Customizing an English-Korean Machine Translation System for Patent Translation*

Sung-Kwon Choi, Young-Gil Kim

Natural Language Processing Team, Electronics and Telecommunications Research Institute,
161 Gajeong-dong, Youseong-gu, Daejon, Korea, 305-350
{choisk, kimyk}@etri.re.kr

Abstract. This paper addresses a method for customizing an English-to-Korean machine translation system from general domain to patent domain. The customizing method consists of following steps: 1) linguistically studying about characteristics of patent documents, 2) extracting unknown words from large patent documents and constructing large bilingual terminology, 3) extracting and constructing the patent-specific translation patterns 4) customizing the translation engine modules of the existing general MT system according to linguistic study about characteristics of patent documents, and 5) evaluating the accuracy of translation modules and the translation quality. This research was performed under the auspices of the MIC (Ministry of Information and Communication) of Korean government during 2005-2006. The translation accuracy of the customized English-Korean patent translation system is 82.43% on the average in 5 patent fields (machinery, electronics, chemistry, medicine and computer) according to the evaluation of 7 professional human translators. In 2006, the patent MT system started an on-line patent 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) tries to launch an English-Korean patent MT service.

Keywords: Machine Translation, Customization, Patent Machine Translation

1. Introduction

An English-Korean machine translation system has been developed in earnest in Korea since 1996. We have applyed it to different areas such as web translation (Choi, 1999) and broadcasting subtitle translation (Choi, 2001). Recently, the natural language processing(NLP) of intellectual property documents is attracting many researchers and NLP-related companies, because NLP techniques associated with specificity of patent domain have promise for improving the translation qualtiy.

It is well known that a sentence style and a dominant translation for a word vary with domains. Therefore, if the domain to be translated is fixed to patents, a adaptation of bilingual dictionary to the patent domain and a customization of natural language analyzer to the linguistic specificity of patent style would be one of 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).

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

Though intensive research has been made on patent MT for the domain-specific advantages, there still remain many issues to be tackled. We focus on the several issues that have continusely been problems in existing English-to-Korean MT systems: (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 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.

According to experience of patent attorneys, it is said that they read about 7 English patent documents to examine one Korean patent document in average. It means that they examine about 1,000,000 English patent documents for new 150,000 Korean patent documents every year. Korean patent attorneys have required any machine translation system to solve language barrier because they prefer reading Korean translated patent documents to reading English patent documents in spite of such linguistic competency as English native speaker.

In this point, th development of the English-Koran patent translation system is closely related to offering of English-to-Korean patent machine translation service through Internet. KIPO (Korean Intellectual Property Office) pushes on with on-line translation service of patent documents by using MT system.

The English-to-Korean patent machine translation system described in this paper was developed by ETRI (Electronics and Telecommunications Research Institute) under the auspices of the MIC (Ministry of Information and Communication, Korea) during 2005-2006. In 2006, the patent MT system started an on-line patent 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) tries to launch an English-Korean patent MT service.

Section 2 describes the customization processes that relate to new terminology construction, patent-specific probabilities of POS tagger, long and complex sentence analysis, and target word selection, respectively. The experimental work is presented in section 3. Lastly, in section 4, we present some conclusions.

2. Customization Process

Some methods of customization to change general MT system to domain-specific MT system have been introduced. For example, the customization process in SYSTRAN as multilingual MT system consists of the following steps: term extraction, dictionary customization, linguistic customization, and testing/evaluation (Zajac, 2003). Hong (2005) applied such an existing customization process to a Korean-English MT system.

In comparison with the existing customization methods above mentioned, the customization process described in this paper is the first worth-mentioning large-scale customization effort of an MT system for English and Korean.

The customization process for an English-Korean patent MT system includes the following steps: 1) linguistically studying about characteristics of patent documents, 2) extracting unknown words from large patent documents and constructing large bilingual terminology, 3) extracting and constructing patent translation patterns 4) customizing the translation engine modules of the existing general MT system according to linguistic study about characteristics of

patent documents, and 5) evaluating the accuracy of translation modules and the translation quality

2.1.Construction of Patent Terminology

The first step of customization process for patent MT system is to gather the existing terms, extract the unknown words from patent documents, and build the bilingual terms. The customization process described in this paper is similar to the method of Kaji(2005), Shimohata(2005), and Kim(2005) in respect of using the monolingual dictionary and the monolingual patent corpus, but our method is different in that it contains a step inverting the existing bilingual terminology with opposite direction. Extraction and construction of terminology might be represented as a following customization process:

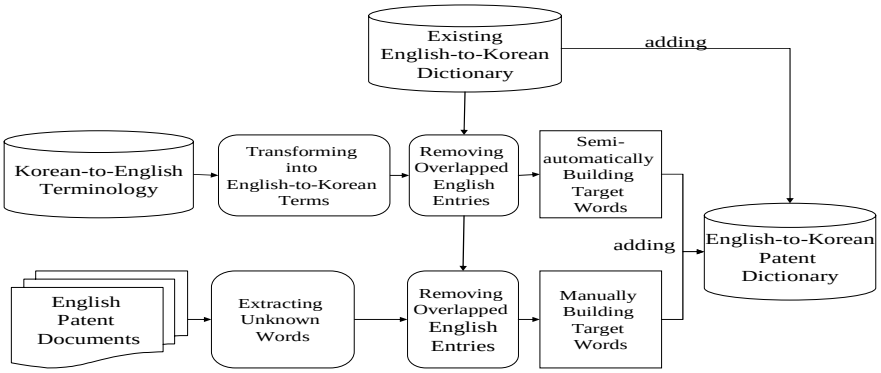

Figure 1: Customization process for building English-Korean patent terminology

As shown in Figure 1, the patent terminology can be built in two ways. One is to extract the unknown words, remove the overlapped entries, and build manually new bilingual terminology. The other is to build semi-automatically new bilingual terminology, assumed we have the existing bilingual terminology with reverse direction (for example, Korean-to-English terminology). By use of the above customization process, we built semi-automatically 801,046 of new bilingual terminoloy and manually 1,039,189 of new bilingual terminology, that is, 1,840,235 English-Korean terms were totally built for 7 months. 23 people as lexicographers have worked to build the new bilingual terms every day.

Table 1: New English-Korean patent terms semi-automatically built by inverting the existing bilingual terminology

Items	Number of entries
Number of existing Korean-English terminolgy	3,052,655
Number of entries of general-purpose English-Korean dictionary	836,000
New English-Korean terms with the exception of the English terms of existing Korean-English terminolgy overlapped with entries of general-purpose English-Korean dictionary	801,046

Table 1 shows that new 801,046 English-Korean terms were semi-automatically constructed from the existing Korean-English terminology. They are consisted of 207,329 single words and 593,717 compounds.

In addition to new patent terms constructed semi-automatically, we had to extract a number of unknown words from English patent documents and manually build new English-Korean terms, because we had no English-Korean patent corpora.

Table 2: New English-Korean patent terms built manually by use of large English patent documents

Items	Number of entries
English patent documents used	1,001,419
Extracted unknown words	9,662,266
New English-Korean terms	1,039,189

In Table 2, we can know that new 1,039,189 English-Korean terms were built from the very large English patent documents. They consisted of 492,295 single terms and 546,894 compound terms.

In the result, we have now 2,676,235 English-Korean terms including existing 836,000 terms of general-purpose English-Korean dictionary.

2.2.Customization of POS Tagger

We define three customization phases for customizing a general POS tagger based on HMM (Hidden Markov Model) to patent domain, according to the characteristics of English patent document mentioned in the section 2.1:

- Customization of surface form analysis
  : a tokenization module and/or a morphological analyzer are modified for tokenizing and/or analyzing the peculiar surface forms found in the specific domain.
- Customization of the lexical information
  : lexical probabilities (output probabilities) are adjusted for holding domain-specific lexical information.
- Customization of the context information
  : contextual probabilities (transition probabilities) are adjusted for holding the domain-specific contextual information.

In the first phase for customization of surface form analysis, the tokenization module is modified to tokenize and/or chunk very complex symbol words, a chemical formula, a mathematical formula, programming codes, and so on. And our morphological analyzer is improved to assign the estimated part-of-speeches into a compound word connected with hyphen or slash. The estimated part-of-speeches are estimated using the part-of-speeches of its components.

The POS tagging module of English-to-Korean patent machine translation system is based on lexicalized HMM (Pla & Molina, 2005). Therefore, the best simple strategy for the second and third customization phase is retrained from a very large tagged patent corpus. However, there is not a tagged patent corpus and it is also very difficult to construct it. Accordingly, for customizing the lexical and contextual probabilities, we used a raw patent corpus consisting of about one million US patent documents applied for from 2001 to 2005. First, the words of the raw corpus are automatically tagged by our general domain POS tagging system, and then the lexical and contextual probabilities are extracted from the machine-tagged patent corpus. Next, we extracted high-frequent lexical having very different probability with that of the general domain. And we extracted the high-frequent contextual n-grams that didn't appear in the general domain. The extracted lexical and contextual n-grams are tuned by the human experts. For customization of our POS tagger, we tuned about 6,000 lexical and about 1,500 tri-grams.

2.3.Customization of Syntactic Analyzer for Long Sentences

The important syntactic characteristics of the patent document are the frequent use of the patent intrinsic translation pattern and abnormally long sentences. With these as central figures, the main contents of customization of syntax analysis are as follows:

- A build-up and application of the patent translation pattern
  : the patent-specific patterns are manually built up and the processing for the recognition of patterns is performed. The general forms of the patent-specific patterns are composed of lexical words and syntactic nodes. Therefore, for the recognition of the patterns, the lexical words are firstly matched, and then the ranges between the lexical words are parsed. If all ranges are parsed into corresponding syntactic nodes in the translation pattern, the pattern is recognized.
- A large amount of lexical pattern collections and application
  : in the patent documents, the high frequency lexical patterns corresponding to the specific part-of-speech patterns are automatically extracted and are applied to syntactic analysis.
- Performing the coordinate construction recognition for long sentences
  : for the coordinate construction recognition, first, the possible site which can become the initial point, the intermediate point, and an endpoint of the parallel construct. Then, the similarity table between each node is constructed. For the all possible coordinate structures, the coordinate weight is calculated using the similarity table. Finally, the coordinate structure having maximum coordinate weight is selected as a final result. The recognized coordinate construction is chunked to one unit, and accordingly the sentence is simplified.
- Performing the sentence segmentation for the long sentence
  : in case of being too long to analyze the sentence at a time in syntactic analyzer, even after the parallel construction is recognized, the sentence segmentation is performed. The sentence is segmented by recognizing participles or simple sentences.
- Reflecting attachment preferences
  : priority for the attachment of ’for’ prepositional phrase and participle is given to the NP attachment than VP attachment.

2.4.Customization of Transfer Module for Target Word Selection

We customized the transfer module for patent document translation. Following customization items for transfer modules were considered:

- The registration of the default target word according to patent technical field
  : In the case that the same source word can be translated into different target word depending on the patent field, the specific value of the 'field' feature is assigned to dictionary.
- The gathering of collocation information for noun/verb with high frequency
  : We use collocation information to select the proper target word depending on the context. The collocation information is used as main knowledge to cope with the problem of the target word selection.
- The implementation of the module to achieve target word selection using collocation information
  : This module carries out the task to select proper target word using collocation information. The approach consists of two levels. In the first step, sense ambiguity of English word is resolved. In the second step, the most suitable Korean target word is selected. To select the most suitable target word, our approach uses multiple knowledge sources such as verb frame patterns, sense vectors based on collocations, statistical Korean local context information and co-occurring POS information. Sense vectors are made using English-Korean parallel corpus. (Lee, 2006)

- The implementation of the interpreter for patent-specific patterns
  : The patent-specific patterns were introduced to translate highly frequent expression. Our parser uses these patent-specific patterns in parsing time and then transfer module interprets the patent-specific patterns applied by the parser.

3. Evaluation

3.1.Evaluation of Morphological Analyzer

We evaluated the performance the POS tagger specialized to the patent domain (PatTagger), compared with the performance of our general-purpose POS tagger (GPTagger). For the evaluation, we used 100 sentences of the electrical and electronics field (EEF) among the whole translation evaluation test set. The EEF test set consists of 2,942 words and the number of words per a sentence is 29.42.

Table 3 shows the word accuracy and sentence accuracy of two taggers. 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 GPTagger. This improvement seems to contribute to the performance improvement of the proposed English-Korean patent translation system.

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

	GPTagger	PatTagger	
Word tagging accuracy	95.85%	99.62%	Up 3.77%
Sentence tagging accuracy	50.00%	91.00%	Up 41.00%

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 2.2 and construction of terminology mentioned in the section 2.1. The construction of terminology 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 4: The performance improvement of PatTagger and the improvement of its word tagging accuracy.

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%

3.2.Evaluation of Syntactic Analyzer for Long Sentences

The evaluation result by the customization of syntactic analyzer is as follows:

Table 5: Evaluation of customization of syntactic analyzer

	Syntactic analysis accuracy	Translation accuracy	Number of translation patterns
General-purpose Syntactic Analyzer	69%	73%	47,413
Customized Syntactic Analyzer	85%	81.6%	75,931
ERR	Up 16%	Up 8.6%	6 months, 3 people per day

In the above table, the syntactic analysis accuracy is calculated by the ratio of the number of correctly analyzed sentences to the number of total sentences[1]. We use the accuracy by the sentence unit instead of the common parsing evaluation metrics by the bracketing match, because the accuracy by the sentence unit shows the direct correlation with the translation accuracy. And the translation accuracy is the comparison result between before and after the customization of syntactic analyzer in the translation system customized for patent documents.

3.3.Evaluation of Transfer Module for Target Word Selection

We compared general-purpose transfer module and patent-specific transfer module for evaluating the performance of target word selection for noun. The test set for the experiment consists of 100 sentences from patent documents. Table 6 shows the experimental results of target word selection of the customized MT system and the non-customized MT system. The performance of customized MT system which has taken customization process for patent document into account overcomes the one of counterpart.

Table 6: Result of target word selection for noun

	Accuracy of target word selection for noun	Percentage of unknown word
General-purpose Transfer Module	71.7%	16.3%
Customized Transfer Module	92.4%	1.5%

3.4.Translation Accuracy

In this chapter, we describe the evaluation about translation quality of English-to-Korean patent MT system. It relates to 5 major patent fields selected from different patent fields. We used the following test sentences, evaluation method and evaluation criterion for translation quality:

- Test sentences
  : translation accuracy was assessed with 100 test sentences randomly extracted from each one of 5 major patent fields (machinery, electronics, chemistry, medicine and computer). The test set was so open that it might reflect a real patent document. Among 100 sentences for each patent field, 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 28.09 words.
- Evaluation criterion:

[1] We consider a sentence as correct when the syntactic analysis result of the sentence has a trivial analysis error that dosen’t affect the translation result.

Table 7: Scoring criteria for translation accuracy

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

- Evaluation method:
  - 7 professional translators were hired for the evaluation. Ruling out the highest and the lowest score, the scores for each sentence were summed. The method for translation accuracy was 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.

The evaluation results for each patent field were as follows:

Table 8: Translation accuracy for each patent field
(Evaluation date: Dec.13, 2006)

Patent field	Average length of a sentence	Translation accuracy higher than 1 score	Translation accuracy higher than 3 scores
Machinery	30.34 words	83.50%	85.00%
Electronics	28.19 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	**28.09 words**	**82.43%**	**87.60%**

Table 8 shows that the translation accuracy of English-Korean patent MT system was 82.43% on the average. The number of the sentence that were rated equal to or higher than 3 points was 438. It means that about 87.60% of all translations were understandable.

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.

Table 9 is the result to compare the translation accuracy before customization with that after customization in the electronic patent document.

Table 9: Comparison of translation accuracy before customization with that after customization in electronic patent document
(Evaluation date: Dec. 13, 2006)

Patent field	Average length of sentence	Translation accuracy before customization	Translation accuracy after customization

Electronics	28.19 words	54.25%	82.20%

In Table 9, 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 qualtiy of English-Korean MT system on patent documents.

4. Conclusion

In this paper we described a method for customizing English-to-Korean machine translation system from general domain into patent domain. The customizing method consists of following steps: 1) linguistically studying about characteristics of patent documents, 2) extracting unknown words from large patent documents and constructing large bilingual terminology, 3) extracting and constructing the patent-specific translation patterns 4) customizing the translation engine modules of the existing general MT system according to linguistic study about characteristics of patent documents, and 5) evaluating the accuracy of translation modules and the translation quality.

The English-Korean patent MT system described in this paper was installed in IPAC (International Patent Assistance Center) under MOCIE (Ministry of Commerce, Industry and Energy) in Korea and provides the patent attorneys and patent examiners with the on-line English-Korean machine translation service for patent documents (http://www.ipac.or.kr). In 2007, KIPO (Korean Intellectual Property Office) is expected to launch its English-Korean MT service.

In near future, we make a plan to evaluate automatically the translation quality like BLEU by building several references and to develop the tool for automatic tuning of bilingual terminology by use of the patent corpus.

References

Brants T. 2000 “TnT – a statistical part-of-speech tagger”. *Proceedings of the Sixth Applied Natural Language Processing*, pp. 224-231.

Choi S.K., Kim T.W., Yuh S.H, Jung H.M., Sim C.M. and Park S.K. 1999. English-to-Korean Web Translator: “FromTo/Web-EK”, *Machine Translation Summit VII.*

Choi S.K., Yang S.I., Roh Y.H., Lee K.Y. and Park S.K. 2001. English-to-Korean Automatic Caption Translation, *International Conference on the Computer Processing of Oriental Languages*.

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

Kaji H. 2005. Domain Dependence of Lexical Translation: A Case Study of Patent Abstract. *Machine Translation Summit X, Workshop on Patent Translation.*

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

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

Pla F. and Molina A. 2005. Improving Part-of-speech Tagging Using Lexicalized HMMs. *Natural Language Engineering*, 10(2), 167-189.

Shimohata S. 2005. Finding Translation Candidates from Patent Corpus. *Machine Translation Summit X, Workshop on Patent Translation.*

Shinmori A., Okumura M., Marukawa Y. and Iwayama M. 2003. Patent Claim Processing for

Readability - Structure Analysis and Term Explanation, *the Association for Computational Linguistics, Workshop on Patent Corpus Processing*.
Zajac R. 2003. MT Customziation. *Machine Translation Summit Workshop*.