How to Overcome the Domain Barriers in Pattern-Based Machine Translation System*

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

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

Abstract. One of difficult issues in pattern-based machine translation system is maybe to find how to overcome the domain difference in adapting a system from one domain to other domain. This paper describes how we have resolved such barriers among domains as default target word of any domain, domain-specific patterns, and domain adaptation of engine modules in pattern-based machine translation system, especially English-Korean pattern-based machine translation system. For this, we will discuss two types of customization methods which mean a method adapting an existing system to new domain. One is the pure customization method introduced for patent machine translation system in 2006 and another is the upgraded customization method applied to scientific paper machine translation system in 2007. By introducing an upgraded customization method, we could implement a practical machine translation system for scientific paper translation within 8 months, in comparison with the patent machine translation system that was completed even in 24 months by the pure customization method. The translation accuracy of scientific paper machine translation system also rose 77.25% to 81.10% in spite of short term of 8 months.

Keywords: Customization, Machine Translation, Pattern-based MT, Scientific Paper MT.

1. Introduction

The use of on-line systems is the biggest growth area in the use of machine translation. People are translating web pages or very large documents by using machine translation system as the solution, as human translation of pages which need to be continually updated or are very large scale is not feasible (Mellebeek et.al., 2005).

Electronics and Telecommunications Research Institute (ETRI, henceforth) in Korea has developed the web-based English-Korean machine translation system till 2004, under assumption that as the size of patterns grows, the performance of the system can be incrementally improved (Hong et. al., 2003). During 2 years (2005- 2006) it implemented an English-Korean patent machine translation system on the basis of the web-based English-Korean machine translation system. The English-Korean patent machine translation system was installed in International Patent Assistance Center (IPAC, henceforth) under Ministry of

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

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) because it helped them understand the existing English patent documents easier and more rapidly (Choi et. al., 2007).

It was due to the customization method (Kwon et. al., 2007) that we could change successfully an existing machine translation system from general domain to patent domain. Figure 1 shows us an example of patent machine translation service at IPAC.

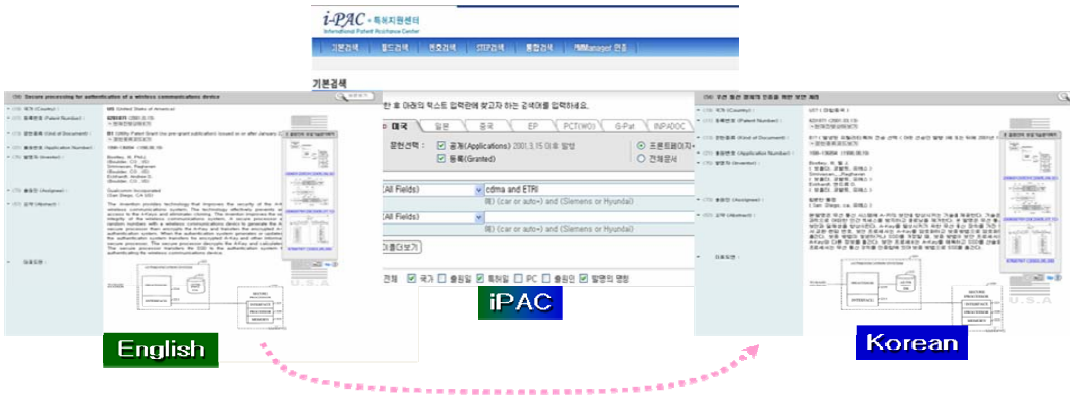

Figure 1: An Example of Patent Machine Translation Service at IPAC

ETRI had upgraded the customization method applied to the English-Korean patent machine translation system since 2007 and completed the practical level of English-Korean scientific paper machine translation system within 8 months. The English-Korean machine translation service for scientific paper translation is expected to be launched for students since September, 2008.

This paper describes how we have resolved such barriers among domains as default target word of any domain, domain-specific patterns, and domain adaptation of engine modules in pattern-based machine translation system, especially English-Korean patterns-based machine translation system. Especially, we will describe a difference between the pure customization method for patent machine translation system and the upgraded customization method applied to scientific paper machine translation system.

The construction of this paper is as follows: in section 2 the pure customization method will be sketched and its experiment will be showed. The limits of the pure customization method will be uncovered in the section 3. To deal with the problems found in the experiment, an upgraded customization method will be proposed in the section 4. In section 5 another experiment will be conducted to evaluate the proposed method. The discussion of the previous sections will be summarized in the concluding section 6.

2. Pure Customization Method

2.1.Customization Steps

As mentioned before, a domain customization method means a method adapting an existing system to new domain. The pure customization method was designed in the course of changing the web-based English-Korean machine transaltion system with general domain into the domain-specific English-Korean machine translation system such as patent domain. It was basically dependent on an idea of Zajac R. (2003) consisting of following steps:

Step 1. Collecting a large scale of domain-specific documents
Step 2. Linguistically studying about characteristics of the collected documents
Step 3. Automatically extracting unknown words and semi-automatically constructing their equivalent words

Step 4. Manually constructing domain-specific translation patterns
Step 5. Customizing the translation engine modules of the existing MT system
Step 6. Human evaluation of translation quality

2.2.Experiment 1

To assess the feasibility of pure customization method, we conducted an experiment. The goal of the experiment was to see how much improvement of translation accuracy can be achieved before and after applying the pure customization method to the web-based machine translation system. Following table shows changes of accuracy before and after applying pure customization steps.

Table 1: Before and after applying pure customization method.

Steps	Item	Before	After	Reference
Step 3	Number of terms	836,000	2,052,604	up 1,216,604 for 13 months
Step 4	Number of patterns	39,127	50,214	up 11,087 for 18 months
Step 5	Accuracy of tagging	95.85%	99.62%	
	Accuracy of parsing	69.00%	85.00%	
	Accuracy of target word selection (noun)	71.70%	92.40%	
Step 6	Translation quality	54.25%	82.20%	

3. Problems of Pure Customization Method

We tried to analyze the automatic translation results of scientific paper translated by patent machine translation system to find the problems of the pure customization method. We evaluated 200 blind test sentences of scientific paper with the patent machine translation system. The average length of the sentences was 21.69 words. The translations were evaluated by 3 translators with the scoring scale from 0 (no translation) to 4 (perfect translation) point. The result was as follows:

Table 2: Error Analysis of Translation Result of Scientific Paper by Patent MT System

Item		Number of Errors	%
Translation Engine	Tagging	10	6.10%
	Parsing	28	17.07%
	Target Word Selection	3	1.83%
	Generation	15	9.15%
Translation Knowledge	Dictionary entries	77	46.95%
	Patterns	23	14.02%
etc		8	4.88%
Total		164	100.00%
Translation Accuracy			78.63%

Table 2 shows that errors of dictionary entries amount to 46.95% and errors of patterns cover 14.02%. That is, 60.97% of total translation errors are caused by translation knowledge.

This means that the extraction of unknown dictionary entries and unknown patterns as well as the adaptation of existing dictionary entries and patterns to new domain are very important to

tune the existing machine translation system to new domain. It would be first problem of the pure customization method not to have the step such as the tuning of existing dictionary entries (Ayan et. al., 2003) and the corpus-assisted expansion of existing patterns (Yamada et. al., 2002).

The second problem of pure customization method is that it has no step of automatic tuning for existing translation engine modules. We corrected the existing system modules such as tagger, parser, transfer and generator whenever we found their errors. Its problem was to spend a long tuning time. Therefore, we needed a new step for semi-automatic tuning of translation engine modules to cut tuning time.

Finally, the weakness of pure customization method was related to only human evaluation for translation assessment. To reduce the expenses and much time, we added an automatic evaluation like BLEU (Papineni et.al., 2002) to the human evaluation.

4. Upgraded Customization Method

To resolve the problems mentioned in above section, we introduce new steps and propose the new process as upgraded customization method as follows:

Step 1. Collecting a large scale of domain-specific documents
Step 2. Linguistically studying about characteristics of the collected documents
Step 3. Automatically extracting unknown words and semi-automatically constructing their equivalent words
Step 4. Semi-automatic tuning of existing terminology
Step 5. Semi-automatic constructing domain-specific translation patterns
Step 6. Semi-automatic customization of the translation engine modules based on answer set
Step 7. Human evaluation and automatic evaluation of translation accuracy

4.1.Overall Customization Flow

The Figure 2 illustrates the overall customization flow.

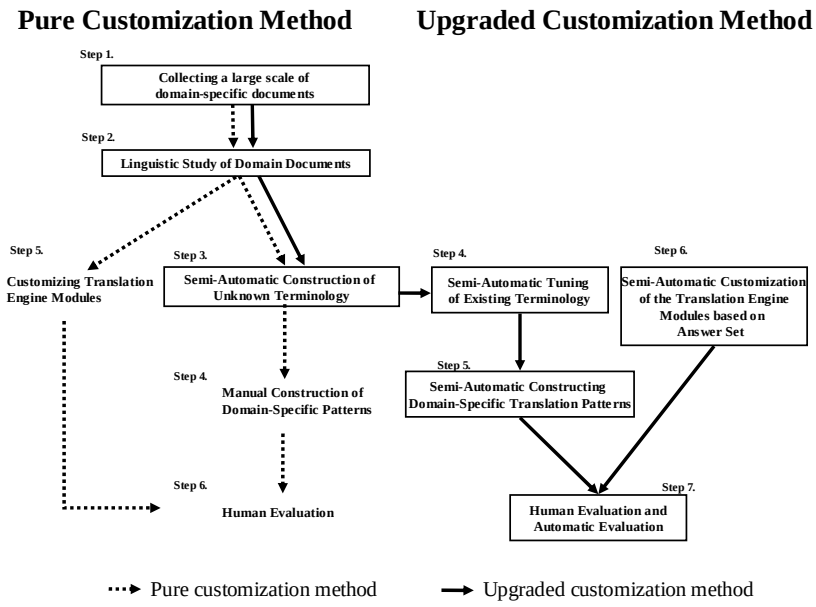

Figure 2: Pure Customization Method vs. Upgraded Customization Method

4.2.Semi-Automatic Tuning of Existing Terminology

We performed the domain tuning for target words of noun, verb and adjective terms adapted to patent domain by using English-Korean comparable corpus. In case of adapting English-Korean

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 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:

$$SV = (w(c_1), w(c_2), w(c_3), ..., w(c_n)) \quad (1)$$

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:

$$w(c_k) = \Pr(s = s_i | w = c_k) \quad (2)$$

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:

$$sim(v, w) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}} \quad (3)$$

4.3.Semi-Automatic Construction of Domain-Specific Patterns

Domain-specific patterns are one of factors to make a translation quality higher. The construction of the unknown domain-specific patterns for scientific papers was performed by automatically extracting the domain-specific patterns from tagged corpus of large scientific papers and manually building theirs target patterns.

Figure 3 outlines the method of automatically extracting the domain-specific patterns from scientific papers.

1. make large raw corpus of scientific papers their tagged corpus.
2. extract from the tagged corpus the following pattern candidates:
   - fixed part-of-speech patterns (e.g. COMMA_CONJUNCTION_PRONOUN_VERB_COMMA such as ",as you know,") or
   - meaningful patterns between boundary conditions (e.g. starting with preposition, verb, and conjunction, and ending with preposition, verb, conjunction, noun, verb, auxiliary verb, and number such as "in reference with", "compared with").

3. filter pattern candidates such as ‘preposition preposition’, ‘preposition noun’, and ‘noun of’ (e.g. “in on”, “for it”, “term of”).
4. count the number of each lexical pattern candidate.
5. conduct a base NP chunking for the lexical pattern candidates (ex. “accuse him of” -> “accuse NP of”)
6. subtract a frequency of long pattern from a frequency of its short pattern to delete unnecessary short patterns (e.g. 1,050 “in spite” – 1,000 “in spite of”).
7. order patterns according to frequency
(e.g. 115334 <I3> in_order_to, 67935 <P4> it_be_shown_that, 61882 <I3> with_respect_to, 60860 <V2> apply[VN]_to, 59730 <V3> paly_NP_in, 53573 <J2> consistent_with)

Figure 3: Method of automatically extracting the domain-specific patterns from scientific papers

4.4.Semi-Automatically Customizing Translation Engine Modules Based on Answer Set

Answer set is a morphologically and syntactically tagged corpus of 5,000 sentences that two human lexicographers constructed. From the answer set we collected the correct answers of morphologically or syntactically ambiguities. On the basis of them, the English part-of-speech tagger and parser were able to be checked automatically.

For customization of the morphological tagger we have first collected the tagging errors with morphological ambiguities of scientific papers that occur frequently. Then the tagging errors were corrected manually if they were matched with the corresponding parts of the answer set. For example, a word with ‘-ing’ can be a noun (NN) or a gerund (VBG). We could find that the par-of-speech of the word with a form ‘-ing’ became noun (NN) or adjective (JJ) before lexical words ‘system’ and ‘method’.

The semi-automatic customization of parser could be achieved by semi-automatically controlling the probabilistic weight of parsing rules including the different attachment ambiguities, such as infinitive phrase attachment and prepositional phrase attachment.

5. Experiment 2

In this section our concern was to see how much the translation accuracy can be enhanced by introducing an upgraded customization method in the place of a pure customization method. To find out this, we evaluated 400 blind test sentences for human evaluation and 1,000 sentences with 5 references for automatic evaluation. The translations for human evaluation were scored at the same manner described in the section 3. In the first experiment described in the section 3, the 616 patterns increased every month, while about 4,000 patterns are growing every month due to a step ‘semi-automatic construction of domain-specific patterns’ in upgraded customization method. After introducing an upgraded customization method, the translation accuracy by human evaluation improved from 77.25% to 81.10% in spite of short term of 8 months. As well, the morphological BLEU score also rose from 0.4946 to 0.5185.

Table 3: Translation Accuracy of Pure and Upgraded Customization Method

Steps	Item	Pure Customization Method	Upgraded Customization Method	Reference
Step 3	Number of terms	2,052,604	2,510,496	up 457,892 for 5 months

Step 5	Number of patterns	50,214	74,337	up 24,123 for 6 months
Step 6	Accuracy of tagging	99.20%	99.27%	
	Accuracy of parsing	72.00%	82.00%	Up 10.00%
	Accuracy of target word selection (noun)	79.00%	87.75%	
Step 7	Human evaluation	77.25%	81.10%	
	Automatic evaluation (Morphological BLEU)	0.4946	0.5185	

6. Conclusion

In this paper we elaborated on the limits and potentials of pure customization method and introduced an upgraded customization method including some of steps of pure customization method. The pure customization method suffers from manually increasing domain-specific dictionary entries and patterns, while the upgraded customization method is oriented at semi-automatic construction of them. By introducing an upgraded customization method, we could implement a practical machine translation system for scientific paper translation within 8 months. The translation accuracy amounts to 81.10%.

We launched a pilot service named “iMT (interactive Machine Translation)” from September 2007 using English-Korean technical document MT system. The pilot service provides Korean-English technical document translation service and English-Korean technical document translation service to users. The English-Korean technical document MT service automatically translates the English PDF file to Korean text as shown in Figure 4. In Figure 4, the service extracts only text fields (right top window of the Figure 4) from the user’s PDF file (left window of Figure 4), next translate into Korean text (right bottom window of Figure 4). In the pilot service, about 50 users among total 2,055 users translate nearly 300 English articles a day. Through the pilot service, we transferred the technology to a machine translation related company at January 2008, and the service will be commercialized at the end of 2008.

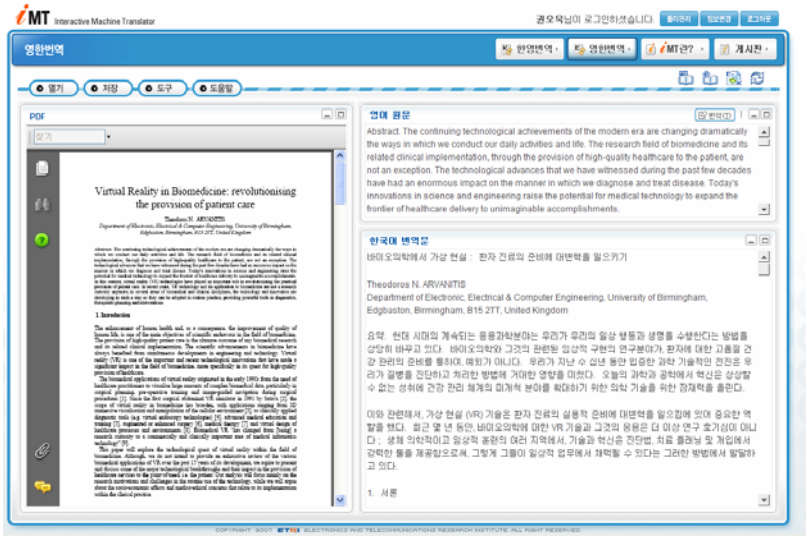

Figure 4: An Example of Machine Translation Service for English-Korean Scientific Paper Translation

In the near future, we are planned to add a new item for customization like a natural generation based on statistical post-editing (Dugast et.al., 2007) to the upgraded customization method.

References

Ayan, N.F., B.J. Dorr and O. Kolak. 2003. Domain Tuning of Bilingual Lexicons for MT. *CS-TR-4449, UMIACS-TR-2003-19, LAMP-TR-096.*

Choi, S.K., O.W. Kwon, K.Y. Lee, Y.H. Roh and Y.G. Kim. 2007. Customizing an English-Korean Machine Translation System for Patent Translation. *The 21st Pacific Asia Conference on Language, Information and Computation (PACLIC 21*), pp. 105-114.

Dugast L., J. Senellart and P. Koehn. 2007. Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. *In Proceedings of the Second Workshop on Statistical Machine Translation.* pp.220-223.

Hong, M.P., K.Y. Lee., Y.H. Roh, S.K. Choi and S.K. Park. 2003. Sentence Pattern-based MT revisited. *In Proceedings of 20st International Conference Computer Processing of Oriental Languages (ICCPOL '03)*. pp. 1-7.

Kwon, O.W., S.K. Choi, K.Y. Lee, Y.H. Roh and Y.G. Kim. 2007. English-Korean Patent Translation System: FromTo-EK/PAT. *MT Summit XI Workshop on Patent Translation*, pp.1-8.

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.

Mellebeek B., A. Khasin, J.V. Genabith and A. Way. 2005. TransBooster: Boosting the Performance of Wide-Coverage Machine Translation Systems. *EMAT 2005 Conference Proceedings.* pp.189-197.

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 Computational Linguistics (ACL)*, Philadelphia, USA, pp.311-318.

Yamada S., K. Imamura and K. Yamamoto. 2002. Corpus-Assisted Expansion of Manual MT Knowledge. *In Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2002)*. pp.199-208.

Zajac R. 2003. MT Customziation. *Machine Translation Summit Workshop*.