technical term.

5. Japio Technical Field Dictionaries

As aforementioned, the 34 technical field-oriented dictionaries, called Japio Technical Field Dictionaries, are created from the Japio Terminology Database for the manual selective use in the machine translation according to the technical field of the source document. Extraction of dictionary terms stands on the Frequency Index of the term regarding the target technical field. When more than one record which have the same Japanese term exist in the Japio Terminology Database, the one with the highest Frequency Index value in the target technical field is selected. If the Frequency Index is less than 0.1, the record is regarded as 'minor translation,' thus excluded. Also in case the Frequency Index is not calculatable (namely, there are no appearance of the Japanese term in the parallel corpus), the record is excluded.

Through this process, only the English term which was actually used most often in the past patent documents is extracted as the correspondent of each Japanese term. Therefore, the average number of terms in the Japio Technical Field Dictionaries is narrowed down to 24,373. The total number of unique records extracted in those dictionaries is 201,445, while the total number of unique Japanese terms is 185,629. The total record number indicates that there are many records that are applicable to more than one dictionary. The difference between the total number of extracted records and that of unique Japanese terms represents there are considerable number of Japanese terms that are translated differently according to the technical field of the source document.

6. Participation to the Task

Japio participates in the Japanese-English intrinsic evaluation. As for the formal run, we switch over the 34 Japio Technical Field Dictionaries manually according to the technical field of the source document as aforementioned (As for the dry run, we just use a single dictionary to all the source documents). We also use a set of pre-installed commercial technical term dictionaries under our original dictionary and switch it over as well. We didn't take any particular measures to tune the system for the Task. We didn't even use the training data for dictionary improvement.

Japio submit two translation results for the formal run. The difference between them is the MT engine to be used. Both of the engines are rule-based commercial engines.

The Japio Technical Field Dictionaries are equally used in both formal run A and B. Consequently, any difference between their outputs is considered to be caused by the difference of the MT engines or their pre-installed dictionaries. In other words, comparing those two outputs doesn't illustrate the impact of the technical field-oriented dictionaries derived from the Japio Terminology Database. The reason why we use plural engines and submit two runs is rather to confirm the order of superiority between rule-based MT method and other methods.

We choose the formal run B to be evaluated by human judgments. It is just because we already had a human evaluation of the MT engine used for formal run A in the course of submission for the dry run.

7. Results

Detailed information on our submissions; dry run, formal run A and formal run B are described in the Table 1 in Page 4. The 'Run-ID' column indicates those three submissions.

The 'TASK' column shows that each of them is Japanese-to-English translation. The 'ID' column shows the "GROUP-ID" specified by the NICTR-7 organizer. The 'TYPE' column indicates each translation is performed by the rule-based machine translation system. The 'RESOURSE' column represents that Japio didn't use the training data provided by the organizer for dry run nor formal runs.

The 'EXTERNAL' column indicates the type of the dictionary Japio used for each run. "Japio-Dic-34" represents the 34 Japio Technical Field Dictionaries, each of which was compiled from the Japio Terminology Database based on the frequency in bilingual patent documents of concerned technical field. The table shows that we used the Technical Field Dictionaries for the formal runs and switched them over according to the technical field of a source document. Switching of the dictionaries were done manually. The "Japio-Dic-1" means that we didn't use those 34 dictionaries for the dry run, but used the single dictionary which was compiled based on the frequency in the whole patent document parallel corpus. Note that the Japio-Doc-1 dictionary is not an aggregation of the 34 technical field-oriented dictionaries. The Japio-Dic-1 dictionary was compiled based on the frequency in the whole patent document parallel corpus, while each of the 34 technical field-oriented dictionaries was compiled based on the frequency in a part of patent document parallel corpus according to concerned technical field.

The 'CONTEXT' column represents that Japio didn't use any context information for any of its translations. As for 'OFFLINE TIME" and 'ONLINE TIME,' each indicates the time used for pretreatment (format of test data and subset generation according to Japio Technical Fields) and for actual machine translation executed upon the Excel. Specification of the computer used for each run ('MACHINE-SPEC') and the outline of the system ('SYSTEM-DESCRIPTION') are also included in the table.

The evaluations for these submissions fed back from the NTCIR-7 organizer are also described in the Table 2. 'Human Judgment Adequacy Avg' and 'Human Judgment Fluency Avg' are the average ratings of 100 sentences evaluated separately by 3 human evaluators. Those columns are left blank for the formal run A since it wasn't evaluated by human experts.

The 'BLEU-JE' column indicates the single reference BLEU values of each submission. 'BLEU-JE-LOW' and 'BLEU-JE-HIGH' are the lowest and highest BLEU values among the whole test data within 95% confidence interval.

8. Review of the results

The human ratings of Japio's submissions, both of the dry run and the formal run, outperform all SMT and example-based methods. The result indicates that the rule-based method is still most suitable for our practical use.

On the other hand, the single-reference BLEU values of the rule-based methods are quite moderate among the participants. It seems that single-reference BLEU evaluation is not always suitable for comparing the performances of different machine-translation methods.

What Japio intends to confirm by participating in the Patent Translation Task is the effectiveness of the Japio Terminology Database, the 34 Technical Field Dictionaries derived from it, and the machine translation system based on these dictionaries. Unfortunately, we cannot find any particular advantage of our system as far as comparing it with the other RBMT entry. In fact, the output of our system is slightly inferior to the other RBMT entry both in BLEU value and human rating.

There are many factors that may cause it; the difference in the used engines, the difference in their versions (Japio doesn't use the latest versions), the lack of tune-ups for the Task. Of course we should examine if there were any unintended adverse effect of using the Japio Technical Field Dictionaries. But at least, we can presume the advantage in switching the dictionaries over according to the technical field of target documents, for the BLEU value of the formal run A outperforms the value of the dry run, in which the same engine and similar test data were used.

References

[1] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop

[2] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

Table 1. Detailed Information on Japio's Submissions

Run-ID	TASK	ID	TYPE	RESOURCE	EXTERNAL	CONTEXT	OFFLINE TIME	ONLINE TIME	MACHINE-SPEC	SYSTEM-DESCRIPTION
Dry Run	JE	D	RBMT	NONE	Japio-Dic-1	NO	1 hours	10min	P4 2.4GHz 512MB memory	A commercial MT system A Merged dictionary
Formal Run A	JE	D	RBMT	NONE	Japio-Dic-34	NO	4 hours	10min	P4 2.4GHz 512MB memory	A commercial MT system A & 34 technical field dictionaries
Formal Run B	JE	D	RBMT	NONE	Japio-Dic-34	NO	4 hours	10min	P4 2.4GHz 512MB memory	A commercial MT system B 34 technical field dictionaries

Table 2: Evaluations for Japio's Submissions

Run-ID	SYSTEM-DESCRIPTION	Human Judgment Adequacy Avg.	Human Judgment Fluency Avg.	BLEU-JE	BLEU-JE-LOW	BLEU-JE-HIGH
Dry Run	A commercial MT system A Merged dictionary	2.95	2.54	18.66	17.92	19.50
Formal Run A	A commercial MT system A 34 technical field dictionaries			20.33	19.63	21.08
Formal Run B	A commercial MT system B 34 technical field dictionaries	3.71	4.02	19.46	18.78	20.14