Terminology Construction Workflow for Korean-English Patent MT

Young-Gil Kim, Seong-Il Yang, Munpyo Hong, Chang-Hyun Kim, Young-Ae Seo, Cheol Ryu, Sang-Kyu Park
ETRI
161 Gajeong-dong
Daejeon, Korea, 305-350
{kimyk,siyang,munpyo,chkim,yaseo,ryuch,parksk}@etri.re.kr

Se-Young Park
KyungPook National University,
Group2, Lab2, Organisation2
1370 Sankyuk-dong
Daegu, Korea, 702-701
seyoung@mail.knu.ac.kr

Abstract

This paper addresses the workflow for terminology construction for Korean-English patent MT system. The workflow consists of the stage for setting lexical goals and the semi-automatic terminology construction stage. As there is no comparable system, it is difficult to determine how many terms are needed. To estimate the number of the needed terms, we analyzed 45,000 patent documents. Given the limited time and budget, we resorted to the semi-automatic methods to create the bilingual term dictionary in electronics domain. We will show that parenthesis information in Korean patent documents and bilingual title corpus can be successfully used to build a bilingual term dictionary.

1 Introduction

In developing a practical MT system, to determine the volume of the linguistic resources is one of the most difficult and important tasks. When there are many similar systems already on the market, it would be relatively easy to estimate the size of the needed resources. The MT developers would simply need to compare the size of the resources of the other systems. However, when the research or the development of an MT system is unprecedented, it would be difficult to "guess" how many words or patterns they would need to have in their dictionary (Dillinger, 2001).

If the domain and the documents to be translated are fixed, the matters could be simple. They would simply extract all the unknown words in the documents, translate them, and add to the dictionary. However, the reality is not always the case. It is quite often the case that the MT developers suffer from the small budget and the short development period. Due to the high cost and the shortage of time, to translate all the unknown words manually would be quite unrealistic in many cases.

In Korea, the fruits of the intensive MT-research since early 1990s have begun to be gathered in many areas (Se-Young, 1999). Especially, given the enormously increasing number of the yearly applied patents, the needs for the high-speed and automatic translation are enormous. Recently, the disputes over intellectual property are happening more and more all around the world. From this reason, Korea, China and Japan have decided to offer the English translation service for their own patens in a few years to each other. In Korea, about 100,000 patents are said to be applied yearly. Each patent including the opinion of the patent examiner consists of 50 pages on average. In other words, about 5 million pages must be translated into English every year. Taking into account the cost and the efforts for the translation, no other solution than MT seems to be feasible.

ETRI (Electronics and Telecommunications Research Institute) has been developing a Korean-English patent MT system since 2004 under auspices of MIC (Ministry of Information and Communication). Last year, a Korean-English patent MT system "FromTo" for electronics domain was developed and installed at KIPO (Korean Intellectual Property Office).

In this paper we present the workflow of constructing patent terminology for electronics domain. The workflow consists of the stage for setting lexical goals and the (semi-) automatic terminology construction stage. In section 2 we show the method to set the lexical goals. Section 3 elaborates on the (semi-) automatic term construction methods. In section 4 we discuss about the evaluation result of the term construction methods. Finally, in section 5 we will conclude our discussion and show the future research directions.

2 Estimating the Number of Terms

There are many aspects and disciplines in terminology research (Sager, 1990). To build a terminology dictionary, the recent works in ATR (Automatic Terminology Recognition) have achieved good results (Dagan, 1995; Oh, 1999). In

this paper, we extracted all the unknown words from a certain volume of patent documents, and used them to estimate the size of terms to be constructed for the patent translation.

In the first phase of the development, we estimated the number of the terms to include in the term dictionary. Although there are some machine-readable term dictionaries available for electronics domain, they contain only the basic terms and sometimes out-of-date terms, so that they don't cover the entire terms appearing in the patent documents. In order to estimate the lexical coverage, we analyzed a Korean patent corpus in the electronics domain which corresponds to all the documents for 9 months. The corpus size was about 340 MB. All images were removed to get pure text data. The corpus consists of 22,756 patent documents that contain 2,667,198 sentences.

We examined the expected lexical coverage in two steps: the coverage of the single terms and the coverage of the compound noun terms. Given the limited time and budget for the lexical resource construction, to the most single noun terms was given the priority of inclusion in the term dictionary. As for the compound noun terms, the priority was given only to the terms with high frequency.

To estimate the coverage of the single terms, we analyzed the corpus using a Korean morphological analyzer. The Korean lexicon for general domain which contains about 160,000 entries, was compiled for the analysis. The POS-tagged result only for the best one candidate showed 92.1% of accuracy for patent documents in the electronics domain. We found 94,724 unknown single word terms. The POS-tagged result had some noise entries because of the limited analysis accuracy. The sampling evaluation of the result of unknown single word terms showed 87.3% of accuracy. From this evaluation, we expected that the actual size of the newly found unknown single word terms would be 82,694 entries. Although the cumulative size of the newly found unknown word terms seems to increase, the number of unique unknown word which is newly found in each document shows a converging point.

When the corpus size is extended to 45,500 documents, we can estimate the size of the unknown single word term, which is newly found in 2,275 documents, decreases to 4,000. The following graph shows the estimation result:

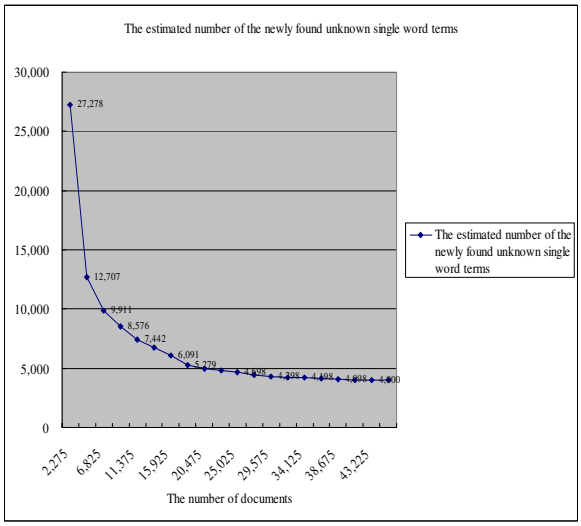

Fig. 1: The estimated number of the newly found u nknown single word terms

The number of the newly found unknown single word terms seems to converge at 4,000 entries per 2,275 documents when constructing about 130,000 single word terms. As for the above graph, 1.74 entries of unique unknown word can be found in a single patent document. But the size of the terms to be constructed can be reduced, because certain terms are frequently used. The relation between the frequency of the terms and the lexical coverage is shown in the following table:

	Unknown word terms newly found in each document	Total size of terms to be constructed
After analyzing 22, 756 documents	2.2 entries per 1 document	82,694 entries
After analyzing 45, 500 documents	1.76 entries per 1 document	136,958 entries

Table1: The relation between the frequency of the terms and the lexical coverage

After empirical study of expert translators, we decided to allow less than 2 unknown words in a document. According to the above coverage calculation, we found out that about 103,920 single word terms are needed to ensure that there are 1.98 unknown entries in a document.

To estimate the size of the compound noun terms, we counted the series of single nouns as a multi-word term. The following graph shows the size of the compound noun terms to be constructed.

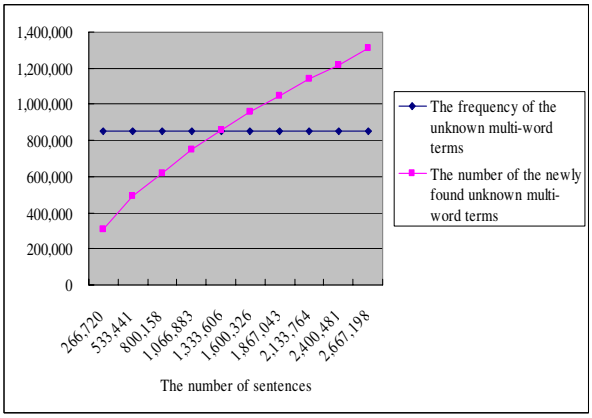

Fig. 2: The increasing number of the newly found unknown multi-word terms

As for the newly found unknown multi-word terms, there seems to be no converging point. But, as for the sampling examination of the newly found unknown multi-word terms, almost all the multi-word terms could be translated to the appropriate target-word based on the word-to-word generation of single noun's target word.

Given the above estimation, we finally decided to construct at least 103,920 single word terms and the multi-word terms with high frequency as our budget allows.

3 Building Korean-English Terms

To build a term dictionary is often a time-consuming and costly task. To cope with the bilingual term dictionary building, we set up the work process as the fig.1 shows. The process largely consists of 3 steps: bilingual term extraction based on parenthesis information of patent documents, bilingual term extraction based on patent bilingual titles, Korean term extraction and human translation.

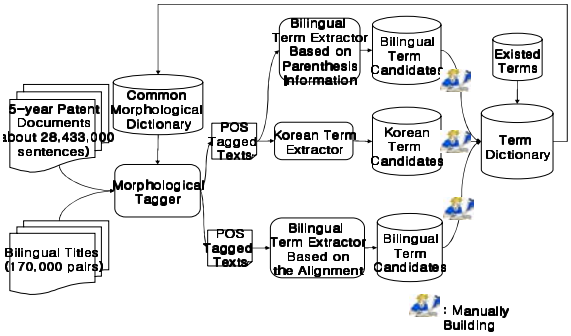

Fig. 3: Workflow for Building the Term Dictionary

Firstly, we have collected the existing technical term dictionaries and used them as an initial term dictionary. We add the entries of the term dictionaries to a common morphological dictionary used for morphological analysis and POS tagging. Secondly, ETRI morphological POS tagger analyzes patent documents and Korean sentences of bilingual titles to build POS tagged texts. Using the tagged texts, a bilingual term extractor based on the parenthesis information and bilingual titles extracts bilingual term candidates. Domain experts simply accepted or rejected the extracted translation pairs and modified some incorrect pairs. Through the semi automatic process, we could get the large-scale terms with relative ease in short time. After applying the automatic term extraction methods, the extracted terms are added to the common morphological dictionary and POS tagger again analyzed the patent documents to produce more correct POS tagged texts. Subsequently, the Korean term extractor extracts Korean term candidates that aren't translated yet, and human translation of the rest extracted terms was performed. Putting all together, we could build about 600,000 entries for electronics domain, thus far exceeding the goals we initially set.

3.1 Using parenthesis information of patent documents

The next step is to translate the extracted unknown terms to English. However, as the budget and the time are limited, we resorted to (semi-) automatic methods for the term dictionary building. The most important clue for the (semi-) automatic construction of the term dictionary was the frequent use of parentheses after a Korean technical term (Hisamitsu, 1998). The following sentence exemplifies the characteristic:

> 이는 크게 **오버헤드 콘솔(Overhead console, 110)**과 **디스플레이 프레임(Display frame, 120)**으로 구성된다.

In the documentation of Korean patents, the authors tend to "elaborate" or to "expatiate" on the technical terms using parentheses. Usually, the English translations of the terms are within the parentheses. Based on this characteristic, we extracted 420,180 Korean-English pairs. The following data shows a Korean-English pair of a term "에치백(*etch back*)". A pair consists of multiple candidates that have their English translations, frequencies, and sentence examples. Domain experts simply accepted or rejected the extracted translation pairs.

> etch back {28} [**에치 백(etch back)** 공정으로]
> etch-back {4} [**에치백(etch-back)** 방법,]
> etched back {1} [**에치백(etched back)**된다.]
> etching back {1} [**에치백(etching back)**을 수행한다.]

3.2 Using patent bilingual titles

Another valuable resource for the semi-automatic term dictionary construction was bilingual patent title corpus. Although a patent is applied in Korean language, the title of the patent must be written both in Korean and English. Using alignment technique, we could extract English translation candidates for Korean compound nouns in patent title sentences. We aligned Korean and English compound nouns, using POS tagged results, common dictionary and the available term dictionary. The following data shows the candidates extracted from the 2 titles in which the English words appear. Using this method we could build about 100,056 compound noun entries from bilingual corpus with relative ease.

<광촉매 박막>
photocatalytic thin film
광촉매 박막 및 이것을 구비한 물품
{ thin photocatalytic film and articles provided with the same }
자외선과 광촉매 박막을 이용한 수중 투입형 광화학 반응장치
{ water immersion type photochemical reaction device using UV and photocatalytic thin film }

3.3 Using bilingual compound nouns

After building the term dictionary semi-automatically, Korean term extractor extracted Korean single noun terms that aren't translated yet. If a Korean term is included in bilingual compound nouns that are built through the automatic process, we could calculate its translation frequency from English compound nouns and present the translation with highest frequency as its most-likely translation candidate. For example, a Korean term "스트로브(*strobe*)" occurs repeatedly in 182 compound nouns, and the occurrence frequency 174 of an English translated word "strobe" is higher than the other English words and is thus selected as the first translation candidate. We also could build about 39,208 terms from bilingual compound nouns relatively easily.

4 Evaluation

In automatically extracting the bilingual term pair candidates, the candidates may contain translation errors caused mostly by morphological analysis, Korean term range detection, and English translation. In case the English translation to the Korean term is wrong, the domain experts could easily correct the errors comparing with the neighbouring translations. As for the method with parenthesis, about 58% of the bilingual term pairs were accepted without any correction. About 42% of the pairs contained an error either in Korean or English side. Among them, 23% could be modified and accepted as bilingual terms and 77% of incorrectly extracted terms were discarded by domain experts. Employing this methodology, we could build about 249,905 entries. The number of the bilingual terms constructed for electronics domain using each methodology is shown in the following table:

Building Method	Term candidates	Building without any correction	Building with any correction
Using parenthesis information	369,354	214,225 (58%)	35,680 (9.66%)
Using patent bilingual titles	115,006	47,152 (41%)	52,904 (46%)
Using bilingual compound nouns	41,839	34,726 (83%)	4,482 (10.71%)
Total	526,199	296,103 (56.27%)	93,066 (17.69%)

Table 2: Bilingual terms that are built without or with any correction

In section 2, we estimate that at least 103,920 entries should be constructed. Among 389,169 terms constructed in the above-mentioned way, the single terms that are included in 103,920 entries were 63,962. As for the rest 39,958 entries, human translation was performed as a last recourse. Adopting the semi-automatic terminology construction workflow, we could reduce the cost over 50% and get 317,207 useful compound noun terms in short time.

5 Conclusion

In this paper we presented the workflow for constructing the linguistic resources for Korean-English patent MT system. To estimate the number of the terms to include in the term dictionary can be difficult, when there is no comparable system. Even if the lexical goals are set, to construct the bilingual term dictionary is often time-consuming and costly. To estimate the number of the terms, we analyzed 45,500 patent documents. The analysis results showed that we need about 104,000 single word terms to ensure that about 1.98 unknown words occur in a document.

To construct the bilingual term dictionary in the limited time and budget, we employed semi-automatic methods. The idea was to make most use of the parenthesis information of patent documents and patent bilingual titles. When there is no semi-

automatic construction method, human translation was performed as a last recourse.

This year we extended the domain from electronics to all patent areas. Having successfully constructed the bilingual terms in electronics last year, we apply the same method to all the areas. We are expecting to have constructed about 2 million bilingual entries by the end of this year.

References

Dagan, I. and K. Church. 1995. *Termight: Identifying and translating technical terminology*. In "Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics", pages 34-40.

Dillinger, M. 2001. *Dictionary Development Workflow for MT: Design and Management*, In "Proceedings of the 8th MT Summit".

Hisamitsu, Toru and Yoshiki Niwa. 1998. *Extraction of useful terms from parenthetical expressions by using simple rules and statistical measures*. In "First Workshop on Computational Terminology Computerm", pages 36-42.

Oh, J.H. and K.S. Choi. *Automatic extraction of a transliterated foreign word using hidden markov model*. In "Proceedings of the 11th Korean and Processing of Korean Conference", pages 137-141.

Sager, J.C. 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins Publishing Company.

Se-Young Park, Gil-Rok Oh. 1999. *Machine Translation in Korea*, In "Proceedings of the 7th MT Summit", pages 100-104.