English-Korean Patent Translation System: FromTo-EK/PAT

Oh-Woog Kwon, Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, Young-Gil Kim

Natural Language Processing Team, Electronics and Telecommunications Research Institute
161 Gajeong-dong, Yuseong-gu,
Daejeon, Korea, 305-350
{ohwoog, choisk, leeky, yhnoh, kimyk}@etri.re.kr

Munpyo Hong
Dept. Of German Language & Literature Sungkyunkwan Univ.
53 Myeongnyun-dong 3-ga, Jongno-gu, Seoul, Korea, 110-745
skkhmp@skku.edu

Abstract

This paper addresses a method for customizing an English-Korean machine translation system from general domain to patent domain. The customizing method includes the followings: (1) extracting and constructing large bilingual terminology and the patent-specific translation patterns, (2) adapting the probabilities of POS tagger trained from general domain to the patent domain, (3) syntactically analyzing long and complex sentences by recognizing coordinate structures, and (4) selecting a proper target word using patent-specific bilingual dictionary and collocation knowledge extracted from patent corpus.

The translation accuracy of the customized English-Korean patent translation system is 82.43% on the average in 5 patent categories (machinery, electronics, chemistry, medicine and computer) according to the evaluation of 7 professional patent translators. A patent MT system for electronics domain was installed and started an on-line MT service in IPAC (International Patent Assistance Center) under MOCIE (Ministry of Commerce, Industry and Energy) in Korea. In 2007, KIPO (Korean Intellectual Property Office) is expected to launch its English-Korean MT service for whole patent domain.

1. Introduction

Given the growing number of foreign language patents filed in the multiple countries, it is feasible that users want to read the patent documents translated to their native language. Such users' demand has become a hot research issue in the MT community. Also because NLP techniques associated with specificity of patent domain have promise for improving the translation quality, patent translation is recently attracting many researchers and MT-related companies.

It is well known that sentence style and dominant translation for a word vary with domains. Therefore, if the domain to be translated is fixed to patents, bilingual dictionary adaptation to the patent domain and customizing natural language analyzers to the linguistic specificity of patent style are effective ways to improve the translation quality of MT system. There have been studies concerned specifically with patent MT using these domain-specific advantages (Shinmori et al., 2003; Hong et al., 2005; Kaji, 2005; Shimihata, 2005).

Though intensive research has been made on patent MT for the domain-specific advantages, there still remain many issues to be tackled. In this paper, we focus on the several issues: (1) new terminology construction, (2) patent-specific probabilities of POS tagger, (3) long and complex sentence analysis, and (4) target word selection.

This paper addresses the customization of an English-Korean MT system for patent translation. The English-Korean patent MT system "FromTo-EK/PAT" described in this paper is based on an English-Korean MT system developed for the web translation in a general domain. English-Korean patent MT system belongs to basically the pattern-based methodology for machine translation. It has the formalism that does English sentence analysis in which English patent-specific patterns are used, matches the English patent pattern with its Korean patent pattern, and then generates a Korean sentence from it. English-Korean patent MT system consists of an English morphological analysis module based on lexicalized HMM, an English syntactic analysis module by pattern-based full parsing, a pattern-based transfer, and a Korean morphological generation.

Section 2 describes the issues of customizing a MT system to the patent domain. In section 3 we will introduce the customization process according to the issues described in section 2. The experimental work is presented in section 4. Lastly, in section 5, we present some conclusions.

2. Issues for Customizing MT System to Patent Domain

It is important to customize translation knowledge and translation modules for adapting the existing general MT system to translation of patent documents. The customization for the translation knowledge is able to be divided into two steps: (1) tuning general translation knowledge to patent-specific translation knowledge, and (2) efficiently constructing the unknown words and the translation patterns found in patent documents. The patent customization of existing translation knowledge is closely related with the customization of the translation knowledge of module. For example, the customization of the module of target word selection is decided by the customization of existing English-Korean bilingual dictionary. The POS tagging knowledge trained from general domain also have an influence on the customization of the POS tagging module. In this respect we consider the method extracting unknown words from patent documents and the method customizing translation modules to patent.

What is firstly necessary for applying a general MT system to patent is to extract the large-scale terms found newly in patent documents and construct their translation knowledge such as the target words. We have built an English-Korean bilingual dictionary by use of exiting Korean-English bilingual dictionary of a Korean-English patent MT system developed in 2005, in order to cut cost and time for building an English-Korean bilingual dictionary. The unknown words could be constructed at maximum effect with little cost and little time by the method, where we preferred selecting the high-frequently and positively necessary words for the English-Korean translation to constructing all unknown words appearing in patent documents.

In relation to POS taggers with good performance and broad coverage, they have recently become available (Brants, 2000; Pla et al., 2004), but have not been trained for patent documents. This means that there is room for doubt that the general POS taggers keep their performance in the patent domain. We can easily find an example to degrade the performance, only looking through any patent document. The example is the word "said": the word is mainly used as a past verb (VBD) in general domain, but is almost used as a adjective (JJ) in patent domain. The words like "said" are retrained from a tagged patent corpus. It is however very difficult to construct the tagged patent corpus because we have no tagged patent corpus. In this paper, we will describe how to adapt the general-purpose POS tagger to the patent domain by using raw patent corpus.

Compared with general documents, one characteristic of patent documents is to use the abnormally long and complex sentences (Kando, 2000), which makes it difficult to apply a parser for general domain to patent domain. A usual method for treating long sentences is to segment a long sentence into several segments and to analyze each segment respectively. However, in case a long sentence is formed by coordination structure, simple segmentation can cause syntactic analysis errors if the coordination structure is not firstly recognized. For this, we will present a method for recognizing the coordination structure in patent documents to enhance parsing efficiency and performance.

Target word selection in English-Korean machine translation is very important factor in that it has a direct influence on the machine translation quality. Particularly, in the case of general domain documents such as web pages, the target word selection problems of English ambiguous words occur very frequently. In general domain documents, many frequently used English words can be translated to various Korean words depending on the contexts. However, in English-Korean patent machine translation, most of words used in patent documents belong to technical terms. These technical terms have relatively low ambiguities of target word selection. Some English words used in patent domain also have a tendency to be translated to specific Korean word according to International Patent Classification (IPC) codes. Although patent documents include many technical terms, target word selection problem still remains an obstacle which should be solved to improve the performance of machine translation system. We customized English-Korean dictionary for patent machine translation to resolve the translation ambiguity of English ambiguous words appearing in patent documents. So, some English ambiguous words contain dominant Korean target word according to specific IPC code. For target word selection ambiguities which did not resolved by dominant Korean target word of translation dictionary, we tried to disambiguate the possible senses of English words by use of other knowledge like sense vectors and Korean bi-gram context information.

3. Customizing Methods

3.1 Construction of Patent Terminology

Terminology construction for English-Korean patent MT system described in this paper is similar to the methods of Kaji(2005), Shimohata(2005), and Kim(2005) in respect of using the existing dictionary and the existing patent corpus, but our method is different in that it contains a step inverting the existing Korean-English bilingual terminology. Extraction and construction of terminology might be represented in Figure 1.

As shown in Figure 1, the patent terminology can be built by two steps. The first step is the step to convert the existing Korean-English terms into the English-Korean terms, to delete the terms overlapped with the terms in the existing English-Korean bilingual dictionary, and to construct the English-Korean bilingual terms semi-automatically. Among inverted English-Korean bilingual terms, if English terms are the nominal phrases including a prepositional phrase, a gerund, and a relative clause, they are deleted. These nominal phrases were constructed for lack of an English compound word suitable to a Korean compound word in Korean-English patent translation. If such nominal phrases are entered in the English-Korean dictionary, the structural errors such as attachment of prepositional phrase or analysis of coordination structure in parsing might be produced. For example, if "method for 1+1 line protecting switching" as an English term equivalent to Korean term "1+1 선로 보호 절체 방법" is made an entry of English-Korean dictionary, it may give rise to the incorrect analysis of coordination structure "(NP (NN device) (CC and) (NN method for 1+1 line protecting switching))" in analysis of a English phrase such as "device and method for 1+1 line protecting switching".

Each English term in the English-Korean terms constructed by the first step may have different Korean target words. To select a dominant one among different Korean target words, we sorted Korean target words automatically according to their frequency occurring in Korean patent documents and made a selection of dominant target word manually. Through this work we could create 801,046 English-Korean terms from 3,052655 Korean-English terms.

The second step is to extract the unknown words from 1,001,419 English patent documents applied to the U.S. Patent Office from 2001 to 2005 and remove the overlapped entries. We extracted about ten million English unknown words from this step, but manually constructed 1,039,189 English-Korean bilingual terminology with high coverage by using the method ‘Setting Lexical Goals’ Hong(2005) presented.

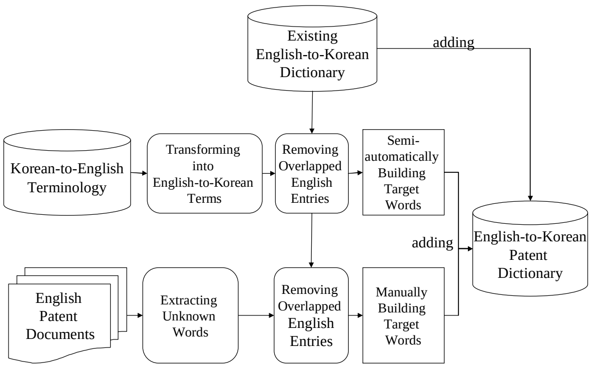
Existing English-to-Korean Dictionary
adding
Korean-to-English Terminology
Transforming into English-to-Korean Terms
Removing Overlapped English Entries
Semi-automatically Building Target Words
adding
English-to-Korean Patent Dictionary
English Patent Documents
Extracting Unknown Words
Removing Overlapped English Entries
Manually Building Target Words

Figure 1: Customization process for building English-Korean patent terminology

3.2 A Domain Adaptation Method for POS Tagger

Three items were tuned for customizing a broad coverage POS tagger based on HMM to patent domain. They are as follows:

- For customization of surface form, a tokenization module and/or a morphological analyzer were modified for tokenizing and/or analyzing the peculiar surface forms found in the specific domain.
- For customization of lexical information, lexical probabilities (output probabilities) were tuned for holding domain-specific lexical information.
- For customization of context information, contextual probabilities (transition probabilities) were controlled for holding the domain-specific contextual information.

In the first step ‘customization of surface form’, the tokenization module was modified to tokenize and/or chunk very complex symbol words, a chemical formula, a mathematical formula, programming codes, and so on. We improved our morphological analyzer to assign the estimated part-of-speeches to a compound word connected with hyphen or slash. The estimated part-of-speeches are estimated by the part-of-speeches of their components.

Our English POS tagger uses a lexicalized HMM (Pla et al., 2004). The process of our POS tagger consists of finding the sequence of POS tags of maximum probability, that is:

$$\overline{T} = \arg\max_{t_1\cdots t_n}\left(\prod_{1\ldots n} P(t_i|t_{i-1},t_{i-2})\cdot P(w_i|t_i)\right) \quad (1)$$

for given sequence of words $w_1, \ldots, w_n$ of length n. $t_1, \ldots, t_n$ are elements of the tagset, the additional tags t_{-1}, t_0, and t_{n+1} are beginning-of-sequence and end-of-sequence markers. In this equation, lexical probability is $P(w_i|t_i)$, and contextual probability is $P(t_i|t_{i-1},t_{i-2})$. The lexical and contextual probabilities are estimated from tagged corpus.

The best simple strategy for the second and third customization phase is to re-estimate lexical and contextual probabilities from very large tagged patent corpus. However, there is not a tagged patent corpus and it is also very difficult to construct it. For customizing the lexical and contextual probabilities, we used a raw patent corpus consisting of about one million U.S. patent documents. First, we tagged automatically the words of the raw corpus with our POS tagger and estimated lexical probability $P'(w_i|t_i)$ and contextual probability $P'(t_i|t_{i-1},t_{i-2})$ from the machine-tagged patent corpus. Next, we extracted the high-frequent lexemes having $abs(P(w_i|t_i)-P'(w_i|t_i))$ greater than arbitrary threshold value and the high-frequent contextual n-grams having $P(t_i|t_{i-1},t_{i-2})$ less than arbitrary threshold value. The extracted lexical and contextual n-grams are tuned by the three human experts for two months. For customization of our POS tagger, we tuned about 6,000 lexemes and about 1,500 tri-grams.

It is difficult that a expert perceives the exact meaning of the output probability, because lexical probability, $P(w_i|t_i)$, corresponds to the output probability in which the word w_i is generated given POS t_i. But, the expert could easily decide whether a word w_i is used as POS t_p more frequently than POS t_q in the patents, or not. In this view point, the expert can more easily and correctly tune $P(t_i|w_i)$ than $P(w_i|t_i)$ for each extracted word w_i. To customize lexical probabilities to patent domain, the experts adjusted $P(t_i|w_i)$ examining the POS tagged sample sentences. Then, we calculated $P(w_i|t_i)$ by using the tuned $P(t_i|w_i)$ as follows:

$$P(w_i|t_i) = P(t_i|w_i)\times f(w_i)/f(t_i) \quad (2)$$

For customization of the context information, the experts selected correct n-grams from the extracted n-grams. To estimate the selected context probabilities $P(t_i|t_{i-1},t_{i-2})$, we first find $P'(t_p|t_{p-1},t_{p-2})$ that is the nearest probability to $P'(t_i|t_{i-1},t_{i-2})$. Then we calculated $P(t_i|t_{i-1},t_{i-2})$ as follows:

$$P(t_i|t_{i-1},t_{i-2}) = P'(t_i|t_{i-1},t_{i-2})\times\frac{P(t_q|t_{q-1},t_{q-2})}{P'(t_q|t_{q-1},t_{q-2})} \quad (3)$$

The representative tri-grams among the extracted n-gram are “NN CD VBZ” and “NNS CD VBP”. They mean that a cardinal number comes before a verb in patent documents, while a cardinal number basically comes before a noun in general documents. In the patent documents, a cardinal number after a noun denotes almost always a reference mark for a diagram or a box in a figure. For example, in the sentence “Another management chip connected to pad 117 controls the parallel port 102b and the serial ports 104c and 104d.”, the cardinal number “117” points out the box corresponding to the pad apparatus in a figure.

3.3 Syntactic Analysis for Patent Document

Two most important ones among peculiar syntactic characteristics of patent documents are the frequent use of patent-specific patterns and the abnormally long sentences (Shinmori et al., 2003). Considering these characteristics as central features, I will describe the main contents of syntax analysis for patent documents in detail.

3.3.1 Application of patent-specific patterns

We applied patent-specific patterns before parsing to reduce a parsing complexity. A general form of the patent-specific patterns is composed of some lexical words and some syntactic nodes as shown in a sample of below pattern.

1) *The method for VP , wherein S*

For the recognition of the patterns, lexical words are firstly matched, and the ranges between the lexical words are recognized as tentative syntactic nodes. Assuming that above pattern is applied to a example sentence 2), "the method for" is matched, the word strings between "for" and "," are recognized as a verbal phrase(VP) and the matching of next lexical symbols ", wherein" is attempted.

> 2) "*The method for controlling the flow in the micro system according to claim 1, wherein the stimulation is a voltage.*"

Actually, we conduct simple condition check to know whether the word strings can be VP or not. If the pattern matches wholly with the input sentence, a parsing with all the tentative nodes is attempted. If all nodes are successfully parsed into the corresponding syntactic nodes in the translation pattern, the syntactic pattern is recognized finally. As a result, the actual parsing ranges are reduced to parsing of two clauses such as "controlling the flow in the micro system according to claim 1" and "the stimulation is a voltage".

3.3.2 Recognizing coordinate construction

The usual method for treating long sentences is to segment a long sentence into several segments by use of syntactic clues or some other conditions (Kim et al., 2001). However, the segmentation method is applicable only in case that segments resulting from segmentation don't have any hierarchical relation between each other. In case of sentences formed by coordination of syntactic nodes such as NP, VP, that-clause, etc., if a sentence is segmented between coordinate constituent nodes, segmentation can cause syntactic analysis errors, because a segment can be dependent on some other node in parse tree.

For example, in the example sentence 3), the sentence can be segmented at the positions such as ", collecting" or ", driving". But verb phrases starting at those positions are objects of the verb "comprising", so such dependency relation is broken by segmentation.

> *3) A method of operating a transaction system which comprises a plurality of currency acceptors, the method comprising installing the acceptors in host machines, performing individual transactions using the machines, collecting performance data from the acceptors, performing a statistical analysis on the performance data from the acceptors, deriving re-configuration data for at least one acceptor as a result of the statistical analysis and re-configuring said at least one acceptor on the basis of the re-configuration data.*

Therefore, we need to recognize coordination structures first before segmentation. Sadao K. and Makoto N. (1994) detected conjunctive structures in a general domain using dynamic programming. Compared with coordinate structures in the general domain, a typical feature of coordination structures in patent documents is that the coordinate structures have a lot of coordinate constituent nodes like VPs in the example sentence 3). Sometimes, each node has very complex structure, which makes the recognition of coordination structure very difficult. So, we have introduced a method of recognizing coordination structure using similarity table. The similarity table is a table which stores similarities between all the possible nodes constituting candidate coordinate structures. All starting positions of possible nodes constituting the candidates of coordination structures are recognized by syntactic clue such as NP or verb followed by "comprise, include, have, etc.". The similarity between nodes is calculated by syntactic similarity and some other factors. Once the similarity table is constructed, all the candidates of coordination structures are searched and their weights are calculated by the similarity table. Finally, the coordinate structure with maximum weight becomes a final result. The sentence is simplified because the recognized coordination construction is chunked to one node. The example sentence 3) is reduced to " A method of operating a transaction system which comprises a plurality of currency acceptors, the method comprising VP."

3.3 Customization for Target Word Selection

We approached target word selection problems in patent machine translation in two ways considering knowledge and engine. For adapting English-Korean bilingual terms to patent domain, we first defined 5 patent categories such as mechanics, chemicals, medicals, electronics and computers and mapped all IPC codes to 5 patent categories. Next, we reconstructed translation dictionary putting the dominant translation word according to 5 patent categories. For this reconstruction process, we made a collection of each 5 patent corpus using a mapping table between IPC codes and 5 categories. And then, we extracted English ambiguous words with high frequency. For these extracted English words, human patent translator registered dominant Korean word by hands considering each category. Our patent machine translation system receives IPC code of an input patent document as a parameter and decides proper Korean target word by it.

For the ambiguous English words which did not resolved by dominant Korean word of translation dictionary, we made a target word selection module using context knowledge constructed from corpus. We extracted context information from English-Korean comparable corpus. The context information was converted to sense vectors. The sense means Korean translation word for the ambiguous English word. The sense vectors were used to disambiguate the possible senses of ambiguous English words (Lee et al., 2006). Sense vector is defined by the following formula:

$$SV = (w(c_1), w(c_2), w(c_3), \ldots, w(c_n)) \quad (4)$$

where $w(c_k)$ is a weighting function for co-occurring word c_k. And $w(c_k)$ can be calculated by the following formula:

$$w(c_k) = \Pr(s = s_i | w = c_k) \quad (5)$$

where s_i is an i-th sense (a group of target words sharing same semantic code) of source word. When $w(c_k)$ is 1, it means that if co-occurring word c_k appears with ambiguous word, the probability that the sense of ambiguous word will be s_i is 1.

In the test phase, the test vector for ambiguous word in input sentence is constructed and has same dimension as the sense vector of the corresponding ambiguous word. The elements of test vector are 0 or 1, where 0 indicates that corresponding co-occurring word c_k does not appear in the input sentence and 1 represents that corresponding co-occurring word c_k appears in the input sentence. The similarity between test vector constructed from input sentence and each sense vector of the ambiguous word is calculated using following formula:

$$sim(v,w) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}} \quad (6)$$

Also, we extracted Korean bi-gram information from Korean monolingual corpus. Korean bi-gram information is used to decide the most proper Korean translation word in final generation phase of our system.

4. Experiments and Evaluation

4.1 Translation Accuracy Evaluation

In this section, we describe the evaluation about translation quality of English-Korean patent MT system. We used the following test sentences, evaluation method and evaluation criterion for translation quality:

- Test sentences: translation accuracy was assessed with 100 test sentences for each one of 5 patent categories (machinery, electronics, chemistry, medicine and computer). Among 100 sentences for each patent category, about 54 sentences were selected from the "detailed description" section of patents, 24 were extracted from the "claim" section, the rest from the "description of the drawing" and the "background of the invention" section. The average length of a sentence was 28.33 words.
- Evaluation criterion:

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, stylistic errors)
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: Scoring criteria for translation accuracy

- Evaluation method:
  - 7 professional translators evaluated the results. Ruling out the highest and the lowest score, the rest 5 scores were used for translation accuracy evaluation. The translation accuracy was defined as follows:
    : translation accuracy(%) = $\sum_{i=1}^{n}(\sum_{j=1}^{5}(score_j/4))/5)/n \times 100.0$, where n is the number of test sentences and $score_j$ is the score evaluated by the j-th professional translator.

Table 2 shows that the translation accuracy of English-Korean patent MT system was 82.43% on the average. Among the patent fields, the translation of the machinery field was best, while the translation of the medicine field scored worst. The reason for the best scoring of the machinery field is that patent-specific patterns were applied to most of sentences. The medicine field contained, as expected, many unknown words and incorrect target word selection. The number of the sentences that were rated equal to or higher than 3 points was 438. It means that about 87.60% of all translations were understandable.

Patent field	Average length of a sentence	Translation accuracy	Translation accuracy higher than 3 scores
machinery	30.34 words	83.50%	85.00%
electronics	29.42 words	82.20%	88.00%
chemistry	29.67 words	82.20%	91.00%
medicine	26.75 words	81.63%	86.00%
computer	25.49 words	82.63%	88.00%
average	28.33 words	82.43%	87.60%

Table 2: Translation accuracy for each patent field

4.2 Evaluation for Customization

We evaluated the performance the modules specialized to the patent domain, compared with the performance of our general-purpose modules. For the evaluation, we used 100 sentences of the electronics category among the whole translation evaluation test set.

Table 3 shows the word accuracy and sentence accuracy of two taggers: the POS tagger specialized to the patent domain (PatTagger) and our general-purpose POS tagger (GPTagger). From these results we can draw the following conclusions. First, the PatTagger reduced significantly the error tagging about 91% with respect to the GPTagger. Second, PatTagger improved the sentence accuracy with 41% compared with GPT3agger.

	GPTagger	PatTagger
Word tagging accuracy	95.85%	99.62%
Sentence tagging accuracy	50.00%	91.00%

Table 3: Comparison of the tagging accuracy between GPTagger and PatTager

Table 4 shows the performance improvement factors of PatTagger and the improved word accuracy according to the factors. The improvement factors of PatTagger are three customization phases mentioned in the section 3.2 and terminology construction mentioned in the section 3.1. The terminology construction is to add unknown words and their part-of-speeches into morphological analysis dictionary. The performance improvement of word supplement is very low because our POS tagger handles unknown words using suffix analysis as proposed in Brants(2000). From the results of table 4, the customization of lexical and context information is surely needed in order to specialize a general-purpose POS tagger based on HMM to a specific domain.

Table 5 shows the evaluation result by the customization of syntactic analyzer. In Table 5, the syntactic analysis accuracy is calculated by the ratio of the number of correctly analyzed sentences to the number of total sentences. We consider a sentence as correct when the syntactic analysis result of the sentence has a trivial error that don't affect the translation result.

Table 6 shows the experimental results of target word selection of the customized MT system and the non-customized MT system. The percentage of unknown word is decreased in customized MT system by registering unknown words to translation dictionary consistently. We can see that how the unknown word can affect target word selection problems. At the same time, the customization of transfer module considering characteristics of patent domain can improve the performance of target word selection.

Table 7 is the result to compare the translation accuracy before customization with that after customization in the electronic patent document. In Table 7, the difference of translation accuracy between before customization and after customization in electronic patent document was 27.95%. This means that the customization process described in this paper made an important role to enhance the translation quality of English-Korean MT system on patent documents.

The performance improvement factor	The # of tagging error correction	The correction rate	The improvement of word tagging accuracy
Customization of surface form analysis	6	5.41 %	0.20%
Customization of the lexical information	81	72.97 %	2.75%
Customization of the context information	22	19.82 %	0.75%
Construction of Terminology	2	1.80 %	0.07%
Total	111	100.00 %	3.77%

Table 4: The performance improvement of PatTagger and the improvement of its word tagging accuracy.

	Syntactic analysis accuracy
General-purpose syntactic analyzer	69.0%
Customized syntactic analyzer	85.0%
ERR (Error Reduction Rate)	51.6%

Table 5: Evaluation of customization of syntactic analyzer

	Accuracy of target word selection for noun	Percentage of unknown word
Non-customized MT System	71.7%	16.3%
Customized MT System	92.4%	1.5%

Table 6: Result of target word selection for noun

Patent field	Translation accuracy before customization	Translation accuracy after customization
electronics	54.25%	82.20%

Table 7: Comparison of translation accuracy before customization with that after customization in electronic patent document

4. Conclusion

In this paper we described a method for customizing English-Korean machine translation system from general domain into patent domain. First, we described the construction method of the large English-Korean bilingual dictionary using the existing Korean-English bilingual dictionary and extracting unknown words from about one million patents. Secondly, to adapt general-purpose POS tagger to the patent domain, we proposed the method for semi-automatically adjusting probabilities trained from general domain to patent context using raw English patent documents. Thirdly, the syntactic analyzer is proposed for segmenting and analyzing long and complex patent sentences by recognizing coordinate structures. Lastly, we proposed the target word selection using patent-specific bilingual dictionary and collocation knowledge extracted from raw patent corpus.

The English-Korean patent MT system "FromTo-EK/PAT described in this paper was developed under the auspices of the MIC (Ministry of Information and Communication, Korea) during 2005-2006. FromTo-EK/PAT was installed in IPAC (International Patent Assistance Center) under MOCIE (Ministry of Commerce, Industry and Energy) in Korea and provides the patent attorneys and the patent examiners with the on-line English-Korean machine translation service for electro-electric patent documents (http://www.ipac.or.kr). In 2007, KIPO (Korean Intellectual Property Office) is expected to launch its English-Korean MT service for whole patent domain.

Acknowledgements

This work was supported by the IT R&D program of MIC/IITA, Domain Customization Machine Translation Technology Development for Korean, Chinese, and English.

Bibliographical References

Brants T. (2000). TnT – a statistical part-of-speech tagger. In Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000) (pp. 224--231).

Hong M.P., Kim Y.G., Kim C.H., Yang S.I., Seo Y.A., Ryu C. & Park S.K. (2005). Customizing a Korean-English MT System for Patent Translation. MT Summit X (pp. 181—187).

Kaji H. (2005). Domain Dependence of Lexical Translation: A Case Study of Patent Abstract. MT Summit X Workshop on Patent Translation.

Kando N. (2000) What Shall we Evaluate? Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and Patent Attorneys. In Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval in conjunction with The 23rd Annual

International ACM SIGIR Conference on Research and Development in Information Retrieval. Athens. Greece

Kim Y.K., Yang S.I., Hong M.P., Kim C.H., Seo Y.A., Ryu C., Park S.K. & Park S.Y. (2005). Terminology Construction Workflow for Korean-English Patent MT. MT Summit X Workshop on Patent Translation.

Lee K.Y., Park S.K. & Kim H.W. (2006). A Method for English-Korean Target Word Selection Using Multiple Knowledge Sources. IEICE TRANS. FUNDAMENTALS, Vol.E89-A, No.6.

Sadao K., Makoto N. (1994). A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structure. Computational Linguistics 20(4): 507-534.

Shinmori A., Okumura M., Marukawa Y. & Iwayama M. (2003). Patent Claim Processing for Readability - Structure Analysis and Term Explanation, ACL-2003 Workshop on Patent Corpus Processing.

Sung-Dong Kim, Byoung-Tak Zhang, and Yung Taek Kim. (2001). Learning-based Intrasentence Segmentation for Efficient Translation of Long Sentences. Machine Translation, 16(3):151-174.

Shimohata S. (2005). Finding Translation Candidates from Patent Corpus. MT Summit X Workshop on Patent Translation.

Pla F. & Molina A. (2004). Improving Part-of-speech Tagging Using Lexicalized HMMs. Natural Language Engineering 10(2) (pp. 167-189).