Customizing a Korean-English MT System for Patent Translation

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

Abstract

This paper addresses a customization process of a Korean-English MT system for patent translation. The major customization steps include terminology construction, linguistic study, and the modification of the existing analysis- and generation-module. To our knowledge, this is the first worth-mentioning large-scale customization effort of an MT system for Korean and English. This research was performed under the auspices of the MIC (Ministry of Information and Communication) of Korean government. A prototype patent MT system for electronics domain was installed and is being tested in the Korean Intellectual Property Office.

1 Introduction

While the beginning of the MT-research can be traced back as early as to 1950s in the United States and to 1980s in Europe and Japan, MT has not been a major research area or a business sector in Korea until 1990s. However, with the widespread of the broadband Internet since mid 1990s, MT has been put into use in several applications and gained recognition as well as notoriety among general users in Korea. As to the Korean-Japanese MT, the users showed relatively high satisfaction in both translation directions. On the contrary, Korean to English MT has been rarely used in serious applications due to its low translation accuracy. Some reasons can be found for the comparably poor translation performance for Korean-English MT. The biggest reason lies in the difficulty of Korean syntactic analysis. Though intensive research has been made on Korean syntactic analysis, there still remain many issues to be tackled. However, in our opinion, one of the biggest reasons is that few noteworthy customizations, if any, have been made for general domain Korean-English MT engines yet.

Recently, the natural language processing of intellectual property documents is attracting many researchers and NLP-related companies. Especially, the multilinguality of patent documents has become a hot research issue in the IR community. The importance of the creation and the dissemination of multilingual patent documents also seems to be gaining much attention from MT community because of its importance and the economic impact. (Kobayashi, 2005)

ETRI (Electronics and Telecommunications Research Institute) has been developing Korean to English patent MT system "FromTo" under the auspices of the MIC since 2004. Last year, a prototype patent MT system for electronics domain was developed and is being tested in the KIPO (Korean Intellectual Property Office). This year, the target of the research was extended to the whole domain of the patents. At the end of this year the development of a Korean-English patent MT system for whole patent domain is scheduled to be finished. Next year, the KIPO is expected to launch its first Korean-English MT service for public users.

This paper addresses the customization process of a Korean-English MT system for patent translation and the related issues. The Korean to English patent MT system "FromTo" is based on a pattern-based Korean-English MT system developed for the web translation for general domain.

Section 2 describes some characteristics of Korean patent documents. In section 3 we will introduce the customization process and share the experiences. The customization was focussed on the linguistic resources, morphological- and syntactic analyzer, and the generator. Section 4 shows the evaluation result of the system. In section 5 we sum up the discussion and present the future research direction.

2 Some Characteristics of Korean Patent Documents

2.1 Long Sentences

It is generally recognized that the patent domain features overwhelmingly long and complex sentences and peculiar style. An analysis of 5-year

Korean patent documents shows that a sentence is composed of 18.45 Eojeols [1] on the average, compared with 12.3 Eojeols in general Korean newspaper articles. Long sentences in the patent documents show the following characteristics:

- Frequent use of the connective endings
- Frequent use of the conjunctions

Extremely long sentences are often found in the detailed description part of a patent. A patent applicant may try to describe his or her invention in full in the detailed description section. A long sentence usually consists of several simple sentences connected with a verbal connective ending.

In long sentences we found not only many verbs with connective endings but also many long NPs. The long NPs are generally an NP connected with the conjunctions like "와 (and)", "및 (and)", "그리고 (and)", "혹은 (or)", and "또는 (or)". These long NPs are often found in the abstracts and the claims of patents.

Such long sentences are one of the biggest obstacles that affect the translation accuracy, because a parser generally has difficulties in parsing such long sentences, which makes the readability of the translated sentences quite low. We use some syntactic clues to partition a long sentence into several "proper sized" sentences so that each of which can be parsed easily. The clues for partitioning a long sentence are exemplified below:

Clue 1: verbal ending morphemes followed by "," such as:

- "verb stem + 지만 (but) + comma"
- "verb stem + 고 (and) + comma"
- "verb stem + 르때 (when) + comma"

Clue 2: conjunctions and specific lexical tokens in NPs such as:

- NP1 와(and); NP2 와(and); … NPn 을 포함하여(including) 이루어진다(be composed of) ➔ It is composed of the following: NP1, NP2, …, NPn

2.2 Sentence- and Phrasal Patterns

The patent documents have their peculiar styles that are widely accepted in the patent offices. In particular, patent claims are formulated according to a set of precise syntactic, lexical and stylistic guidelines (Sheremetyeva, 2003). After the linguistic study of 1,000 sample Korean patent documents, we extracted sentence patterns based on certain syntactic, lexical and stylistic features. In each section of a patent document, there are different types of sentence patterns as follows:

Abstract: the introduction about the invention is described in a specific form like:

- 본 발명은 ~에 관한 것이다: the present invention relates to ~
- 본 발명은 ~를 개시한다: the present invention discloses ~

Detailed description of the invention: the idiomatic adverb phrases are frequently repeated:

- 종래 기술에 따르면…: according to prior art …
- 도 n 에 도시되어 있는 바와 같이…: as shown in Fig. N …

Brief descriptions of the drawing: it is mainly composed of noun phrases that explain the drawings:

- 도 n 은 …를 설명하기 위한 도면: Fig. n is a view for explaining …

The effect of the invention: the sentences are mainly for explaining the effects of the invention:

- 본 발명은 …는 효과가 있다: the present invention has the effect that …

Claims : it is mainly composed of noun phrases that describe the patent claims:

- 제 n 항에 있어서, …를 더 포함하는 것을 특징으로 하는…: … of claim n, further comprising …

3 Customization Process

3.1 Setting Lexical Goals

At the beginning stage of the project we set the lexical goals. To set the lexical goals, there are some approaches, citing the terms introduced in (Dillinger, 2001), "market approach", "resource approach", and "sample approach". As this is the first Korean-English patent MT system, there is no comparable MT system to assess the lexical coverage of our system. The resource approach is not suitable for our case, as there is no complete list of words or glossary for electronics terms used in patent documents.

To estimate the number of the terms to include in the term dictionary, we analyzed a patent corpus

[1] An Eojeol is a spacing unit. It corresponds to a bunsetsu in Japanese.

of the size of 340 MB.[2] It corresponds to the size of the patent documents for 9 months in the electronics domain. The corpus consisted of 22,756 patent documents that contained 2,667,198 sentences. Given the limited time and budget for the lexical resource construction, we examined the expected lexical coverage in two steps: the coverage of the single noun terms and the coverage of the compound noun terms. To the most single noun terms was given the priority of inclusion in the term dictionary. As to the compound noun terms, the priority was given only to the terms with high frequency.

The number of the newly found unknown single word terms seems to converge after constructing about 130,000 single word terms. As to the newly found unknown multi-word terms, there seems to be no converging point.[3]

Given the above estimation, we decided to construct at least 130,000 single word terms and the multi-word terms with high frequency as our budget allows.

3.2 Building Lexical Resources

The next step is to translate the extracted unknown terms to Korean. 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 term dictionary was the frequent use of parentheses after a Korean technical term. The following sentence shows the characteristic:

> 접철이 가능한 플립-다운형(Flip-down type) 카텔레비젼(100)이며, 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 about 420,000 Korean-English pairs. Domain experts simply accepted or rejected the extracted translation pairs. Using this methodology, we could build about 250,000 entries.

Another valuable resource for the semi-automatic term dictionary construction was bilingual patent title corpus. Even 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 build about 100,000 entries from bilingual corpus relatively easily. Korean and English compound nouns were aligned using POS tagged results, common dictionary and the available term dictionary.

After applying all these semi-automatic methods, human translation of the rest extracted terms was performed as a last recourse. Putting all together, we could build about 600,000 entries for electronics domain with the given budget, thus far exceeding the goals set initially.

The following figure summarizes the described terminology construction workflow:

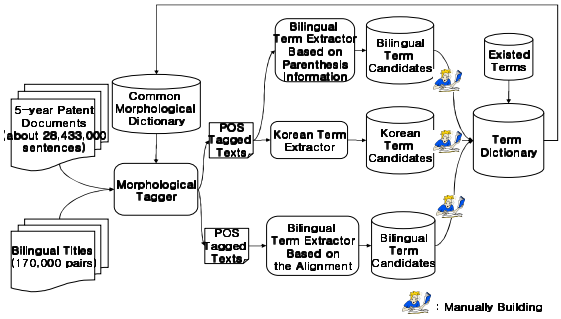

Fig. 1: Terminology construction workflow

3.3 Customization of POS Tagger

To cope with the morphological peculiarity of patent documents, a general morphological analyzer for Korean was modified with respect to the following two characteristics: Frequent use of derived words and the biased POS tagging tendency of ambiguous words. The proper treatment of a prefix and suffix is especially important in the morphological processing of patent documents, as many technical terms are derived nouns. To filter out POS candidates during the POS tagging, we designed a rule table which shows the connectivity between the morphemes in an eojeol.

As was the case with the sentence- and phrasal patterns, there are many eojeols that occur repeatedly in almost all patents. The eojeols are inherently ambiguous w.r.t. their POS. However, as they are used for one meaning in almost every case, by simply fixing their POS without calculating the probability, the tagging accuracy can be improved. For example, the following

[2] Images in the patents were removed during the preprocessing of the corpus.

[3] About the detailed information about the coverage estimation, please refer to "Terminology Construction Workflow for Korean-English Patent MT" to be presented at the workshop on patent translation at the 10th MT Summit.

expression "본 발명 (the present invention)" is actually ambiguous w.r.t. its POS. The one possible tagging result is "보 (to see)/verb + ㄴ/prenominal ending + 발명 (invention)/noun", the interpretation of which is "*the invention that I saw before*". The other possible tagging result is "본 (present)/prenominal modifier + 발명 (invention)/noun", which means "*the current invention*". The first interpretation is possible in a general domain, while only the second interpretation is possible in patent documents. We found actually no case where it was interpreted as in the first POS tagging in patents. We have found about a hundred such expressions. Although the number of such expressions is not high, its impact on the tagging accuracy is not small, because they are frequently used ambiguous words. The POS tagger for patent documents shows 98.7% tagging accuracy.

3.4 Customization of Syntactic Analyzer

The Korean syntactic analysis consists of two steps:

- Predicate-argument-adjunct structure analysis for each predicate using the so-called verb patterns
- Structure analysis between predicates employing predicate-predicate structure patterns.

When analyzing the predicate-argument-adjunct structure, each verb on a dependency tree is compared with verb patterns. Verb patterns contain the information not only about arguments of the predicates but also about the adjuncts. In case of a prenominal clause, the modifee of the clause can fill in an argument position or an adjunct of the predicate. In such cases, the prenominal clause including the modifee is compared with verb patterns. When more than one verb patterns are matched to them, each verb pattern is evaluated according to the below criteria:

- the number of the matched arguments
- the number of the mismatched arguments
- the locality of the verb patterns

Verb patterns with more matched arguments and smaller mismatched arguments score higher points. Also, the verb patterns with shorter distance are preferred to those with longer distance.

Predicate-predicate structure analysis determines the structure between predicates in a sentence. To determine the syntactic structure between the predicates, the semantic relation between them needs to be explained. It requires an amount of semantic information, for which a processing in a deeper level and rich linguistic resource are needed. Thus we are considering only statistical information between verbal connective endings. Connective endings relate the predicates in question to specific discourse relationship such as cause, reason, expectation, or condition. From this reason, we conclude that connective endings can play a role as an indicator about semantic relation information.

With respect to syntactic analysis, the patent domain shows several peculiarities compared with the texts in a general domain.

Firstly, the frequency of topic markers is relatively low.[4] A topic marker distinguishes a topic noun. The NP with a topic marker is usually underspecified w.r.t. its case. The authors of parent documents generally try to describe their invention as explicit as possible and to avoid the misunderstandings as much as possible. A frequent use of a topic marker may cause a misunderstanding of the content or at least make it more difficult to understand the content. From this reason topic markers are seldom used. When they are used, their cases are limited to nominative or accusative cases in most cases. It is valuable information that can help to improve the accuracy of the syntactic analysis.

Secondly, adverbs are also not frequently used and its vocabulary is limited. Adverbs pose serious problems in the syntactic analysis because of their characteristic to modify several syntactic categories. But in the patent translation setting, it does not seem that we have to pay much attention to the treatment of adverbs.

Thirdly, unknown predicates are often encountered. Even if we set the correct lexical goals and construct the term dictionary, we cannot cover all the possible predicates in patent documents. The unknown predicates belong to an open class like technical terms. For unknown technical predicates there is no information about their valency. But, deeper investigation of many unknown predicates revealed that the locality can be a good clue to the solution. The application of the locality clue to the syntactic analysis is also satisfying.

Fourthly, noun phrases show very complex structures. When testing the performance of a syntactic analyzer for general purpose, it showed the worst performance in dealing with the noun phrases. The complexity of noun phrases in patent domain is mainly attributed to the coordination and technical terms. From this reason, compared with

[4] 'nun' is a representative topic marker in Korean. It corresponds to the Japanese topic marker 'wa'.

the general domain, the window size for coordination detection needed to be extended. For the technical terms, we use lexical and/or structural information. Especially, suffix and prefix information is quite useful and effective. Nonetheless, it's still a very difficult task and has its limits. We believe that certain kind of ontology or any form of semantic information is urgent for each technical domain to cope with NP analysis properly.

After adapting the syntactic analyzer to the patent domain, the accuracy of dependency detection rose from 87.4% to 93.4%.

3.5 Customization of Generation Module

This section introduces three major customization efforts with regard to the generation module: introducing the sentence patterns, customizing the link patterns and customizing the word sense disambiguation module.

- Introducing the sentence patterns

The sentences in the patent documents are described with the specific style and words that ordinary people find difficult to read and understand (Shinmori et al., 2003). For the proper treatment of such peculiar sentences, we employed a sentence pattern matching algorithm. After the morphological analysis and noun phrase chunking, tokenized words of an input sentence are matched with the pre-compiled tokens of sentence patterns. The sentence patterns are pre-compiled using a morphological analyzer and a noun phrase chunker, and are stored as forms of morphological tokens. The compound nouns that have no structural ambiguity are chunked to raise the matching coverage of sentence patterns. The sentence patterns are especially useful in translating the "detailed description of the drawings", "the effects of the invention" and "the claims" of patent documents.

- Customizing link patterns

As mentioned in previous chapters, entremely long sentences are often found in patent documents. A long sentence usually consists of several simple sentences, each of which is connected with a connective ending. The relationship between the connective endings, which roughly correspond to conjunctions in English, is directly encoded in the so-called link pattern. The investigation of the patent documents revealed that there are many complex connective endings that are generally not accepted as a connective ending in the tranditional Korean grammar. For example, "ㄴ것과_같이" is a complex connective ending composed of a prenominal ending ("ㄴ"), a dependent-noun ("것"), and a complex postposition ("과같이"). It is a fixed expression which corresponds to "as" in English, is often used to combine two simple sentences in patent documents. We collected the frequently used complex connective endings in patent documents such as "려는것인바", "ㄴ경우에있어서".

As a connective ending may have several semantic roles, no 1-to-1 mapping between a Korean connective ending and an English conjunction can be made. Furthermore, because of the big structural and stylistic difference between two languages, a simple combination of the English verbal phrases corresponding to the Korean verbal phrases cannot produce the correct English translations.

The verbal phrase linker in our system solves this problem using link patterns. These patterns contain the generation information such as the relative order of English verbal phrases, the correct English conjunction, and the syntactic information of the phrases for generation. Figure 1 shows an example of the link patterns.

On the analyzed parse tree, the verbal phrase linker detects the dependency relation among verbal phrases. A matched link pattern provides the English generation information to the verbal phrase generator. Using this generation information, the verbal phrase generator transfers the proper English expressions corresponding to Korean verbal phrases.

```
[KEY]
ㄴ것과같이:2_다:1
[CONTENT]
{ VP1[] VP2[] >
VP2 VP1[SCONJ:[eroot := [as]]
VERB:[eform := [sprp]]] }

/*도 2 에 도시된 것과 같이 전극이 설치되어 있다. */
/* As shown in Figure 2, the electrode is set up. */
```

Fig. 2: A link pattern "ㄴ것과같이:2_다:1"

- Customizing Word Sense Disambiguation

When selecting a proper target word, domain-specific lexical and semantic information within certain local syntactic relations is employed. The details about the target word selection algorithm are described in (Kim et al., 2004). Many

ambiguous words incline to have a specific meaning in a certain domain, as pointed out in (Streiter et al., 1999). For example, a Korean word "자극" has two different meanings: "magnetic pole" and "stimulus". In the electronics domain, "자극" is mostly used as "magnetic pole", while in the medical domain, it is mostly used as "stimulus". We have constructed the DB of the domain-specific semantic and lexical co-occurrence information from semantically tagged sentences for such ambiguous words in the electronics domain, and are now constructing DBs for every patent domain. The semantically tagged corpus was constructed for 1,000 most frequent ambiguous words in electronics domain. For every word 100 sentences were extracted and semantically tagged by domain experts. If an input sentence contains an ambiguous word, its context is considered. If an exactly matching context is found in the DB, its semantics is calculated according to the DB information. Otherwise, the most widely used target word for the ambiuous word in the domain is selected.

4 Evaluation

The goal of the evaluation was to see:

i) how accurate does the system deliver the meaning of the source language? (accuracy)
ii) how natural do the users find the translation? (understandability)

Translation accuracy was assessed with 200 test sentences randomly extracted from patent corpus. The test set was organized in such a way that it reflects a real patent document. Among 200 sentences, about 120 sentences were selected from the "detailed description" section of patents, 40 were extracted from the "claim" section, the rest from the "description of the drawing" and the "effects of the invention" section. The average length of a sentence was 23.7 words. The length was normalized in order to reflect the length of the real patent sentences. The accuracy was scored according to the following criteria:

Score	Criterion
4	The meaning of a sentence is perfectly conveyed
3.5	The meaning of a sentence is almost perfectly conveyed except for some minor errors (e.g. wrong article)
3	The meaning of a sentence is almost conveyed (e.g. some errors in target word selection)
2.5	A simple sentence in a complex sentence is correctly translated
2	A sentence is translated phrase-wise
1	Only some words are translated
0	No translation

Table 1: Accuracy scoring criteria

Six professsional translators were hired for the evaluation. They were all native Korean speakers who translate Korean technical documents to English professionally. To make the judgment as fair as possible, we ruled out the highest and the lowest score for every sentence. That is, if 4 evaluators gave 3 points, 1 gave 4 points, and the 1 gave 2 points for a translation, the highest and the lowest score 4 and 2 were ruled out for the summation. In this way the scores for each sentence were summed. The translation accuracy was 79.51%. The number of the sentences that were rated equal to or higher than 3 points was 132. It means that about 66% of all translations were understandable.

On the contrary to the accuracy evaluation, the style of the translation was evaluated by 2 English native speakers. Both of them were US patent experts. As a test suite, 2 patent documents were randomly selected. The style or the understandability of the translation was evaluated according to the following criteria:

Score	Criterion
4	I can understand the sentence after reading it just once. The sentence contains almost no error and is natural.
3	I can understand the sentence. But to understand it, I need to read it a couple of times. The sentence contains some (non-critical) errors.
2	The sentence is ungrammatical. I can understand it only partly. (phrase-wise) It is not so difficult to guess what it is about, because some translated chunks deliver meaningful information, even though they are ungrammatical.
1	The sentence delivers almost no information but some word-to word translations. I can only guess what it is about due to some word translations
0	It's useless

Table 2: Understandability scoring criteria

The evaluation result for understandability was somewhat lower than the accuracy (71%). One of the most striking differences between the perspectives of the Korean evaluators and the American evaluators was their opinions about the translations which are grammatically correct but resemble the structure of Korean source sentence too much. Most Korean evaluators did not find them problematic, but the American evaluators found them awkward.

Among the sections in patent documents, the translation of the description of the drawings part was best in both the accuracy and the understandability evaluation, while the translation of the detailed description part scored worst. The difference between them was about 4%. The reason for the best scoring of the description of the drawing section is that to the most sentences were applied sentence patterns. The detailed description part contained, as expected, many long sentences. In some long sentences, the wrong analysis results lead to the poor translation.

5 Conclusion

In this paper we presented the customization process of a general Korean-English MT system for patent translation. Setting the lexical goals and the linguistic study of patent documents belong to the first and probably the most important steps of the customization. When constructing the lexical resources for the patent translation, semi-automatic methods were employed to reduce the time and the cost. The customization of each engine module was performed based on the linguistic study of the patent documents. The Korean-English patent MT system for electronics domain was installed and is being tested at KIPO. The extension of the lexical resource to all areas in the patents as well as the improvement of each engine module is made in this year. KIPO is expected to launch the MT service for all patents at the end of next year.

To improve the translation quality, not only the improvement of the MT engine itself but also the controlling of the authoring of patent documents is very important. To our surprise, we often encounter many syntactically as well as stylistically awkward sentences in patent documents. To solve the problems we are planning to introduce the concept of controlled language in authoring a patent in cooperation with KIPO from next year. The development of the controlled language checker for patent documents will be our next research topic.

References

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

A. Kobayashi. 2004. *Machine Translation and Japio's role in disseminating Japanese information*. http://www.european-patent-office.org/epidos/conf/jpinfo/2004/_pdf/pres/japio_kobayashi_machine_translation_and_japio_role.pdf

Y. Kim, M. Hong, C. Kim, S. Park. 2004. *Word Sense Disambiguation Using Lexical and Semantic Information within Local Syntactic Relations.* In "Proceedings of the 30th Annual Conference of the IEEE Industrial Electronics Society".

S. Sheremetyeva. 2003. *Natural Language Analysis of Patent Claims.* In "Proceedings of ACL 2003 Workshop on Patent Corpus Processing Workshop".

A. Shinmori, M. Okumura, Y. Marukawa, M. Iwayama. 2003. *Patent Claim Processing for Readability.* In "Proceedings of ACL 2003 Workshop on Patent Corpus Processing Workshop".

O. Streiter, L. Iomdin, M. Hong, U. Hauck. 1999. *Statistical Support for Rule-Based MT*. In "Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)".