Customizing an English-Korean Machine Translation System for Patent/Technical Documents Translation *

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

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

Abstract. This paper addresses a method for customizing an English-Korean machine translation system from general domain to patent or technical document domain. The customizing method includes the followings: (1) adapting the probabilities of POS tagger trained from general domain to the specific domain, (2) syntactically analyzing long and complex sentences by recognizing coordinate structures, and (3) selecting a proper target word using domain-specific bilingual dictionary and collocation knowledge extracted from patent or technical document corpus. The translation accuracy of the customized English-Korean patent translation system is 82.43% on the average in 5 patent categories according to the evaluation of 7 professional patent translators. The translation accuracy of the customized English-Korean technical document translation system is 81.10% and its BLEU score is 0.5185 in the evaluation test set where the average BLEU score of cross-evaluation between references is 0.6615.

Keywords: Machine Translation, Customization for MT, Patent Translation, Technical Document Translation.

1 Introduction

We often look for the foreign patents or technical documents for acquiring the current trends and new information. When we try to translate the foreign patents or documents in order to just acquire the information, we want to require the rapidity of the translation and the understandable translation quality, rather than the completeness of the translation quality. Such users' demand has become a hot research issue in the MT community.

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 or technical documents, bilingual dictionary adaptation to the domain and customizing natural language analyzers to the linguistic specificity of the domain's 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).

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

This paper addresses the customization of an E-K(English-Korean) MT system for patent and technical documents translation. The E-K patent MT system "FromTo-EK/PAT" and The E-K technical paper MT system "FromTo-EK/PAP" described in this paper is based on an E-K MT system developed for the web translation in a general domain. We first customized our general E-K MT system for patent translation, and then customized E-K patent MT system to technical document domain.

* The work reported in this paper was supported by the IT R&D program of MKE, "Development of Machine Translation Technology for Korean/Chinese/English Spoken Language and Business Documents".

Our E-K MT system belongs to basically the pattern-based methodology for machine translation. It has the formalism that does English sentence analysis in which English domain-specific patterns are used, matches the English domain-specific pattern with its Korean domain-specific pattern, and then generates a Korean sentence from it. E-K 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.

2 Issues for Customizing MT System to Specific Domain

It is important to customize translation knowledge and translation modules for adapting the existing general MT system to translation of patent documents and technical documents. The customization for the translation knowledge is able to be divided into two steps: (1) tuning general translation knowledge to patent-specific or technical document specific translation knowledge, and (2) efficiently constructing the unknown words and new domain-specific translation patterns found in patent documents and technical documents. The customization of existing translation knowledge is closely related with the customization of modules using the translation knowledge.

What is firstly necessary for customizing a general MT system to a specific domain is to extract the large-scale terms found newly in patent documents or technical documents, and construct their translation knowledge such as the target words. 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 E-K translation to constructing all unknown words appearing in domain-specific 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 or technical documents. This means that there is room for doubt that the general POS taggers keep their performance in the specific 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 an 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 domain by using raw domain-specific corpus.

Compared with general documents, one characteristic of patent documents is to use the abnormally long and complex sentences, 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 E-K machine translation is very important factor in that it has a direct influence on the machine translation quality. Particularly, in the documents of non-specific domain such as web pages, the target word selection problems of English ambiguous words occur very frequently, and many frequently used English words can be translated to various Korean words depending on the contexts. However, in E-K patent machine translation, most of words used in patents or technical 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. In the case of technical document translation, the ambiguities of target word selection are higher than in patent translation, but the

ambiguities are much lower than in the general domain. Although patent or technical 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. For target word selection, we tried to disambiguate the possible senses of English words by use of other knowledge like sense vectors and Korean bi-gram context information. If the target word selection module didn't make the decision, the target word is selected with dominant Korean target word. To improve the translation accuracy, we reconstructed the E-K bilingual dictionary whose English words contain their dominant Korean target word according to specific domain.

3 Customizing Methods

3.1 A Domain Adaptation Method for POS Tagger

Three items were tuned for customizing a broad coverage POS tagger based on HMM to specific domain (patents or technical documents 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. The surface forms of the words appearing in the patents are very more various than the words of the technical documents.

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:

$$\bar{T} = \underset{t_1 \cdots t_n}{\operatorname{argmax}} \left(\prod_{1 \ldots n} P(t_i \mid t_{i-1}, t_{i-2}) \cdot P(w_i \mid t_i) \right) \qquad (1)$$

for given sequence of words w_1, …, w_n of length n. t_1, …,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 and simplest strategy for the second and third customization phases is to re-estimate lexical and contextual probabilities from very large tagged patents or technical documents corpus. However, there is not a tagged patent or technical document 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 for patent domain and a raw technical document corpus consisting of about 20 million abstracts of English technical articles. 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 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 in each domain customization. For customizing our general POS tagger to patent or technical document domain, we tuned about 6,000 lexemes and about 1,500 tri-grams in each case.

The representative tri-grams among the extracted n-gram in the patent domain 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.

3.2 Syntactic Analysis for Domain-Specific Document

Two most important ones among peculiar syntactic characteristics of patent or technical documents are the frequent use of patent or technical document specific patterns and the abnormally long sentences (Shinmori et al., 2003). In patent documents, abnormally long sentences are frequently appeared, but are less appeared in technical documents compared to patents. Considering these characteristics as central features, we will describe the main contents of syntax analysis for patent or technical documents in detail.

Application of domain-specific patterns

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

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. If a sentence formed by coordination of syntactic nodes such as NP, VP, that-clause, etc., is segmented between coordinate constituent nodes, the segmentation can cause syntactic analysis errors because a segment can be dependent on some other node in the 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. Kurohashi and Nagao (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 domain-specific machine translation in two ways considering knowledge and engine. For adapting E-K 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.

In case of adapting E-K bilingual terms to technical document domain, we didn't define the categories. We extracted English ambiguous words with high frequency in the technical document corpus, and then we sorted their Korean equivalents with Korean word frequency extracted from Korean technical document corpus. Next, human translator selected dominant Korean word from the sorted Korean word list.

For the ambiguous English words which couldn't be 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 E-K 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:

$$SV = (w(c_1), w(c_2), w(c_3), ..., w(c_n)) \quad (2)$$

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:

$$w(c_k) = \Pr(s = s_i | w = c_k) \quad (3)$$

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:

$$sim(v,w) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}} \qquad (4)$$

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 Evaluation Methods

To evaluate our E-K MT system, we used a human MT evaluation and the BLEU method (Papineni et al., 2002), which is based on comparison of n-gram models in MT output and in a set of human reference translation. In our human MT evaluation, human translators yield the score shown in Table 1 to evaluate the machine translation results. In our evaluation, 7 professional translators evaluated the results. Ruling out the highest score and the lowest score, the rest 5 scores were used for translation accuracy evaluation. The translation accuracy was defined as follows:

$$\text{: translation accuracy(\%)} = \sum_{i=1}^{n}(\sum_{j=1}^{5}(score_j/4))/5)/n \times 100.0 \text{ ,}$$

where n is the number of test sentences and $score_j$ is the score evaluated by the j-th professional translator.

Table 1: 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

4.2 Evaluation for E-K Patent MT System

In this section, we describe the evaluation about translation quality of E-K patent MT system. In case of the patent translation evaluation, we only used the human MT evaluation method, because we didn't construct an evaluation set for the BLEU method. We used the following test sentences for the human MT evaluation:

- 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.

Table 2 shows that the translation accuracy of E-K 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.

Table 2: Translation accuracy for each patent field

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 3 is the result to compare the translation accuracy before customization with that after customization in the electronic patent document. In Table 3, 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 E-K MT system on patent documents.

Table 3: Comparison of translation accuracy before customization with that after customization in electronic patent document

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

4.3 Evaluation for English-Korean technical document MT System

E-K technical document MT system, we used two test sets for the human MT evaluation and the automatic MT evaluation, respectively. The test set of human MT evaluation consists of 400 sentences extracted randomly from about 100,000 English articles and the average length of a sentence is 18.33. The test set of the BLEU method consists of 1,000 sentences with 8 reference translations and the average length of a sentence is 18.37. Several kinds of n-grams can be used in the BLEU, we used 4-gram in this paper. In the evaluation set, the average BLEU score of cross-evaluation between 8 references(a leave-one-out cross-evaluation) is 0.6615.

Table 4: Translation evaluation results in the technical document domain

Test Date	Translation accuracy	Translation accuracy higher than 3 scores	BLEU Score
FromTo-EK/PAT	74.39%	65.00%	0.4793
Customizing tagger	77.25%	67.50%	0.4946
Customizing parser	78.40%	70.75%	0.5152
Adding unknown words	80.78%	75.00%	0.5169
Customizing target word selection	81.10%	75.75%	0.5185

Table 4 shows the increase of the translation accuracy and BLEU scores as customizing the patent MT system to the technical document domain. First, we evaluated the patent MT system(FromTo-EK/PAT) in two test sets as the baseline test. Although the translation accuracy of the system was 82.43% in the patent domain, the system provides only 74.39% in the technical document domain. According to customizing the modules and adding unknown words into bilingual dictionary, the system improved the performance in the new domain. The best contribution enhanced the performance was customized by adapting technical-document-specific POS tagger. Then, the next contribution was the construction of bilingual dictionary with adding unknown words extracted from technical document corpus. Because long sentences are less appeared in the technical documents compared to patents, and the parser was customized after customizing the tagger, the improvement of customizing the parser is falling short of expectation. From table 4, we can speculate that domain-specific target word selection didn't provide a significant contribution to translation accuracy in the technical document domain.

5 Conclusion

In this paper we described a method for customizing E-K machine translation system from general domain into patent and technical document domain. First, to adapt general-purpose POS tagger to the patent or technical document domain, we proposed the method for semi-automatically adjusting probabilities trained from general domain to domain-specific context using raw English patent or technical documents. Secondly, 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 domain-specific bilingual dictionary and collocation knowledge extracted from raw patent or technical document corpus.

References

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

Kurohashi, S. and M. Nagao. 1994. A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structure. *Computational Linguistics, 20(4)*: 507-534.

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

Papineni, K., S. Roukos, T. Ward and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics* (ACL), Philadelphia, July 2002, pp.311-318.

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

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

Kim, S.-D., B.-T. Zhang and Y.T. Kim. 2001. Learning-based Intrasentence Segmentation for Efficient Translation of Long Sentences. *Machine Translation, 16(3)*:151-174.