Patent Translation as Technical Document Translation: Customizing a Chinese-Korean MT System to Patent Domain

Yun Jin, Oh-Woog Kwon, Seung-Hoon Na and Young-Gil Kim
NLP Research Team, Electronics and Telecommunications Research Institute
218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, Korea
{wkim1019, ohwoog, nash, kimyk}@etri.re.kr

Abstract

The purpose of patent translation is to correctly translate patent documents from one language to another language semantically and syntactically. In this paper, we view *patent translation* as *technical document translation* given their domain similarity in terms of their terminologies and writing styles. From this viewpoint, we simply perform patent translation using a *technical* domain MT system without any further domain adaptation. Experimental results in a Chinese-to-Korean MT system shows that the improved translation performance in technical domain leads to a further improvement in patent translation.

1 Introduction

It is time consuming and laborious for human translators to translate a particular patent document from source language to target language, because it requires the human translators not only need to know both languages in professional levels but target patent related technologies.

Since intellectual property becomes important on these days and a vast and growing number of foreign language patents can be easily accessed via internet, many people want to swiftly review and refer to those related foreign language patents in their native language. For this reason, the patent translation is more spotlighted than any before. To meet the large degree of the patent translation need, it is arguably necessary to design an automatic patent translation system (i.e., patent MT system), which automates the human translation process and provides an automatically translated patent document to target language.

However, since patent texts have many long sentences and bilingual patent corpus hard to obtain, Statistical Machine Translation (SMT) approach(Brwon et al., 1991) seems not suitable for patent translation; instead, several previous researches have been focused on customizing an existing rule/pattern-based MT system to patent domain (Ehara 2007; Choi et al., 2008; Kwon et al., 2009).

In this paper, we address the issues for customizing a general-purpose Chinese-Korean MT system to patent domain. Our key idea is that we view *patent translation* as *technical document translation*. This is based on the assumption that patent and technical documents are very similar in terms of their terminologies and writing styles. Taking into account this viewpoint, we first customize a general-purpose MT system to technical domain to improve our technical *domain* MT system by automatically enhancing translation knowledge. Then, we simply apply the improved technical domain MT system to translate patent documents without further adaptation. Experimental results in a Chinese-to-Korean MT system shows that the improved translation performance in the technical domain MT system leads to a further improvement for translating patent documents.

2 Our Chinese-Korean MT System

In this section, we briefly describe our existing general-purpose MT system, as it is used as backbone system for customizing to technical domain. Our Chinese-Korean MT system is a typical rule-based MT system. Our MT system consists of Chinese words segmentation, POS tagging, Chinese clause segmentation, Chinese syntactic analysis, Chinese-Korean transfer, and Korean generation. As the most distinctive feature of our MT system, we use *clause-based*

Proceedings of the 5th Workshop on Patent Translation, Nice, September 2, 2013. Yokoyama, S., ed.

translation; we first segment a Chinese sentence to clauses, translate Chinese clauses to Korean clauses and then combine the translated clauses to finally generate a Korean sentence. Because the clauses of a written Chinese sentence are easily identified by the syntactic symbols (space, comma, colon, semi-colon, etc.) and some clue words, the clause-based translation reduces the complexity of the syntax analysis and syntax transfer and improves the efficiency of the speed and quality of translation. Also, the clause-based translation is effective in improving the translation quality of long sentences that are frequently appeared in patents and technical documents.

Our Chinese word segmentation is composed of three processing components: (1) as the main algorithm, we adopted Longest Length Matching (LLM), the effectiveness of which was already verified in previous researches (Chen et al., 1992; Ma et al., 2003). (2) To resolve the segmentation ambiguity problem, we deployed probability based disambiguation approach. For example, the string "高一点(little high)" has two possible segment cases "高(high)|一点(little)" and "高一(high school)|点(dot)". Based on our system, the first case was selected because the probabilistic score of first case (12.8055)[1] is greater than that of second case (8.40111). (3) To handle unknown words, we used two different approaches: a) for general words, we used CRF-based unknown words detection to extract word candidates and insert them with their lexical information to our Chinese word dictionary. b) for proper noun words, we used context-based heuristic detection and chunking approach. For this, we used the list of 20,236 possible Chinese proper name characters.

Our Chinese POS tagging is based on the lexicalized trigram HMM approach. As proposed in (Brants 2000), he applied this approach to English POS tagging. In Chinese, the most ambiguous POS words are Chinese functional words such as "在(in/at/exist)" and "有(have/be/exist)", they can be either general or functional. In our lexical dictionary, the average number of POS tags for functional words are 4.3. So, we use those functional words as lexical features and use their collocation POS to construct trigram lexical POS features, like "在/PO_各/DT_NN ", and then combine those features apply to HMM model.

Our Chinese clause segmentation module decomposes a sentence into a number of clauses, only by using the clause segmentation rules. The rules consist of symbols like space, comma, colon, semi-colon and the clause segmentation clues. The clause segmentation clues are either single Chinese words like verb or phrases, which usually appear before or after the segmentation symbols such as comma, colon, etc.

Our Chinese syntactic analysis is based on chart parsing method which uses fully syntactic grammatical rules and knowledge. The rules are heuristically scored by the grammatical knowledge. The grammatical knowledge consists of 5 fields as a dictionary form; compound word, syntactic pattern information, syntactic feature information, semantic information and collocation information. The most appropriate syntax tree of a given input clause is selected by the sum of scores of the rules which are used to generate the tree.

Our transfer module transfers Chinese clause parse tree to Korean clause parse tree using tree-to-tree transfer rules and bilingual dictionary. The transfer module traverses the Chinese input tree in the head-first manner, searches the transfer rules matching the traversing node and its constituent, and then generates Korean tree using matching transfer rule. The transfer rules consist of a Chinese tree pattern and the corresponding Korean tree pattern. The patterns represent the dependency-based syntax tree with a head, its dependents, and their syntactic relation. The node of the dependency-based syntax pattern are phrases (NP, VP, etc.), POS tags, or lexical and constrained by the syntax and semantic features.

Our Korean generation module is to morphologically generate Korean clauses from the transferred Korean trees, and combine the clauses using Korean connective words. In this module, we focus on morphologically ordering the nodes of the transferred tree with locating adverb and on generating surface forms of each node with case marker and modality generation.

3 Customization of MT system to a Technical Domain

3.1 Customization Steps

We first studied previous researches to find out commonly used customization steps. The

[1] Those values are calculated by sum of two word log frequency

purpose of previous studies was twofold: 1) to figure out whether previous customization steps would help our situation. 2) to explore the possibility of treating other similar resources as patent domain resource.

Zajac (2003) and Choi (2007) proposed customizing a general MT system to specific domain. Two previous studies consider the whole steps of customization as follows:

- Step1: Collecting a large scale of domain-specific documents
- Step2: Linguistically studying about characteristics of the collected documents
- Step3: Automatically extracting unknown words and semi-automatically constructing their equivalent words
- Step4: Manually/Semi-automatically tuning or constructing domain-specific translation knowledge (pattern, terminology etc.).
- Step5: Customizing the translation engine module.
- Step6: Human evaluation and automatic evaluation of translation performance.

In our work, we also followed the above six steps to customize our MT system to technical domain, but with some modification or extension of each step.

3.2 Collecting a large number of domain documents

Based on the above steps, we can understand the first task of customization approach is to collect enough domain documents. We use our web crawler to collect the Chinese technical documents such as technical reports, manual and papers from the Chinese web. Technical web news are also one of the good candidates because they can easy to collect and have a lot of similar vocabularies and their writing style is different from that of other document. In our approach the customized technical domain MT system is used to test Chinese patent domain; thus in this paper, we only use the technical news documents to construct the bilingual dictionary.

Since we don't have explicit URL resource and search keywords at the beginning, we first simulated by using manual documents with search keywords which are done by 2~3 persons; they give us explicit clue for automatically crawling similar documents. As result, the number of technical documents collected is almost 378,000, the number of the collected manual documents is 82,746; the number of technical reports is 154,900; and the number of papers is 140,164.

3.3 Extracting and Constructing Bilingual dictionary

Even we collected a large number of technical documents, but we still faced a big task that is how we extracting and constructing Chinese-Korean bilingual dictionary. We use our CRF-based unknown word detection tool to extract OOV(Out-Of-Vocabulary) candidates from pre-collected technical documents. As result, we get almost one million OOV candidates from OOV tool. Even we select OOV words from them we also need to get equivalent Korean words.

Kwon (2009) used an existing Korean-English bilingual dictionary to build an English-Korean bilingual dictionary in effective way. We also use our English-Korean bilingual dictionary that is extracted and constructed from huge number of English patent documents. We use English as pivot language to translate via Google translator[2]. The English terminologies are used from English-Korean bilingual dictionary. We choose English as pivot language because English to Chinese translation more correct than Korean.

Our English-Korean dictionary has almost 2 million technical terms, if we each time only use an English word, it is very time consuming work. Instead, we choose 50,000 English term list as a target translating document per times.

The next thing is we choose OOV word. We use E-C bilingual terminology filter OOV candidates and finally to get C-K bilingual dictionary by merge E-C OOV word list and E-K dictionary. As result, we constructed 518,306 of C-K bilingual dictionary.

4 Experiments on Patent Translation via the Customized Technical domain MT System

To gauge the performance of customizing a general-purpose Chinese-Korean MT system to Chinese patent domain, we carried out a series of experiments based on 200 news documents, 300

[2] http://translate.google.com

technical documents and 100 patent documents as test set.

All of translations were evaluated by 5 human translators with the scoring criteria given in Table 1. For the evaluation method, we rule out the highest and the lowest score, the scores for each sentence were summed. The method for translation accuracy (TA) was as follows:

$$TA(\%) = \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{3}(Score_j/4)\right)/3\right)/n \times 100$$

where n is the number of test sentences and $Score_j$ is the score evaluated by the *j*-th human translator.

Table 1: Evaluation criterion

Score	Criterion
4	The meaning of a sentence is perfectly conveyed
3.5	The meaning of a sentence is almost perfectly translated, except for some minor errors(e.g. wrong stylistic errors)
3	The meaning of a sentence is almost conveyed (e.g. som 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

4.1 Evaluation of Customizing to Technical domain

In this experiment, we conduct two experiments to evaluate the performance of customizing a general-purpose MT system to technical domain.

Based on the above setting we first compared our general-purpose Chine-Korean MT system on two different domains by using two test sets; news test set for news domain and technical test set for technical domain. Table 2 shows the performance comparison between two domains. The performance of translation accuracy in technical domain is decreased by 3.8% than that of news domain at the first. It makes sense because our developed system focuses on news domain.

Table 2: Comparison of two domains

Domain	Translation Accuracy
News domain	77.5%
Technical domain	73.7%

We briefly analyzed the result of the first customization. We found that only 26 documents got 4 point score and those documents rates is only 8.7%. We named this version as the baseline system.

For customizing a general-purpose MT system to technical domain, we added automatically constructed bilingual C-K dictionary, and also selected 3000 technical sentences from technical corpus for tuning set. For evaluating the trend of our tuning result, we also use automatic evaluation method using BLEU (Papineni et al., 2002). The Figure 1 shows BLEU trend at tuning period. The BLEU score increased from 0.2108 to 0.2210. In the tuning period, we focus on technical word insertion and Korean terms adaptation, modified proper noun detection, measurement and conjunction word processing, each of them increased BLEU score 0.52, 0.18 and 0.33 %.

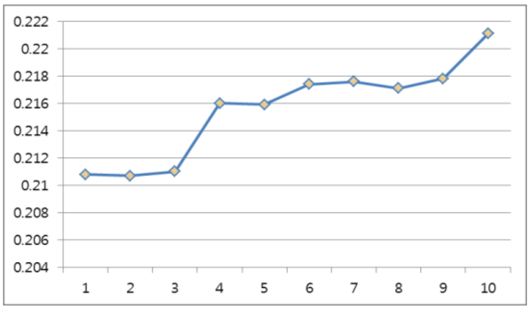


Figure 1: BLUE evaluation trend

After tuning the customization of technical domain MT system, we re-evaluate the system by using same technical documents test set. Table 3 shows the performance of customized technical domain MT system further improved baseline MT system by 7.1%.

Table 2: Comparison of two MT systems

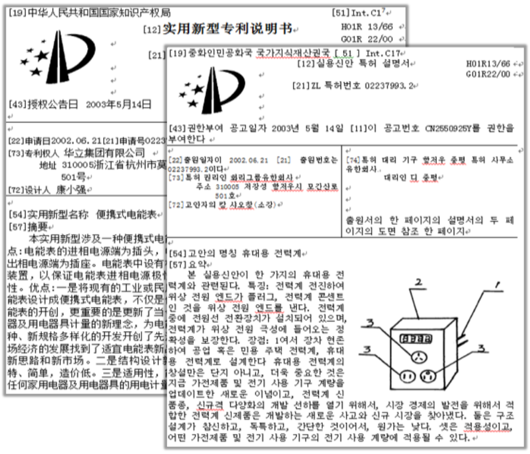
MT system	Translation Accuracy
Baseline MT system	73.7%
Improved MT system	80.8%

4.2 Evaluation of adapting to Patent Domain

In this experiment, we evaluate patent documents by using improved technical documents MT system. We only used technical documents as *pseudo-patent documents*. Nonetheless, the upgraded system shows 78.21% of translation accuracy. Even we feel that the performance over customization to patent domain is not as high as that of technical domain, but the translation accuracy score indicates that our improved MT system is suitable for patent domain.

Figure 2 shows an example of patent document translation. The main reasons of paten domain success are as follows:

- Korean terminologies originally come from Chinese words. The most common Korean nouns and verbs are directly transliterated from Chinese.
- Chinese words less ambiguous than Korean words. Due to this characteristic leads, we do not to be equipped with the full-process of word sense disambiguation.
- The tense of patent and technical domain is almost declarative, it lead us simply generate the lack of Chinese tense information and to simply generate declarative form.


[19]中华人民共和国国家知识产权局 [51]Int.Cl⁷

[12]实用新型专利说明书

H01R 13/66
G01R 22/00

[21]

[43]授权公告日 2003年5月14日

[22]申请日2002.06.21[21]申请号0223
[73]专利权人 华立集团有限公司
地址 310005浙江省杭州市莫
501号
[72]设计人 康小强

[54]实用新型名称 便携式电能表
[57]摘要
本实用新型涉及一种便携式电
点:电能表的进相电源端为插头，
出相电源端为插座。电能表中设有
装置，以保证电能表进相电源极
性。优点:一是将现有的工业或民
能表设计成便携式电能表，不仅是
能表的开创，更重要的是更新了当
器及用电器具计量的新理念，为电
种、新规格多样化的开发开创了先
场经济的发展找到了适宜电能表新
新思路和新市场。二是结构设计
特、简单，造价低。三是适用性，
任何家用电器及用电器具的用电计

[19]중화인민공화국 국가지식재산권국 [51] Int.C17
[12]실용신안 특허 설명서
H01R13/66
G01R22/00
[21]ZL 특허번호 02237993.2

[43]권한부여 공고일자 2003년 5월 14일 [11]이 공고번호 CN2550925Y를 권한을 부여한다

[22]출원일자이 2002.06.21 [21] 출원번호는 02237993.2이다
[73]특허 권리인 화리그룹유한회사
주소 310005 저장성 항저우시 모간산로
501호
[72]고안자의 캉 샤오창(소강)

[74]특허 대리 기구 항저우 춘평 특허 사무소 유한회사.
대리인 디 춘평

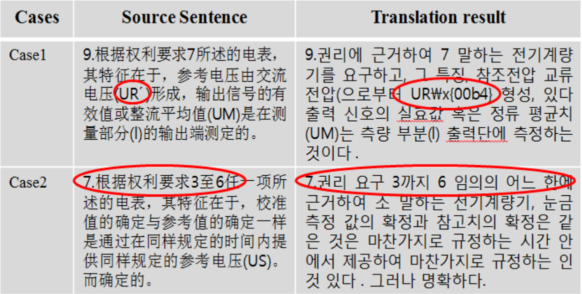
출원서의 한 페이지의 설명서의 두 페이지의 도면 참조 한 페이지

[54]고안의 명칭 휴대용 전력계
[57]요약
본 실용신안이 한 가지의 휴대용 전력계와 관련된다. 특징: 전력계 전진하여 위상 전원 엔드가 플러그, 전력계 콘센트인 것을 위상 전원 엔드를 낸다. 전력계 중에 전원선 전환장치가 설치되어 있으며, 전력계가 위상 전원 극성에 들어오는 정확성을 보장한다. 장점: 1여서 장차 현존하여 공업 혹은 민용 주택 전력계, 휴대용 전력계로 설계한다 휴대용 전력계의 창설만은 단지 아니고, 더욱 중요한 것은 지금 가전제품 및 전기 사용 기구 계량을 업데이트한 새로운 이념이고, 전력계 신품종, 신규격 다양화의 개발 선하를 열기 위해서, 시장 경제의 발전을 위해서 적합한 전력계 신제품은 개발하는 새로운 사고와 신규 시장을 찾아냈다. 둘은 구조 설계가 참신하고, 독특하고, 간단한 것이어서, 원가는 낮다. 셋은 적용성이고, 어떤 가전제품 및 전기 사용 기구의 전기 사용 계량에 적용될 수 있다.

Figure 2: An example of Patent document Translation

The main reasons of patent domain MT system still have gap with upgraded technical domain MT system are as follows:

- Some of Chinese patent sentences contain special figures, marks, or symbols, it makes the translation hard to analyze the input sentence. For example, in the case 1 of Figure 3, the source sentence has special symbol (UR´), which leads to the translation error “UR\x{00b4}”.
- The Chinese patent claim sentences often contain compound reference claim number and figure terms, and the extremely long sentence and phrase, it inevitably cause the translation error, as case 2 of Figure 3.



Cases	Source Sentence	Translation result
Case1	9.根据权利要求7所述的电表，其特征在于，参考电压由交流电压(UR´)形成，输出信号的有效值或整流平均值(UM)是在测量部分(I)的输出端测定的。	9.권리에 근거하여 7 말하는 전기계량기를 요구하고, 그 특징, 참조전압 교류전압(으로부터 UR₩x{00b4} 형성, 있다 출력 신호의 실효값 혹은 정류 평균치(UM)는 측량 부분(I) 출력단에 측정하는 것이다.
Case2	7.根据权利要求3至6任一项所述的电表，其特征在于，校准值的确定与参考值的确定一样是通过在同样规定的时间内提供同样规定的参考电压(US)。而确定的。	7.권리 요구 3까지 6 임의의 어느 한에 근거하여 소 말하는 전기계량기, 눈금 측정 값의 확정과 참고치의 확정은 같은 것은 마찬가지로 규정하는 시간 안에서 제공하여 마찬가지로 규정하는 인 것 있다. 그러나 명확하다.

Figure 3: The cases of Error translation

5 Conclusion

In this paper, we addressed the issues related customization of patent domain that often might suffer from the lack of monolingual patent documents. In this paper, we view *patent translation* as *technical document translation* given their domain similarity in terms of their terminologies and writing styles. We first customized general-purpose Chinese-Korean MT system to technical domain. We then simply used the customized technical domain MT system to translate patent translations without any further domain adaptation. The experiment shows the customized and improved technical MT systems leads to improvements in patent domain translation.

References

Brants, T. 2000. *TnT—Statistical Part-of-Speech Tagging*. In proceedings of the sixth conference on Applied natural language processing, 224-231.

Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R. 1991. *The Mathematics of Statistical Machine Translation*. Computational Linguistics, 19(2), 263-311.

Chen, K.-J. and S.-H. Liu. 1992. *Word Identification for Mandarin Chinese Sentences*. In proceedings of the COLING92, 101-107.

Ehara Terumasa. 2007. *Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation*. MT Summit XI workshop on Patent Translation, Copenhagen, Denmark, 13-16.

Ma, W.-Y. and K.-J. Chen. 2003. *Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff*, In Proceedings of SIGHAN′03.

Oh-Woog Kwon, Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh and Young-Gil Kim. 2009. *Customizing an English-Korean Machine Translation*

System for Patent/Technical Documents Translation. PACLIC 2009, 718-725.

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

Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, Oh-Woog Kwon, Young-Gil Kim. 2008. *How to Overcome the Domain Barriers in Pattern-Based Machine Translation System*. The 22nd Pacific Asia Conference on Language information and Computation(PACLIC22), 460-466.

Zajac Remi. 2003. *MT Customization*. Machine Translation Summit IX Tutorials, New Orleans USA.