Use of the Japio Technical Field Dictionaries for NTCIR-PatentMT

Tadaaki Oshio
Japan Patent Information Organization
1-7, Toyo, 4-Chome, Koto-Ku
Tokyo 135-0016, JAPAN
+81-3-3615-5513
t_oshio@japio.or.jp

Tomoharu Mitsuhashi
Japan Patent Information Organization
1-7, Toyo, 4-Chome, Koto-Ku
Tokyo 135-0016, JAPAN
+81-3-3615-5513
t_mitsuhashi@japio.or.jp

Tsuyoshi Kakita
Japan Patent Information Organization
1-7, Toyo, 4-Chome, Koto-Ku
Tokyo 135-0016, JAPAN
+81-3-3615-5513
t_kakita@japio.or.jp

ABSTRACT

Japio performs various patent-related translation businesses, and owns the original patent-document-derived bilingual technical term database (Japio Terminology Database) to be used by the translators. Currently the database contains more than 1,000,000 J-E technical terms.

The Japio Technical Field Dictionaries (technical-field-oriented machine translation dictionaries) are created from the Japio Terminology Database based on each entry's frequency in the bilingual patent document corpus compiled by Japio.

Japio applied the Japio Technical Field Dictionaries to a commercial machine translation engine for the NTCIR9-PatentMT (JE and EJ subtasks).

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – Machine Translation

General Terms

Management, Performance, Experimentation, Verification.

Keywords

Japio Terminology Database, Japio Bilingual Sentence-alignment Corpus, Japio classification, Japio Technical Field Dictionaries, Frequency Index

TeamName: [Japio]

Subtasks/Languages: [Japanese-to-English], [English-to-Japanese]

External Resources Used: a commercial rule-based translation engine

1. INTRODUCTION

Japio (Japan Patent Information Organization) is a noncommercial 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 espacenet. 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. Consequently, Japio was required to provide a common technical term dictionary to be shared among the translators. 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.

On the other hand, Japio has created the sentence-aligned bilingual (JE) corpus by compiling the translations of patent documents produced via its translation activities and their originals. We call this corpus 'Japio Bilingual Sentence-Alignment Corpus' in this paper.

2. MOTIVATIONS

Japio's main motivation for participating in the "patent machine translation" (PatentMT) at NTCIR-9 is to evaluate the advantage/disadvantage of the Japio Terminology Database, the technical field-oriented machine translation dictionaries (the Japio Technical Field Dictionaries) extracted from the database, and the machine translation system based on these dictionaries. We also have interests in the correlation between automatic evaluation 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 further improve them.

3. JAPIO TERMINOLOGY DATABASE

Japio is not a developer of MT engine; the system we used for the PatentMT 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 classification. Currently the dictionary contains more than 1 million records. All records of the database are extracted from patent-related source documents. See examples shown on Table 1.

The Japio classification is based on the International Patent Classifications (IPC) and it currently consists of 34 classes. Each class corresponds to a unit of the technical translators in charge, and this classification has been used for distributing patent documents to be translated to the appropriate translators. For example, the Japio Class 'C01' corresponds to the IPC classes A01 to A24 except for the subclasses A01N and A01P, as well as classes C12 and C13 except for the subclass C12N and covers the fields of agriculture, /forestry, fishery and the like.

Table 1. Examples of Japio Terminology Database entries

Japanese	English	Japio class
２サイクルエンジン用気化器	carburetor for two-cycle engine	M09
２サイクルエンジン用軸受付きコンロッド	connecting rod with bearing for two-cycle engine	M10
２サイクルエンジン用鋳造軽金属ピストン	cast light metal piston for a two-stroke engine	M09
２サイクルディーゼルエンジン	two-cycle diesel engine	M09
２サイクル火花点火エンジン	two-stroke spark ignition engine	M09
２サイクル内燃エンジン	two cycle internal combustion engine	M09

*Only the Japanese and English terms are copied to the Japio Technical Field Dictionaries.

In each of the JE/EJ PatentMT subtasks we used 34 machine translation dictionaries that were created based on the Japio classification. We call those dictionaries 'Japio Technical Field Dictionaries.' Please note that Japio classification information stored in each record of the Japio Terminology Database is used for selecting the entries of those dictionaries, but the classification information itself is deleted when the record is copied to those dictionaries, thus the records of those dictionaries do not contain any classification information.

4. MEASURES FOR TRANSLATIONAL VARIATIONS

One Japanese term may correspond to multiple English translations. Such overlapping of terms often happens in our Japio Technical Field 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 7.55 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 Japio Bilingual Sentence-Alignment Corpus. As aforementioned, Japio owns the sentence-aligned parallel corpus of the Japanese and English patent documents. This corpus is subdivided into 34 subsets based on Japio classification, and each subset is used for calculating the Frequency Index of a Japio Technical Field Dictionary of the same Japio class. 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.

5. JAPIO TECHNICAL FIELD DICTIONARIES

As aforementioned, the Japio Technical Field Dictionaries, 34 dictionaries for each of JE and EJ, are created from the Japio Terminology Database using the Japio Bilingual Sentence-Alignment Corpus, to be used in the machine translation selectively 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 calculable (namely, there are no appearance of the Japanese term in the parallel corpus), the record is excluded.

Through this process, the most-frequently used English translation is selected based on the certain class's subset of the Japio Bilingual Sentence Alignment Corpus, and contained in the JE Japio Technical Field Dictionary of the certain class. Of course vice-versa can be said for the process of the EJ dictionaries.

In such manner we created 34 JE Japio Technical Field Dictionaries as well as 34 EJ ones each of which corresponds to 34 classes of Japio classification.

The average number of records in the 34 types of JE Japio Technical Field Dictionaries is therefore narrowed down to 17,925, while the total number of unique records extracted in those dictionaries is 212,781 and that of unique Japanese terms is 192,781. As for the EJ dictionaries, their average number of records is 28,696, the total number of unique records is 244,735, and that of unique Japanese terms is 219,066.

6. TRANSLATION PROCESS

In each of the JE/EJ Patent Translation Tasks, Japio used a commercial machine translation engine and added the Japio Technical Field Dictionaries to it. We divided the set of source documents into 34 subsets according to the Japio classification in advance, and we switched the Japio Technical Field Dictionaries as well as the technical dictionaries of the engine manually according to the Japio class to which the source document subset belongs. The translation process consists of the following 3 steps:

1. We divide the set of source documents into 34 subsets according to the Japio classification. Japio class of each source documents is determined based on its initial IPC code which is obtained using its document number.

2. We translate each subset using a commercial machine translation engine. The Japio Technical Field Dictionary which corresponds to the Japio class of the subset is added to the technical dictionaries of the engine manually. Combination of the technical dictionaries of the engine is also determined preliminarily according to the Japio classification. The Japio Technical Field Dictionary and the set of the technical dictionaries of the engine are switched manually for every subset.

3. We merge all the translated subsets and sort back the documents to original order.

Combination and order of the technical dictionaries of the engine are determined experimentally. Word selection among dictionaries totally depends on the undisclosed mechanism of the machine translation engine used, although we assume that it is governed by the hierarchy of the dictionaries.

7. RESULTS

Japio submitted the translations for JE and EJ PatentMT subtasks. Tables 2 and 3 shows the parameters of each JE/EJ submissions and their evaluations returned from the organizers.

Table 2. Parameters and evaluations of JE subtask

SYSTEM-ID	JAPIO	
DIRECTION	JE	
PRIORITY	1	Only 1 submission
TYPE	RBMT	A commercial machine translation engine was used.
RESOURCE_BILINGUAL	NO	No bilingual training data was used.
RESOURCE_MONOLINGUAL	NO	No monolingual training data was used.
RESOURCE_EXTERNAL	YES	A commercial RBMT software and JE Japio Technical Field Dictionaries (avg. 17925 entries) were used.
RESOURCE	NO	No training data was used.
EXTERNAL	YES	JE Japio Technical Field Dictionaries (avg. 17,925 terms) were used.
CONTENT	YES	Japio class was used via IPC code and document number.
ONLINE-TIME	10min	Approximate estimate
MACHINE-SPEC		Intel® Core™2 Duo E8500 @3.16GHzCPU 3.21GB memory
BLEU	0.2035	
NIST	6.6180	
RIBES	0.7146	
ADEQUACY	3.6666	Average
ACCEPTAILITY	0.7117	Average

Table 3. Parameters and evaluations of EJ subtask

SYSTEM-ID	JAPIO	
DIRECTION	EJ	
PRIORITY	1	Only 1 submission
TYPE	RBMT	A commercial machine translation engine was used.
RESOURCE_BILINGUAL	NO	No bilingual training data was used.
RESOURCE_MONOLINGUAL	NO	No monolingual training data was used.
RESOURCE_EXTERNAL	YES	A commercial RBMT software and JE Japio Technical Field Dictionaries (avg. 17925 entries) were used.
RESOURCE	NO	No training data was used.
EXTERNAL	YES	EJ Japio Technical Field Dictionaries (avg. 28,696 entries) were used.
CONTENT	YES	Japio class was used via IPC code and document number.
ONLINE-TIME	10min	Approximate estimate
MACHINE-SPEC		Intel® Core™2 Duo E8500 @3.16GHzCPU 3.21GB memory
BLEU	0.2272	
NIST	6.2892	
RIBES	0.7088	
ADEQUACY	3.4633	Average
ACCEPTAILITY	0.6515	Average

8. REVIEW OF THE RESULTS

The translations submitted by Japio gained quite high evaluation scores in either of the JE and EJ subtasks. As for the JE subtask, the evaluation scores of Japio's translation outgo all the other ones in both adequacy and acceptability. As for the EJ subtask, Japio’s translation is ranked fourth in both adequacy and acceptability.

As aforementioned, Japio used a RBMT engine. RBMT method was obviously superior to other methods in NTCIR-7. However, in NTCIR-9 EJ subtask, one submission (G07-1) that used SMT method outran all the other submissions including the ones using RBMT method. It may show the fact that SMT method is catching up RBMT method.

Japio had submitted two translations to the JE PatentMT subtask in NTCIR-7 as well. The parameters and the evaluations for those two translations (RUN-A, RUN-B) are shown in the Table 4 below.

Table 4. Parameters and evaluations of NTCIR-7 JE subtask

SYSTEM-ID	JAPIO	
DIRECTION	JE	
PRIORITY	RUN-B	2 submissions
TYPE	RBMT	A commercial machine translation engine was used.
RESOURCE	NO	No training data was used.
EXTERNAL	YES	JE Japio Technical Field Dictionaries (avg. 24,373 terms) were used.
CONTENT	YES	Japio class was used via IPC code and document number.
ONLINE-TIME	10min	Approximate estimate
MACHINE-SPEC		Intel® P4 2.4GHz CPU 512MB memory
BLEU	20.33	RUN-A
ADEQUACY	-	
FLUENCY	-	
BLEU	18.78	RUN-B
ADEQUACY	3.71	
FLUENCY	4.02	

Japio used different commercial MT engines for RUN-A and RUN-B. We use the same Japio Technical Field Dictionaries for both translations. The engine we used for RUN-A was the same one, but older version of the one that we used in NTCIR-9 PatentMT subtasks

In NTCIR-7, only the RUB-B translation was evaluated by human evaluators, and it was ranked second in both adequacy and fluency. The highest scores in adequacy and fluency are 3.81 and 3.94 respectively, and they are both given to the same translation (TSBMT).

Japio has conducted extensive enhancement and revision of JE Japio Technical Field Dictionaries during the period between NTCIR-7 and 9. More than 200,000 terms are added to the source database (Japio Terminology Database), while many problems caused by the entries of those dictionaries were detected and corrected by proofreading the actual MT outputs. The main reason why the average number of records of the Japio Technical Field Dictionaries decreased from 24,373 at the time of NTCIR-7 to 17,925 at NTCIR-9 is that a lot of entries that cause ill effects to MT results in high ratio were excluded from the dictionaries.

On the contrary, not much enhancement was made for the EJ dictionaries before the NTCIR-9, when compared with the JE ones. The average number of EJ dictionaries is 28,696 and it represents that many entries that have ill-effects to MT outputs were included in the EJ dictionaries then. This is one of the reasons of the difference in evaluation scores between JE and EJ subtasks. Japio has already started the extensive revision of EJ dictionaries as well.

9. ACKNOWLEDGMENTS

We would like to thank the NTCIR-9 organizers for giving us a precious opportunity to get objective and comparative evaluations of our machine translation system outputs.

10. REFERENCES

[1] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou, Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, NTCIR-9, 2011.

[2] Oshio, T., Mitsuhashi, T, and Kakita, T. 2008. Use of the Technical Field-Oriented User Dictionaries. *Proceedings of NTCIR-7 Workshop Meeting*, 462-465. DOI= http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C3/PATMT/14-NTCIR7-PATMT-OshioT.pdf