Machine Translation Summit XIV

2-6 September 2013, Nice, France

Workshop Proceedings:

The 5th Workshop on Patent Translation

Editor: Shoichi Yokoyama

EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

IAMT

Workshop Proceedings for:

The 5th Workshop on Patent Translation

(Organized at the 14th Machine Translation Summit)

Editor: Shoichi Yokoyama

Published by

The European Association for Machine Translation

Schützenweg 57

CH-4123 Allschwil / Switzerland

ISBN: 978-3-9524207-1-3

MT Summit 2013 Workshop Chair:

Svetlana Sheremetyeva

Workshop Chair:

Shoichi Yokoyama, Yamagata University, Japan

Co-Chair:

Hiroyuki Kaji, Shizuoka University, Japan

Steering Committee:

Shoichi Yokoyama, Yamagata University, Japan
Hiroyuki Kaji, Shizuoka University, Japan
Jun'ichi Tsujii, Microsoft Research Asia, China
Svetlana Sheremetyeva, Lanaconsult, Denmark
Hans Uszkoreit, Saarland University, Germany
Terumasa Ehara, Yamayashi Eiwa College, Japan
Akira Ushioda, Nara Institute of Science and Technology, Japan
Takashi Ninomiya, Ehime University, Japan
Takashi Tsunakawa, Shizuoka University, Japan
Toshimichi Moriya, Japan Patent Information Organization, Japan
Shigemasa Matsuda, Japan Patent Information Organization, Japan

Program Committee:

Steering Committee Members
Hiroshi Echizen-ya, Hokkai-Gakuen University, Japan
Xiao Rong Fan, University of Tokyo, Japan
Isao Goto, NICT, Japan
Olivier Hamon, Evaluations and Language Resources Distribution Agency, France
Kinji Hanawa, Japan Patent Information Organization, Japan
Munpyo Hong, Sungkyunkwan University, Korea
Naoto Hoshiyama, Japan Patent Information Organization, Japan
Hideki Isozaki, Okayama Prefectural University, Japan
Tsuyoshi Kakita, Japan Patent Information Organization, Japan
Philipp Koehn, University of Edinburgh, UK
Sadao Kurohashi, Kyoto University, Japan
Akira Kumano, Toshiba Solutions Corp., Japan
Jong-Hyeok Lee, Pohang University of Science and Technology, Korea
Bente Maegaard, University of Copenhagen, Denmark
Tomoharu Mitsuhashi, Japan Patent Information Organization, Japan
Mitsugu Miura, NEC Corp., Japan
Shinichiro Miyazawa, Shumei University, Japan
Tomoki Nagase. Fujitsu Labs Ltd., Japan
Tadaaki Oshio, Japan Patent Information Organization, Japan
Sayori Shimohata, Oki Electric Industry Co. Ltd., Japan
Katsuhito Sudo, NTT Communication Lab., Japan
Hirokazu Suzuki, Toshiba Corp., Japan
Benjamin Tsou, The Hong Kong Institute of Education, Hong Kong
Masashi Tsuchiya, Japan Patent Information Organization, Japan
Takehito Utsuro, University of Tsukuba, Japan
Xiangli Wang, Japan Patent Information Organization, Japan

Workshop Program

Monday, September 2, 2013

09:00-09:10 Opening Remarks: Professor Shoichi Yokoyama (Yamagata University, Japan)

09:10-09:55 Invited Talk I: Mr. Paul Schwander (Director, PD 2.8, European Patent Office) (TBD):
Machine Translation at the EPO - Removing the language barrier for patent data

09:55-10:40 Invited Talk II: Mr. Hitoshi Honda (Assistant Director, Patent Information Policy Planning Office, General Coordination Division, Japan Patent Office) (TBD):
Current Status of Machine Translation System in the JPO

10:40-11:00 Break

11:00-12:00 General Session I
11:00-11:20 Svetlana Sheremetyeva: On Integrating Hybrid And Rule-Based Components For Patent MT With Several Levels Of Output
11:20-11:40 Itsuki Toyota, Zi Long, Lijuan Dong, Takehito Utsuro, and Mikio Yamamoto: Compositional Translation of Technical Terms by Integrating Patent Families as a Parallel Corpus and a Comparable Corpus
11:40-12:00 Shoichi Yokoyama: Analysis of Parallel Structures in Patent Sentences, Focusing on the Head Words

12:00-14:00 Lunch Break

14:00-15:00 Keynote Speech: Professor Philipp Koehn (University of Edinburgh, Great Britain):
Advances in Machine Translation for Patent Translation and Beyond

15:00-15:20 Break

15:20-16:00 General Session II
15:20-15:40 Yun Jin, Oh-Woog Kwon, Seung-Hoon Na, and Young-Gil Kim: Patent Translations as Technical Document Translation: Customizing a Chinese-Korean MT System to Patent Domain
15:40-16:00 Rahma Sellami, Fatiha Sadat, and Lamia Hadrich Belguith: Exploiting Multiple Resources for Japanese to English Patent Translation

16:00-16:15 Closing Remarks

Table of Contents

Welcome from the MT Summit 2013 Workshop Chair

Svetlana Sheremetyeva 6

Preface for Fifth Workshop on Patent Translation

Shoichi Yokoyama 7

On Integrating Hybrid And Rule-Based Components For Patent MT With Several Levels Of Output

Svetlana Sheremetyeva 8

Compositional Translation of Technical Terms by Integrating Patent Families as a Parallel Corpus and a Comparable Corpus

Itsuki Toyota, Zi Long, Lijuan Dong, Takehito Utsuro, Mikio Yamamoto 16

Analysis of Parallel Structures in Patent Sentences, Focusing on the Head Words

Shoichi Yokoyama 24

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

Yun Jin, Oh-Woog Kwon, Seung-Hoon Na, Young-Gil Kim 28

Exploiting Multiple Resources for Japanese to English Patent Translation

Rahma Sellami, Fatiha Sadat, Lamia Hadrich Belguith 34

Welcome from MT Summit 2013 Workshop Chair

As a person responsible for workshop events in conjunction with the 14th Machine Translation Summit I have the great privilege, in Nice this September in welcoming the 5th Workshop on Patent Translation.

Machine translation of patents is one of the most user-required MT applications. The wealth of technology contained in patents cannot be rated high enough. Patent MT is one of the most demanding tasks for both researchers and developers due to the high complexity of patent texts.

Holding the 5th Workshop on Patent Translation does not only maintain a 2-year cycle of such events now traditionally held in conjunction with MT-Summit conferences, but more importantly address deeper topics reflecting current trends in patent world and machine translation technology. This is clearly seen in the front-end perfect workshop program.

I would like to thank a lot the chair of the 5th Workshop on Patent Translation Professor Shoichi Yokoyama, the Co-Chair Professor Hiroyuki Kaji , the members of the steering and program committees for the great amount of work and skill which has gone into the organization of this event.

Let me express my most sincere belief that the information exchange and discussions resulting from this very important full-day workshop will most positively contribute not only to the research and development in patent translation but in the whole area of MT.

And last but not least, I wish you to further enjoy the main conference and the city of Nice with its charm, history and lots to do.

Svetlana Sheremetyeva

Preface for Fifth Workshop on Patent Translation

Shoichi Yokoyama
Graduate School of Science and Engineering (Informatics), Yamagata University
4-3-16 Jonan, Yonezawa, Yamagata 992-8510, Japan
yokoyama@yz.yamagata-u.ac.jp

Welcome to the fifth workshop on patent translation. The fifth workshop will be held with MT Summit XIV on September 2, 2013, in Nice. The Chair will be Shoichi Yokoyama and the Co-Chair Hiroyuki Kaji. The keynote speech by Prof. Philipp Koehn from Edinburgh University, two invited speeches by representatives from major patent offices, and five presentations in the general session are planned.

Patent documents have become one of the major application areas of machine translation. The workshop aims to foster research and development of the technology for patent translation by providing a forum where researchers and practitioners can exchange their ideas, approaches, perspectives, and experiences from the work they are undertaking.

Topics of interests include, but are not limited to:

- MT and translation aids for patent documents
- Language resources for patent translation
- Evaluation techniques for patent translation
- Multilingual patent classification and retrieval

The workshop has been held four times as a part of every MT Summit since 2005, increasing its presence and drawing an ever larger number of participants. The brief history of the past workshops is as follows:

The first workshop (Chair: Shoichi Yokoyama) started with MT Summit X, on September 16, 2005, in Phuket, in order to discuss the special interest on patent translation from the viewpoint of users (patent offices), researchers, and makers. In the workshop, three invited speakers gave addresses, one was by Prof. Bente Maegaard from the University of Copenhagen, and two were user viewpoints from patent offices (EPO and JPO). In the general session, nine papers were published with a wrap-up meeting.

The second workshop (Co-Chair: Jun'ichi Tsujii and Shoichi Yokoyama) took place with MT Summit XI, on September 11, 2007, in Copenhagen. In the workshop, one invited talk by WIPO and two users' viewpoints from EPO and JPO were presented. In the general session, six papers were published.

The third workshop (Co-Chair: Terumasa Ehara and Shoichi Yokoyama) took place with MT Summit XII, on August 30, 2009, in Ottawa. There were two invited talks, one by EPO, and the other by CPIC. In the workshop, three talks from users' viewpoints were presented by USPTO, KIPO, and JPO. In the general session, five papers were published. In the panel discussion, five panelists under the moderator Dr. Akira Ushioda from Fujitsu Labs. participated.

The fourth workshop was held with MT Summit XIII on September 23, 2011, in Xiamen. The Chair was Shoichi Yokoyama with Co-Chairs Terumasa Ehara and Dang Wang. Three invited talks were given, the first by EPO, the second by CPIC, and the third by JPO. About 70 participants attended. Seven papers were presented in the general session. Prof. Ehara organized a panel discussion with five panelists.

The importance of patent translation has increased, along with the globalization of patents. The forthcoming fifth workshop will definitely be an excellent opportunity to share and exchange information regarding the latest technologies and trends in the patent translation domain.

On Integrating Hybrid And Rule-Based Components For Patent MT With Several Levels Of Output

Svetlana Sheremetyeva
South Ural State University/ 76 Lenin pr. Chelyabinsk 454080, Russia
LanA Consulting ApS / Moellekrog 4, Vejby, Copenhagen, Denmark
lanaconsult@mail.dk

Abstract

We present a methodology integrating hybrid and rule-based components for speeding up the development of a patent MT system. The methodology is suitable for highly inflecting languages and described on the example of translating patent claims from Russian into English. Based on different combinations of hybrid and rule-based components the system performs shallow or/and deep parsing and provides for several complementary levels of output, - (i) translation of terminology, that only involves shallow MT procedures, and (ii) full translation that is based on both shallow and deep parsing integrated either automatically, or in an interactive environment. Full translation of the patent claim is output in two formats, - a legal one sentence format and a better readable set of simple sentences. To control the quality of claim translation by better understanding the input, the system also outputs a SL claim decomposed into simple sentences.

1 Introduction

The wealth of technology contained in patents cannot be rated high enough. With ever exploding volume of patent documentation machine translation contributes a lot to strengthening the innovation process worldwide, removing language as a delimiting factor. In patent domain machine translation is a very challenging task. Only high quality patent translation could be used as a basis for important decisions on, e.g., novelty or the scope of inventor's rights. Quality requirement prompted the development of patent RBMT (Shimohata, 2005; Hong et al., 2005; Sheremetyeva, 2007; Wen and Jin, 2011) whose techniques promise correct translation but demand huge linguistic resources.

In an attempt to speed up the process of MT development and make it more robust, SMT and hybrid technologies (Ceausu et al., 2011; Eisele et al., 2008; Ehara, 2011; Espana-Bonet et al., 2011; Enache, 2012) came into patent domain. Though years of R&D in MT have resulted in great progress, the output of machine translation still cannot provide for required quality without human judgment (Koehn, 2009). In addition to traditional postediting recent work investigated the inclusion of interactive computer-human communication at each step of the translation process by, e.g., showing the user various "paths" among all translations of a sentence (Koehn, cf), or keyboard-driving the user to select the best translation (Macklovitch, 2006). One of the latest publications reports on Patent SMT from English to French where the user drives the segmentation of the input text (Pouliquen et.al, 2011). Popular SMT and hybrid techniques are problematic when dealing with inflecting languages. Statistical components of MT systems working well on configurational and morphologically poor languages, such as English, fail on non-configurational languages with rich morphology (Sharoff, 2004).

This paper reports on a novel hybrid methodology for developing an efficient patent MT system that can cope with translating patent claims. The methodology focuses on a highly inflecting SL and based on different combinations of hybrid and rule-based components provides for several complementary levels of output, - translation of terminology and full text translation in different formats. To improve the quality of full translation, the system includes an interactive module. To support quality control of

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

claim translation the system helps the user to better understand the input by decomposing a SL claim into simple sentences thus improving its readability. Different levels of output and possible interactivity make the MT system useful for different types of users: TL-only speakers, SL-only speakers, people with some knowledge of both languages and professional translators.

The methodology is described on the model of a Russian-to-English MT system. In selecting Russian as our first inflecting SL we were motivated by two major considerations. Firstly, Russia has a huge pool of patents which are unavailable for non-Russian speakers without turning to expensive translation services. The situation is of great disadvantage for international technical knowledge assimilation, dissemination, protection of inventor's rights and patenting of new inventions. Secondly, Russian is an ultimate example of a highly inflecting language with a free word order. A typical Russian word has from 9 (for nouns) up to 50 forms (for verbs), which makes Russian a good testbed for hybrid MT covering inflecting languages.

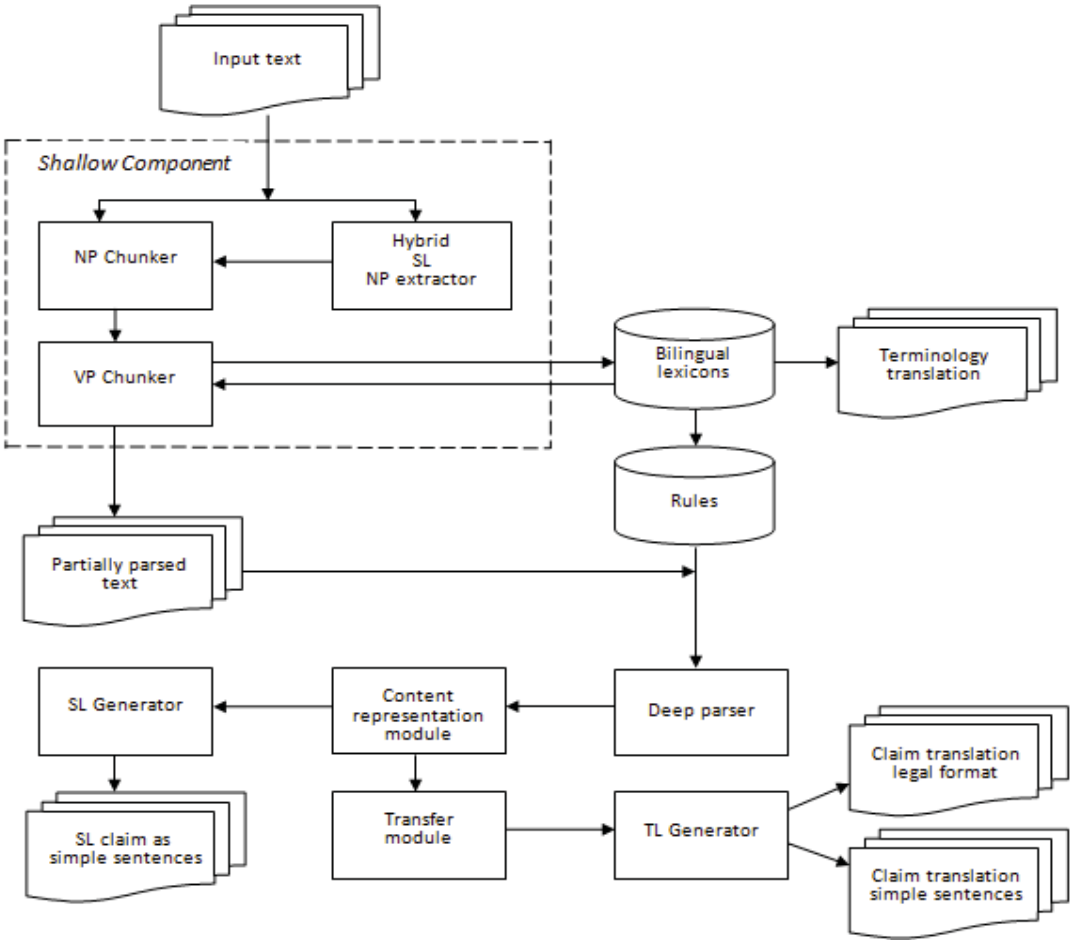

Figure 1. An overall architecture of the hybrid patent MT system with different levels of output.

2 System Overview

The system takes a Russian patent claim as input and produces translations at two major levels, - translation of claim terminology (not just any chunks), and full translation of a patent claim. Full translation of a patent claim is output in two formats, - in the form of one sentence meeting all legal requirements to the claim text, and as a better readable set of simple sentences. In addition, for the translator/posteditor to better control the quality of translation the system also improves the readability of a SL claim by decomposing it into a set of simple sentences. Partial output of the first translation level is

useful for a non-Russian speaker for a quick patent digest to make a decision whether a full translation of a patent is needed. A list of translated terms is also useful for improving readability of a full claim translation[1].

This research extends our previous work on an RBMT system for translating patent claims between the low inflecting configurational English and Danish languages (Sheremetyeva, 2007). It partially reuses the program shell and some of the linguistic knowledge of its RB components. Necessary updates are made for the Russian language. The top architecture of the system follows the traditional RBMT schema, - SL analyzer – transfer -TL generation, but instead of a fully rule-based analyzer the current system includes a hybrid parser with shallow and deep components that lifts a lot of ambiguity problems and makes the whole parsing easier, less resource consuming and more robust. The core of the shallow parser is a hybrid Russian NP extractor which is a standalone tool that was integrated into the system. The full Russian parser and transfer module are designed so as to produce a final parse of a Russian patent claim in the format acceptable by the English claim generator from the earlier application. The architecture of the system is shown in Figure 1.

3 Knowledge

Patent claims must be formulated as specified by the German Patent Office and commonly accepted in Europe, the U.S., Russia and other countries. The claim must describe essential features of the invention in the obligatory form of a single extended nominal sentence with a well-specified conceptual, syntactic and stylistic/rhetorical structure.

For successful translation of patent claims two distinct types of expert knowledge are necessary: knowledge about the sublanguage of patents as legal documents and knowledge about the technical field of the invention. Both kinds of knowledge are mainly encoded in the lexicons:

(i) **a shallow bilingual (Russian/English) lexicon**, where the units are listed with their morphological features. This is the type of resource that, once build for some other purpose, can be simply fed into the system. We had a successful experience of pipelining such knowledge into an MT system in our Japanese-English project (Neumann, 2005). Acquisition of this type of knowledge for every new pair of languages is what existing SMT tools can provide either in advance or on the fly, as reported in (Enache et al., cf). We, therefore, do not dwell on acquisition of this type of resource. To demonstrate the viability of the methodology we will use our own limited semi-manually acquired set of bilingual terminological data.

(ii) **a deep (information-rich) bilingual lexicon of predicates** used in the English and Russian language patent claims; this lexicon has been specifically constructed for the current application and is meant for a multifunctional use in the modules of the system.

3.1 Deep lexicon and content representation language

The core of the system knowledge is a deep bilingual Russian-English predicate lexicon which is organized as a set of cross-referenced set of monolingual entries and contains lexical, morphological, syntactic and semantic knowledge. Syntactic and semantic zones are as follows:

CASE_ROLEs, - a set of lexeme case roles such as *agent, theme, place, instrument*, etc.

FILLERs – lexical categories that can fill case-role slots of a lexeme;

PATTERNs - code both the co-occurrences of predicates with their case-roles, and their linear order in the claim text.

Figure 2 presents a fragment of the entry of the predicate "mounted" with the case-roles and pattern zones.

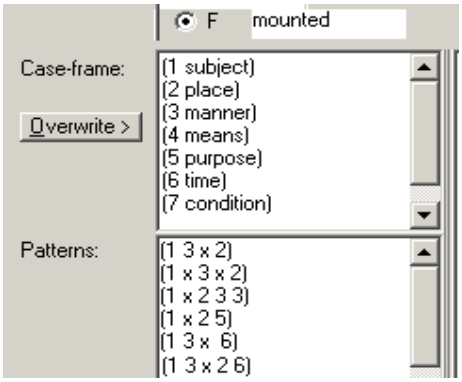

Figure 2. a fragment of the entry of the predicate "mounted".

1 It is well known that due to an extremely complex syntactic structure of the patent claim that can run for a page or so, the problem of patent readability is an issue even in a SL (Shinmori et al.,2003), let alone in translation.

The pattern (1 3 x 2), for example, can trig the realization of such clam fragment as

1:devices 3:rotatably x:mounted 2:on the leg.

3.2 Content representation language

The knowledge in predicate entries is used to support the claim content representation language as shown below.

*Sentence::={ template}{template}**
template::={label predicate-class predicate ((caserole)(case-role))}*
case-role::= (rank status value)
value::= phrase{(phrase(word tag))}*,*

where "label" is a unique identifier of a predicate/argument structure, "predicate-class" is a label of a semantic class, "predicate" is a string corresponding to a predicate, "case-roles" are "ranked" according to the frequency of their co-occurrence with a certain predicate in the training corpus, "status" is a semantic status of a case-role (*place, instrument*, etc.,) and "value" is a case-role filler.

4 Hybrid parser

4.1 Shallow component

Russian NP extractor. The core of the shallow parsing component is a hybrid Russian NP extractor which is a standalone tool[2] pipelined to the system. It was built following the methodology of keyword extraction for the English language described in (Sheremetyeva 2009). The extractor does not rely on a preconstructed corpus, works on small texts, does not miss low frequency units and can reliably extract all NPs from an input text. The extraction methodology combines statistical techniques, heuristics and very shallow linguistic knowledge that includes a number of shallow lexicons (sort of extended lists of stop words) forbidden in a particular (first, middle or last) position in the typed unit (Russian NP in our case) to be extracted.

NP extraction starts with n-gram calculation and then removes n-grams which cannot be NPs by matching components of calculated n-grams against the stop lexicons. The extraction itself thus neither requires such NLP procedures, as tagging, morphological normalization, etc., nor does it rely on statistical counts (statistical counts are only used to sort out keywords). The latter makes this extraction methodology suitable for inflecting languages (Russian in our case) where frequencies of n-grams are low.

Porting the NP extractor from English to Russian consisted in substituting English stop lexicons of the tool with the Russian equivalents. We did this by translating each of the English stop lists into Russian using a free online system PROMT followed by manual brush-up. The extracted NP phrases are of 1 to 4 components due to the limitations of the extractor 4-gram model. We did not lemmatize the output of the extractor. All extracted Russian NP strings keep their text forms. This allows straightforward bracketing of these NPs in the claim text by simple matching the extracted NPs against the text. The remaining unbraketed text of the input is then matched against the morphological fields of the predicate entries in the predicate lexicon and, in case of a match, a predicate is chunked (bracketed) in the input text. This practically lifts the problem of lexical ambiguity between the forms of verbs and other parts-of-speech. Being identified as NP components and enclosed in brackets a lot of ambiguous words are simply not submitted to the VP chunker. The order of NP and VP (predicates) chunking is relevant. Noun phrases are chunked first as they are the most frequent types of phrases and thus leave less "residue text" for VP identification reducing the ambiguity.

The output of the shallow parsing components is then submitted to the bilingual lexicon with the help of which the first (partial) level of translation is performed. An example of such partial translation is shown in Figure 3 (right, bottom). If the goal of the user is just a digest, the MT procedure can stop right here. Otherwise, the shallow parse is input to the deep component that have two modes, - automatic and interactive.

4.2 Deep component

Automatic mode. The deep component takes a partially parsed claim from the shallow component as input and automatically completes the parsing procedure. It uses the knowledge from the deep lexicon and rules of our application-specific grammar, - a mixture of context free lexicalized Phrase Structure Grammar and Dependency Grammar.

[2] This tool can be used for different purposes, e.g., we also used its English and Russian versions for the acquisition of Russian-English lexicon by running it on available parallel and comparable corpora

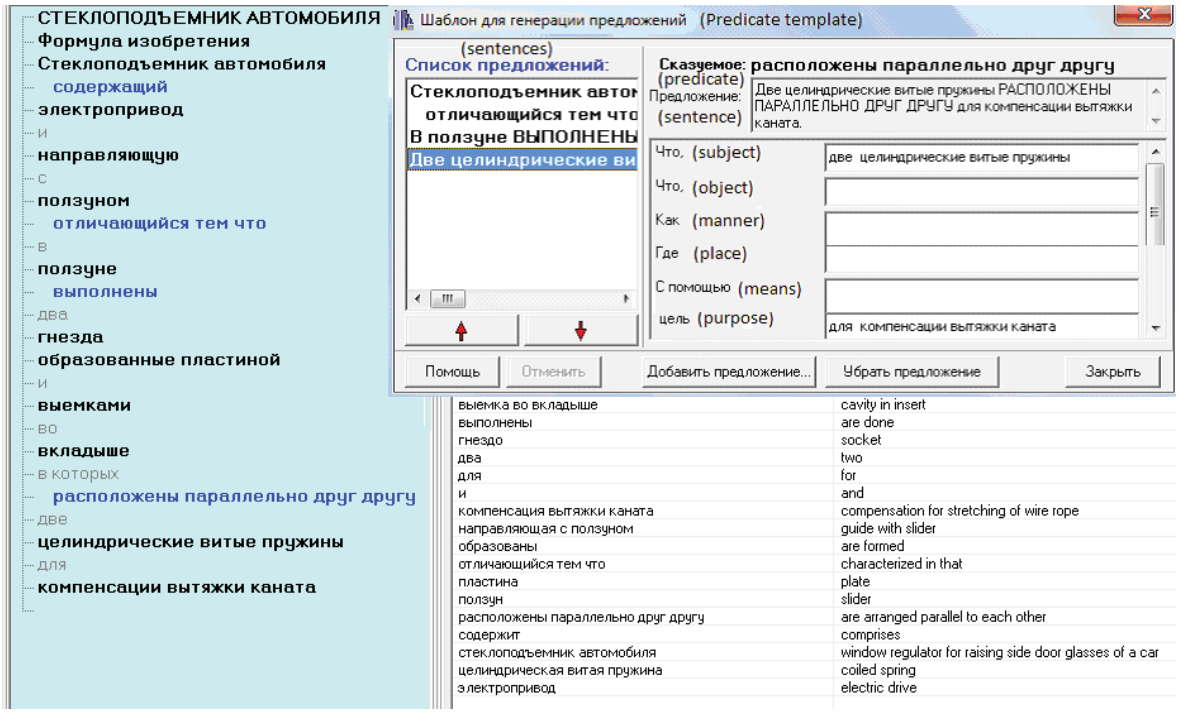

Figure 3. A screenshot of the user interface in the interactive mode. In the left pane a chunked input with highlighted NPs and VPs is displayed. On the bottom of the right pane the first level of translation results are shown. On the top in the right pane an interactive predicate template is presented which pops-up after the user clicks on a corresponding highlighted predicate.

The deep parser includes newly developed components as a Russian disambiguating tagger, a Russian bottom-up heuristic parser with a recursive pattern matching technique to recursively chunk all types of Russian phrases (NPs[3], PPs, AdvP, etc.). It preserves the inner structure of the longer chunks and marks the head of every noun phrase with its "singular/plural" feature.

At its last run the parser determines the semantic dependency relations between the identified chunks and predicates, - and assigns to the chunks their semantic status as particular case-roles of the governing predicate (see Sheremetyeva, 2003 for details).

Shallow "pre-parsing" significantly reduces all kinds of ambiguity at all stages of processing and decreases the number of rules. The final parse is then supplied into the Transfer and SL generation module to get a full translation of the claim. The final parse of the claim text displayed in the left pane of Figure 3 is shown in Figure 4 (left pane).

Interactive mode. In the interactive mode of the deep parsing component the system guides the user through the paces of "understanding" the structure of the SL claim by decomposing a complex input text into predicate phrases (simple sentences) describing individual features of the invention "disguised" in the complex telescopic claim structure. The system supports user elicitation decisions with instructions, and highlighted SL noun terms and predicates (see Figure 3, left pane). Once the user clicks on a highlighted predicate, a corresponding elicitation template is displayed in a separate pop-up window (Figure 3, top). The template is based on the knowledge in the case-role zone in the lexicon entry of the selected predicate. The user then fills the slots of the template with text elements by simply clicking on them in the interactively marked (chunked) input document. The slot fillers can be edited by supplying chunks of the text into the slots of predicate templates the user determines the dependency relations between the predicates and other chunks and defines the semantic status of these chunks as case-roles of the governing predicate. The output of this interaction proce-

[3] The deep parser combines NPs chucked by the shallow component into longer nominal phrases.

dure is a set of predicate/argument structures with partially parsed case-role fillers, which are further input into the deep parsing component. The deep parser automatically completes case-role fillers tagging and recursive chunking and outputs a set of predicate/argument structures as shown in Figure 4 (left pane). This content representation is then submitted into two system modules, – the Russian generator that outputs a Russian claim in a more readable format of simple sentences, and to the Russian-English transfer module.

4.3 Transfer module

The transfer module is fully automatic. It takes the deep parser output, - a SL set of predicate templates as input and outputs a set of TL predicate templates whose slots are filled with translated TL phrases (case-role fillers). The transfer procedure is a combination of interlingual and lexical-syntactic transfer. The interlingual transfer is based on the knowledge about predicate case-roles in the deep lexicon. It finds structural TL equivalents for every SL predicate/argument structure. The TL predicate gloss is substituted with its TL equivalent. Then the SL fillers of the case-roles are translated (See Figure 4). A "real" translation procedure is thus reduced to the phrase level which, though not without problems, is still much simpler than machine translation of a full patent claim. Translation of case-role fillers can be outsourced to a foreign MT system and then put back into a predicate-argument structure. As was mentioned above this is where SMT techniques can be particularly useful.

4.4 Generation module

The claim text generation stage takes an English-oriented text representation (Figure 4, right pane) as input, and submits it to an automatic text planner which outputs a hierarchical structure of predicate templates.

The planning stage is guided both by constraints on the patent claim sublanguage and the general constraints on style. The former determines the global ordering of the claim text while the latter deals with local text coherence.

The realization stage of the generator linearizes the hierarchy of TL predicate templates and takes care of the ellipsis, conjoined structures, punctuation and morphological forms. The generic part and novelty part of the claim are generated separately.

The two completely ready parts of the claim text are bound by the intermediate expression "characterized in that", the generic and novelty parts being put correspondingly before and after this expression. The output is an English text of the claim in a legal format (see Figure 5, top). Parallel to this a better readable translation of the same claim in the form of simple sentences is also generated (Figure 5, bottom).

5 Status and Discussion

The methodology we have described in this paper has been implemented in a Russian-English hybrid MT system for patent claims. The system is in the late stages of development as of June 2013. The static knowledge sources have been compiled for the domain of patents about vehicles. The programming shell of the system is completed and provides for knowledge administration in all modules of the system to improve their performance. The extractor of Russian nominal terminology currently preforms with 98,4 % of recall and 96,1% precision.

The shallow clunker based on the extraction results and predicate knowledge shows even higher accuracy. This is explained, on the one hand, by the high performance of the Russian extractor, and, on the other hand, by the nature of highly inflecting languages. Rich morphology turns out to be an advantage in our approach. Great variety of morphological forms significantly lowers ambiguity between NP components and verb paradigms. We have tested shallow chunking on patents in English and, though the efficiency of English and Russian NP extractors is practically the same, chunking of the English NPs in claim texts is rather problematic due to much higher ambiguity of wordforms in English. Our conclusion is that shallow chunking based on unlemmatized extraction results better suites inflecting languages.

The interactive semantic and syntactic analysis module for the Russian language and the English generator are fully developed using the technology of earlier applications. The Russian-to-English transfer module responsible for lexical transfer and case-role translation is workable.
In the deep parsing component the morphological analysis of Russian and syntactic chunking are operational and well tested. The case-role dependency detection in the automatic mode is being currently tested and updated. We have not yet made a large-scale evaluation of our deep

analysis module. This leaves the comparison between other parsers and our approach as a future work. In general preliminary MT results show a reasonably small number of failures that are being improved by brushing up the shallow knowledge and by larger involvement of predicate knowledge. This proves the viability of the suggested MT methodology. We intend to a) improve the quality of the automatic mode of our MT system by updating system knowledge based on extensive testing; b) develop a patent search and extraction facility on the basis of the patent sublanguage and our parsing strategy.

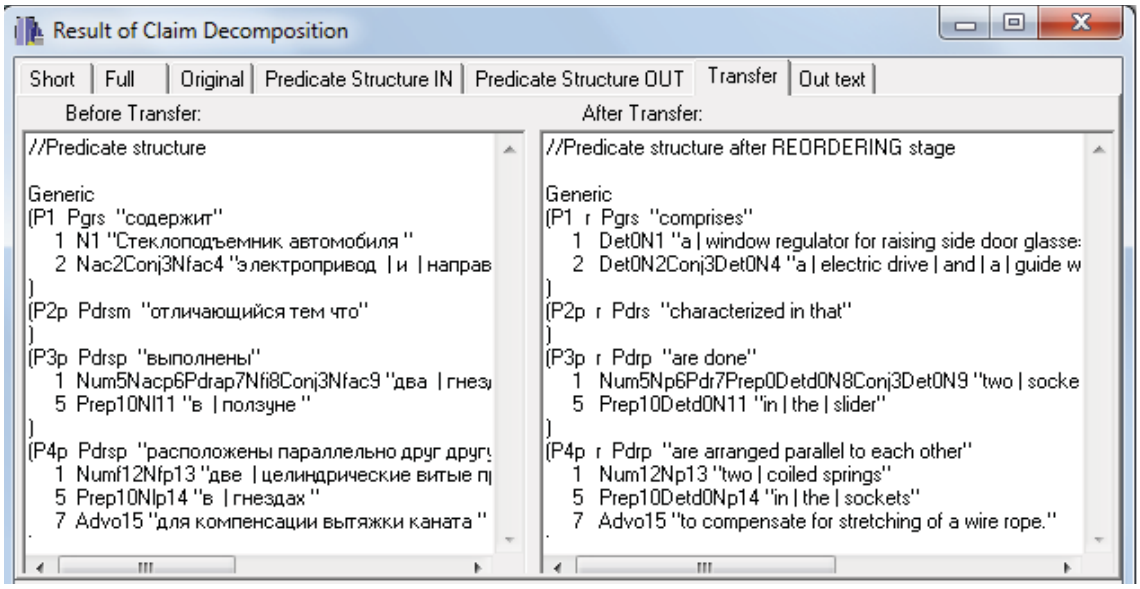

Figure 4. A screenshot of the developer's interface. On the left pane shown is the final parse of the example claim shown in Figure 3 (left pane). The output of the transfer module is shown on the right.

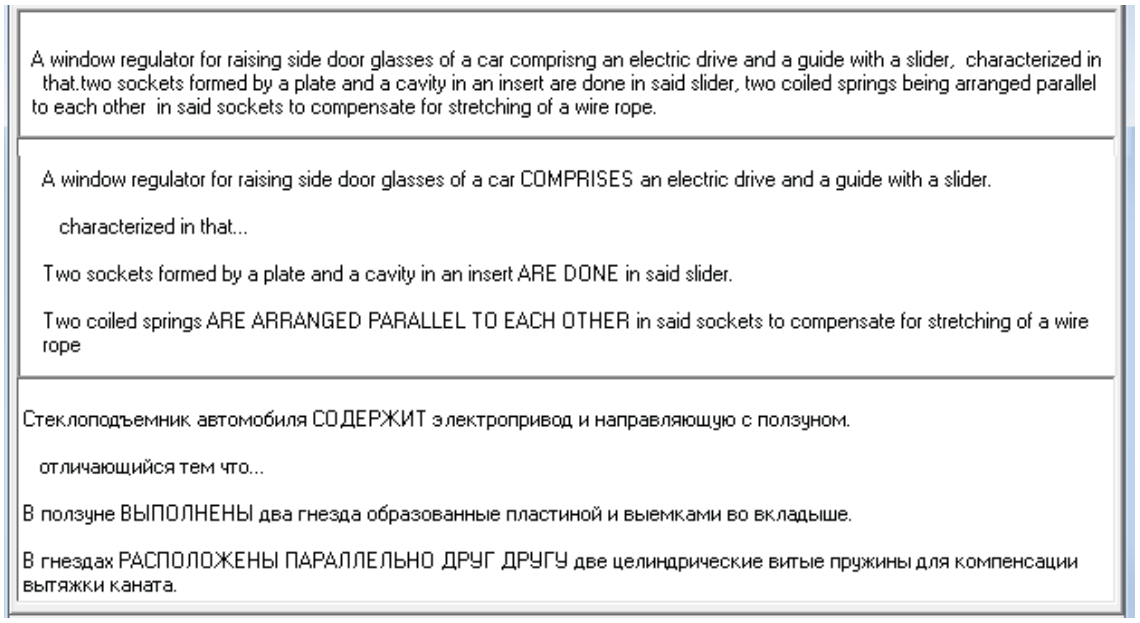

Figure 5. Examples of MT output. On the top a full claim translation into English in the legal format is shown. In the middle the "better readable" claim translation in the form of simple sentences is shown. In the bottom a decomposed Russian input claim is given.

References

Ceausu, Alexandru, John Tinsley, Jian Zhang, and Andy Way. 2011. *Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project*. In Proceedings of EAMT 2011:

Ehara Terumasa, 2011 Machine translation system for patent documents combining rule-based translation and statistical postediting applied to the Patent MT Task. *Proceedings of NTCIR-9 Workshop Meeting,* 2011, Tokyo, Japan

Eisele, A., C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. 2008. Using moses to integrate multiple rule-based machine translation engines into a hybrid system. *In Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08.

Enache Ramona, Cristina Espa˜na-Bonet Aarne Ranta Llu´ıs M`arquez. 2012. A Hybrid System for Patent Translation. *Proceedings of the EAMT Conference.* Trento..Italy, May

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.

Koehn Philipp. 2009. A process study of computer-aided translation, *Philipp Koehn*, Machine Translation Journal, 2009, volume 23, number 4, pages 241-263

Macklovitch, Elliott. 2006. TransType2: The last word. *In proceedings of LREC06*, May 2006,Genoa, Italy

Neumann Ch. 2005. A Human-Aided Machine Translation System for Japanese-English Patent Translation. *Proceedings of the Workshop on Patent Translation MT Summit*, Phuket, Thailand,.

Pouliquen Bruno, Christophe Mazenc Aldo Iorio. 2011. Tapta: A user-driven translation system for patent documents based on domain-aware Statistical Machine. *Proceedings of the EAMT Conference*. Leuven, Belgium, May.

Sharoff, Serge . 2004. What is at stake: a case study of Russian expressions starting with a preposition. *Proceedings of the ACL Workshop on Multiword Expressions: Integrating Processing*, July.

Sheremetyeva Svetlana. 2003. Natural Language Analysis of Patent Claims. *Proceedings of the workshop "Patent Corpus Processing" in conjunction with ACL 2003,* Sapporo. Japan, July.

Sheremetyeva Svetlana. 2007. On Portability of Resources for a Quick Ramp up of Multilingual MT of Patent Claims. *Workshop on Patent Translation. In conjunction of Machine Translation Summit XI.Copenhagen*.Denmark. September

Sheremetyeva, Svetlana. 2009. On Extracting Multiword NP Terminology for MT. *Proceedings of the EAMT Conference.* Barcelona, Spain, May

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

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

Wen Xiong, Yaohong Jin. 2011. A new Chinese-English machine translation method based on rule for claims sentence of Chinese patent; In Proceedings of 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE).

Compositional Translation of Technical Terms by Integrating Patent Families as a Parallel Corpus and a Comparable Corpus

Itsuki Toyota Zi Long Lijuan Dong
Grad. Sch. Sys. & Inf. Eng.,
University of Tsukuba,
Tsukuba, 305-8573, JAPAN

Takehito Utsuro Mikio Yamamoto
Fclty of Eng., Inf.& Sys.,
University of Tsukuba,
Tsukuba, 305-8573, JAPAN

Abstract

In the previous methods of generating bilingual lexicon from parallel patent sentences extracted from patent families, the portion from which parallel patent sentences are extracted is about 30% out of the whole "Background" and "Embodiment" parts and about 70% are not used. Considering this situation, this paper proposes to generate bilingual lexicon for technical terms not only from the 30% but also from the remaining 70% out of the whole "Background" and "Embodiment" parts. The proposed method employs the compositional translation estimation technique utilizing the remaining 70% as a comparable corpus for validating translation candidates. As the bilingual constituent lexicons in compositional translation, we use an existing bilingual lexicon as well as the phrase translation table trained with the parallel patent sentences extracted from the 30%. Finally, we show that about 3,600 technical term translation pairs can be acquired from 1,000 patent families.

1 Introduction

For both high quality machine and human translation, a large scale and high quality bilingual lexicon is the most important key resource. Since manual compilation of bilingual lexicon requires plenty of time and huge manual labor, in the research area of knowledge acquisition from text, automatic bilingual lexicon compilation have been studied. Techniques invented so far include translation term pair acquisition based on statistical co-occurrence measure from parallel sentences (Matsumoto and Utsuro, 2000), translation term pair acquisition from comparable corpora (Fung and Yee, 1998), transliteration (Knight and Graehl, 1998), compositional translation generation based on an existing bilingual lexicon for human use (Tonoike et al., 2006), and translation term pair acquisition by collecting partially bilingual texts through the search engine (Huang et al., 2005).

Among those efforts of acquiring bilingual lexicon from text, Morishita (2008) studied to acquire technical term translation lexicon from the phrase translation table, which are trained by a phrase-based statistical machine translation model with parallel sentences automatically extracted from patent families. We further studied to require the acquired technical term translation equivalents to be consistent with word alignment in parallel sentences and achieved 91.9% precision with almost 70% recall. This technique has been actually adopted by a Japanese organization which is responsible for translating Japanese patent applications published by the Japanese Patent Office (JPO) into English, where it has been utilized in the process of semi-automatically compiling bilingual technical term lexicon from parallel patent sentences. In this process, persons who are working on compiling bilingual technical term lexicon judge whether to accept or not candidates of bilingual technical term pairs presented by the system. According to our personal communication with the organization, under a certain amount of budget for the labor of judging the correctness of bilingual technical term pairs suggested by the system, the organization collected about 500,000 bilingual technical term pairs per year. The orga-

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

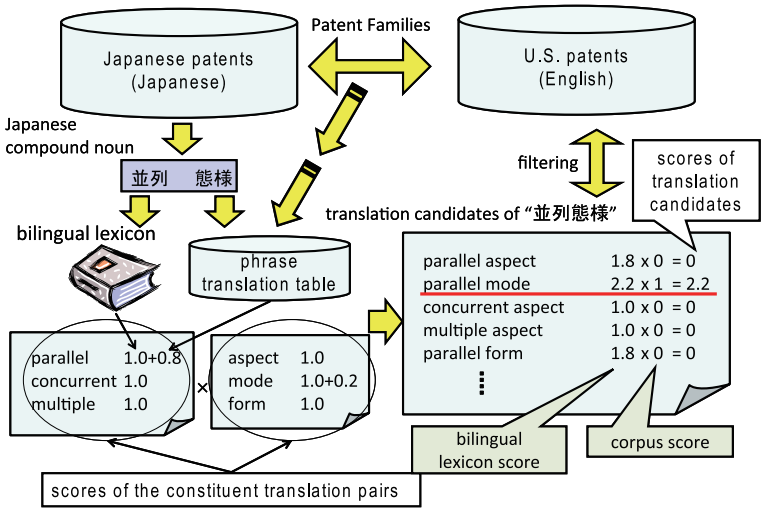

Figure 1: Proposed Framework of Compositional Translation Estimation for the Japanese Technical Term "並列態様" *(parallel mode)*

nization is also working on the task of compiling a Japanese-Chinese bilingual technical term lexicon from Japanese-Chinese patent families, where they claim that, under a certain amount of budget, they are able to compile 1,000,000 bilingual technical term pairs per year.

In Morishita (2008), the portion from which parallel patent sentences are extracted is composed of the parts of "Background" and "Embodiment". However, this portion is about 30% out of the whole "Background" and "Embodiment" parts and about 70% are not used. Considering this situation, this paper proposes to generate bilingual lexicon for technical terms not only from the 30% but also from the remaining 70% out of the whole "Background" and "Embodiment" parts. As shown in Figure 1, the proposed method employs the compositional translation estimation technique utilizing the remaining 70% as a comparable corpus for selecting translation candidates that actually appear in the target language side of the comparable corpus. As the bilingual constituent lexicons, the compositional translation procedure uses an existing bilingual lexicon as well as the phrase translation table trained with the parallel patent sentences extracted from the 30%. Through the experimental evaluation, we show that about 3,600 technical term translation pairs can be acquired from 1,000 patent families.

2 Related Work

Lu and Tsou (2009) and Yasuda and Sumita (2013) studied to extract bilingual terms from comparable patents, where, as we studied in Morishita (2008), they first extract parallel sentences from comparable patents, and then extract bilingual terms from parallel sentences. As we discussed in section 1, in this paper, we concentrate on generating bilingual lexicon for technical terms not only from the parallel patent sentences extracted from patent families, but also from the remaining parts of patent families.

Liang et al. (2011) considered situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents. They then studied the issue of identifying synonymous translation equivalent pairs. The technique proposed in this paper can be easily integrated into the achievement presented in Liang et al. (2011) in the task of identifying synonymous translation equivalent pairs.

The task of translation term pair acquisition from comparable corpora (e.g., (Fung and Yee, 1998)) has been well studied, where most of those works rely on measuring contextual similarity of translation term pair candidates across two languages. Compared with those techniques, our proposed method relies on the compositional translation approach utilizing patent families. Patent families can be regarded as a partially parallel and partially comparable corpus, where a relatively large portion of technical terms are compositionally translated across two languages, and in those cases, translation candidates can be easily detected without introducing contextual similarity.

3 Japanese-English Patent Families

In the NTCIR-7 workshop, the Japanese-English patent translation task is organized (Fujii et al., 2008), where patent families and sentences are provided by the organizer. Those patent families are collected from the 10 years of unexamined Japanese patent applications published by the Japanese Patent Office (JPO) and the 10 years patent grant data published by the U.S. Patent & Trademark Office (USPTO) in 1993-2000. The numbers of documents are approximately 3,500,000 for Japanese and 1,300,000 for English. Because the USPTO documents consist of only patent that have been granted, the number of these documents is smaller than that of the JPO documents.

From these document sets, patent families are automatically extracted and the fields of "Background of the Invention" and "Detailed Description of the Preferred Embodiments" are selected. This is because the text of those fields is usually translated on a sentence-by-sentence basis. Then, the method of Uchiyama and Isahara (2007) is applied to the text of those fields, and Japanese and English sentences are aligned (about 1.8M sentences in total).

4 Compositional Translation of Technical Terms

As the procedure of compositional translation of technical terms, translation candidates of a term are compositionally generated by concatenating the translation of the constituents of the term (Tonoike et al., 2006)[1] [2].

4.1 Bilingual Constituents Lexicons

First, the following sections describe the bilingual lexicons we use for translating constituents of technical terms, where Table 1 shows the numbers of entries and translation pairs in those lexicons.

[1]Tonoike et. al (2006) studied how to compositionally translate technical terms using an existing bilingual lexicon as well as bilingual constituent lexicons constructed from the constituents collected from the existing bilingual lexicon. Compared to Tonoike et. al (2006), this paper proposes how to optimally incorporate constituent translation pairs collected from the phrase translation table trained with the parallel patent sentences introduced in section 3 into the procedure of compositional translation.

[2]As "constituents", we do not consider "syntactic constituents", but simply consider a word or a sequence of two or more consecutive words.

4.1.1 A Bilingual Lexicon (Eijiro) and its Constituent Lexicons

As an existing Japanese-English translation lexicon for human use, we use Eijiro (`http://www.eijiro.jp/`, We merged two versions Ver.79 and Ver. 131.).

We also compiled bilingual constituents lexicons from the translation pairs of Eijiro. Here, we first collect translation pairs whose English terms and Japanese terms consist of two constituents into another lexicon P_2. We compile the "bilingual constituents lexicon (prefix)" from the first constituents of the translation pairs in P_2 and compile the "bilingual constituents lexicon (suffix)" from their second constituents[3].

4.1.2 Phrase Translation Table of an SMT Model

As a toolkit of a phrase-based statistical machine translation model, we use Moses (Koehn and others, 2007) and apply it to the whole 1.8M parallel patent sentences described in section 3. In Moses, first, word alignment of parallel sentences are obtained by GIZA++ (Och and Ney, 2003) in both translation directions and then the two alignments are symmetrised. Next, any phrase pair that is consistent with word alignment is collected into the phrase translation table and a phrase translation probability is assigned to each pair (Koehn et al., 2003). We finally obtain 76M translation pairs with 33M unique Japanese phrases, i.e., 2.29 English translations per Japanese phrase on average, with Japanese to English phrase translation probabilities $P(p_E \mid p_J)$ of translating a Japanese phrase p_J into an English phrase p_E. For each Japanese phrase, those multiple translation candidates in the phrase translation table are ranked in descending order of Japanese to English phrase translation probabilities.

4.2 Score of Translation Candidates

This section gives the definition of the score of a translation candidate in compositional translation.

First, let y_S be a technical term whose translation is to be estimated. We assume that y_S is de-

[3]Tonoike et. al (2006) reported that those two bilingual constituent lexicons compiled from the translation pairs of Eijiro improved the coverage of compositional translation from 49% up to 69%.

Table 1: Numbers of Entries and Translation Pairs in Lexicons

lexicon	# of entries		# of translation pairs
	English	Japanese	
Eijiro	1,631,099	1,847,945	2,244,117
bilingual constituents lexicon (prefix) B_P	47,554	41,810	129,420
bilingual constituents lexicon (suffix) B_S	24,696	23,025	82,087
phrase translation table	33,845,218	33,130,728	76,118,632

composed into their constituents as below:

$$y_S = s_1, s_2, \cdots, s_n \quad (1)$$

where each s_i is a single word or a sequence of words. For y_S, we denote a generated translation candidate as y_T:

$$y_T = t_1, t_2, \cdots, t_n \quad (2)$$

where each t_i is a translation of s_i, and is also a single word or a sequence of words independently of s_i. Then the translation pair $\langle y_S, y_T \rangle$ is represented as follows[4].

$$\langle y_S, y_T \rangle = \langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle, \cdots, \langle s_n, t_n \rangle \quad (3)$$

The score of a generated translation candidate y_T is defined as the product of a bilingual lexicon score and a corpus score as follows.

$$\prod_{i=1}^{n} q(\langle s_i, t_i \rangle) \cdot Q_{corpus}(y_T) \quad (4)$$

The bilingual lexicon score $\prod_{i=1}^{n} q(\langle s_i, t_i \rangle)$ is represented as the product of the score $q(\langle s_i, t_i \rangle)$ of a constituent translation pair $\langle s_i, t_i \rangle$, while the corpus score is denoted as $Q_{corpus}(y_T)$. Here, the bilingual lexicon score measures the appropriateness of the translation of each constituent pair $\langle s_i, t_i \rangle$ referring to bilingual lexicons provided as a resource for term translation, while the corpus score measures the appropriateness of the translation candidate y_T based on the occurrence of y_T in a given target language corpus.

More specifically, when the technical term y_S of the source language is decomposed into a sequence of constituents, the variation of the constituent sequence could be more than one. Then, this situation could lead to the case where a translation candidate y_T can be generated from more than one variations of the constituent sequence $s_1, s_2, \cdots, s_n$ of y_S. Considering such a situation, the overall score $Q(y_S, y_T)$ of the translation pair $\langle y_S, y_T \rangle$ is denoted as the sum of the score for each variation of the constituent sequence $s_1, s_2, \cdots, s_n$ of y_S.

$$Q(y_S, y_T) = \sum_{y_S = s_1, s_2, \ldots, s_n} \prod_{i=1}^{n} q(\langle s_i, t_i \rangle) \cdot Q_{corpus}(y_T) \quad (5)$$

4.2.1 Bilingual Lexicon Score

The bilingual lexicon score $q(\langle s, t \rangle)$ of a constituent translation pair $\langle s, t \rangle$ is defined as the sum of the score q_{man} for the pairs included in Eijiro, B_P, or B_S, as well as the score q_{smt} for those included in the phrase translation table:

$$q(\langle s, t \rangle) = q_{man}(\langle s, t \rangle) + q_{smt}(\langle s, t \rangle)$$

$$q_{man}(\langle s, t \rangle) = \begin{cases} 1 & \text{(if } \langle s, t \rangle \text{ in Eijiro, or } B_P \text{, or } B_S\text{)} \\ 0 & \text{(otherwise)} \end{cases}$$

$$q_{smt}(\langle s, t \rangle) = \begin{cases} P(t|s) & \text{(if } \langle s, t \rangle \text{ in the phrase translation table and } P(t|s) \geq p_0\text{)} \\ 0 & \text{(otherwise)} \end{cases}$$

In this definition, When the pair $\langle s, t \rangle$ is in Eijiro, B_P, or B_S, the score $q_{man}(\langle s, t \rangle)$ is defined as 1, while it is defined as 0 otherwise[5]. When the pair $\langle s, t \rangle$ is in the phrase translation table, on the other hand, we introduce the lower bound p_0 of

[4] Those bilingual constituents lexicons we introduced in section 4.1 have both single word entries and compound word entries. Thus, each constituent translation pair $\langle s_i, t_i \rangle$ could be not only one word to one word, but also one word to multi words, or multi words to multi words.

[5] In Tonoike et. al (2006), the score $q_{man}(\langle s, t \rangle)$ is defined to be a function of the number of constituents in s and t when the pair $\langle s, t \rangle$ is included in Eijiro, while it is defined to be a function of the frequency of the pair $\langle s, t \rangle$ in Eijiro when the pair is included in B_P or B_S. However, in our preliminary tuning phase, this definition achieves almost the same performance than the one we present in this paper. Thus, we prefer a simpler definition of q_{man} in this paper.

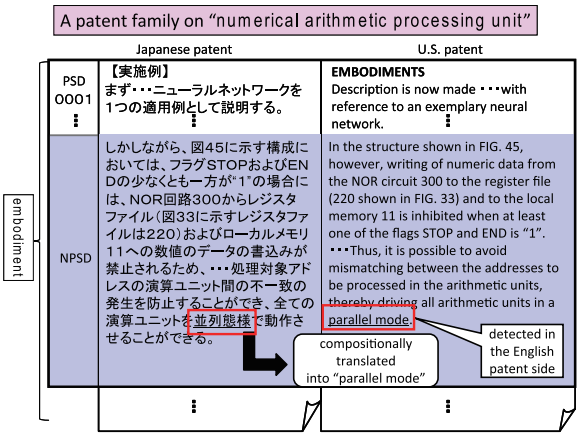

Figure 2: An Example of "Embodiment" Part with No Parallel Sentences Extracted

the translation probability. In this definition, when the translation probability $P(t|s)$ is more than or equal to the lower bound p_0 ($P(t|s) \geq p_0$), then the score $q_{smt}(\langle s, t\rangle)$ is defined as $P(t|s)$, while it is defined as 0 otherwise. In the evaluation in section 6, the parameter p_0 is optimized with a tuning data set other than the evaluation set.

4.2.2 Corpus Score

The corpus score measures whether the translation candidate y_T does appear in a given target language corpus:

$$Q_{corpus}(y_T) = \begin{cases} 1 & y_T \text{ occurs in the corpus of the target language} \\ 0 & y_T \text{ does not occur in the corpus of the target language} \end{cases} \quad (6)$$

5 Translation Estimation with the Part of No Parallel Sentences Extracted as a Comparable Corpus

This section describes how to estimate translation of technical terms using the part of patent families from which no parallel sentences are extracted, regarding it as a comparable corpus.

First, as we denote below, the Japanese part D_J of a Japanese-English patent family consists of the "Background of the Invention" part B_J, the "Detailed Description of the Preferred Embodiments" part M_J, and the rest N_J. B_J and M_J are then decomposed into the part PSD_J from which parallel sentences are extracted, and that $NPSD_J$ from which parallel sentences are NOT extracted. Similarly, the English part D_E of a Japanese-English patent family consists of the "Background of the Invention" part B_E, the "Detailed Description of the Preferred Embodiments" part M_E, and the rest N_E. B_E and M_E are then decomposed into the part PSD_E from which parallel sentences are extracted, and that $NPSD_E$ from which parallel sentences are NOT extracted. Figure 2 shows an example of "Embodiments" part, along with its PSD part and $NPSD$ part.

$$\begin{aligned} D_J &= \langle B_J, M_J, N_J\rangle \\ B_J \cup M_J &= \langle PSD_J, NPSD_J\rangle \\ D_E &= \langle B_E, M_E, N_E\rangle \\ B_E \cup M_E &= \langle PSD_E, NPSD_E\rangle \end{aligned}$$

In this paper, we extract a Japanese technical term t_J to translate into English from $NPSD_J$. This is mainly because we assume that Japanese technical terms appearing in PSD_J are expected to be translated into English by referring to the phrase translation table trained with parallel sentences extracted from PSD_J and PSD_E.

Then, considering the "Background" part B_E and the "Embodiment" part M_E in the English side as the target language corpus, we apply the compositional translation procedure of section 4 to t_J and collect the candidates of English translation which have the positive score $Q(t_J, t_E)$ into the set $TranCand\Big(t_J, B_E \cup M_E\Big)$:[6]

$$\begin{aligned} &TranCand\Big(t_J, B_E \cup M_E\Big) \\ &= \Big\{ t_E \in B_E \cup M_E \;\Big|\; t_J \text{ is compositionally translated into } t_E \text{ by the procedure of section 4 and (equation (5)) } Q(t_J,t_E) > 0 \Big\} \end{aligned}$$

Finally, out of the set $TranCand(t_J,\ B_E \cup M_E)$ of the translation candidates, we have t_E with the maximum score by the following function

[6] As the target language corpus, we also evaluate the part $NPSD_E$ (of B_E and M_E) from which parallel sentences are NOT extracted. However, in this case, we had a lower rate of correctly matching the translation candidates in the target language corpus. From this result, we prefer to have B_E and M_E as the target language corpus.

Table 2: Classification of the Japanese Compound Nouns in the 1,000 Japan-US Patent Families

(1) for the whole 61,133 Japanese noun phrases

Categories		Bilingual Constituent Lexicons		
		Eijiro ONLY	phrase translation table ONLY	Eijiro AND phrase translation table
(a)	Its English translation listed in Eijiro appears in the target language corpus	5,449 (8.9%)		
(b)	Included in the phrase translation table as one of the Japanese entries	32,516 (53.2%)		
(c)	Its compositional English translation (by the proposed method) appears in the target language corpus	4,004 (6.6%) (set E)	14,310 (23.4%) (set P, when maximizing $\|P\|$ ($p_0 = 0$))	14,575 (23.8%) (set EP, when maximizing $\|EP\|$ ($p_0 = 0$))
(d)	An English translation can be generated by Eijiro or compositional translation (by the proposed method), which does not appear in the target language corpus	397 (0.6%)	993 (1.6%)	1,041 (1.7%)
(e)	No English translation can be generated by Eijiro nor compositional translation (by the proposed method)	18,767 (30.7%)	7,865 (12.9%)	7,552 (12.4%)
total		61,133 (100%)		

(2) the set of whole 61,133 Japanese noun phrases $-$ the set (a) $-$ the set (b) $-$ the set E

Categories		Bilingual Constituent Lexicons	
		phrase translation table ONLY	Eijiro AND phrase translation table
(c)	Its compositional English translation (by the proposed method) appears in the target language corpus	10,375 (17.0%) (set $P - (E \cap P)$)	10,571 (17.3%) (set $EP - (E \cap EP)$)

$TranCand(t_J,\ B_E \cup M_E)$.

$$\mathrm{CompoTrans}_{\max}(t_J, B_E \cup M_E) = \underset{t_E \in TranCand(t_J,\ B_E \cup M_E)}{\arg\max} Q(t_J, t_E)$$

6 Evaluation

In order to evaluate the proposed method, we compare the following three cases:

(i) *Eijiro ONLY* $\cdots$ As bilingual constituents lexicons, Eijiro and its constituent lexicons are employed.

(ii) *Phrase translation table ONLY* $\cdots$ As bilingual constituents lexicons, the phrase translation table is employed.

(iii) *Eijiro AND phrase translation table* $\cdots$ As bilingual constituents lexicons, Eijiro and its constituent lexicons as well as the phrase translation table are employed.

First, we pick up 1,000 patent families, from which we extract 61,133 Japanese noun phrases. Then, we apply the compositional translation procedure of section 4 to those 61,133 Japanese noun phrases, and classify them into the following five categories (as shown in Table 2-(1)):

(a) The Japanese noun phrase is included in Eijiro as one of the Japanese entries, and its English translation appears in the target language corpus.

(b) The Japanese noun phrase is not in (a), and is included in the phrase translation table as one of the Japanese entries.

(c) The Japanese noun phrase is not in (a) nor (b), and by applying the proposed method of compositional translation to it, its English translation appears in the target language corpus.

(d) The Japanese noun phrase is not in (a), (b),

Table 3: Result of Evaluating Compositional Translation and Estimated Numbers of Bilingual Technical Term Translation Pairs to be acquired by the Proposed Method (per 1,000 Patent Families)

(1) for each case of bilingual constituent lexicons in compositional translation

	Bilingual Constituent Lexicons		
	Eijiro ONLY	phrase translation table ONLY	Eijiro AND phrase translation table
Evaluation Sets	$E' \subset E$, $\|E'\| = 93$	$P' \subset P - (E \cap P)$, $\|P'\| = 224$	$EP' \subset EP - (E \cap EP)$, $\|EP'\| = 230$
recall (%) precision (%) F-measure (%)	97.8 97.8 97.8	30.1 / 88.3 / 44.9 ($p_0 = 0.07$, when maximizing precision with recall $> 20\%$)	32.6 / 93.8 / 48.4 ($p_0 = 0.15$, when maximizing precision with recall $> 30\%$)
estimated numbers of term translation pairs	1,957 ($= 4,004 \times 0.5 \times 0.978$) (for the set E, $\|E\| = 4,004$)	1,561 ($= 10,375 \times 0.5 \times 0.301$) (for the set $P - (E \cap P)$, $\|P - (E \cap P)\| = 10,375$)	1,723 ($= 10,571 \times 0.5 \times 0.326$) (for the set $EP - (E \cap EP)$, $\|EP - (E \cap EP)\| = 10,571$)

(2) for the whole 61,133 Japanese noun phrases

	translation estimation for the set E with Eijiro ONLY + translation estimation for the set $P - (E \cap P)$ with phrase translation table ONLY	translation estimation for the set E with Eijiro ONLY + translation estimation for the set $EP - (E \cap EP)$ with Eijiro AND phrase translation table
estimated numbers of term translation pairs	3,518 (= 1,957+1,561)	3,680 (= 1,957+1,723)

nor (c), and from it, an English translation can be generated by Eijiro or by the proposed method of compositional translation, while the English translation does not appear in the target language corpus.

(e) The Japanese noun phrase is not in (a), (b), (c), nor (d), and from it, no English translation can be generated by Eijiro nor by the proposed method of compositional translation, simply because one or more constituents of the Japanese noun phrase can not be found in any constituent lexicons.

As in Table 2-(1), the number of the Japanese noun phrases of category (c) is 4,004 when *Eijiro ONLY* (denoted as the "set E"). The number is 14,310 when *phrase translation table ONLY* and the lower bound p_0 of the translation probability is equal to 0 (denoted as the "set P"), which becomes about 3.5 times larger. Furthermore, the number is 14,575 when *Eijiro AND phrase translation table* and the lower bound p_0 of the translation probability is equal to 0 (denoted as the "set EP"), which then becomes about 3.6 times larger compared with the set E.

Next, Table 3 shows the results of measuring recall / precision / F-measure of the proposed method, where we compare the three cases of bilingual constituent lexicons. First, we construct evaluation sets E', P', and EP' from the sets E, $P - (E \cap P)$, and $EP - (E \cap EP) = EP - E$, respectively[7]. Since we can mostly correctly estimate translation of the Japanese compound nouns within the set E when *Eijiro ONLY*, we exclude those members of E from the evaluation sets P' and EP'. Second, with tuning data sets other than those evaluation sets P' and EP', we optimize the

[7]We examined the sets E, $P - (E \cap P)$, and $EP - (E \cap EP) = EP - E$ in advance, and found that only 50% of their members are Japanese technical terms, while the remaining 50% consist of general compound nouns other than technical terms, terms with errors in segmentation of morphemes, and those not translated in the English patent side in the patent family. Thus, we construct the evaluation sets E', P', and EP' only from the Japanese technical terms portion of E, $P - (E \cap P)$, and $EP - (E \cap EP)$, i.e., 50% of them.

lower bound p_0 of the translation probability individually for both P' and EP'. Requiring that the recall is to be around 20∼30%, while the precision is to be around 80∼90%, we have the lower bounds p_0 as 0.07 for P' and as 0.15 for EP'

As shown in Table 3-(1), for the evaluation set E', we achieve high recall / precision / F-measure (97.8%), and the estimated number of technical term translation pairs to be acquired is more than 1,900[8]. This result is very impressive compared with the relatively low recalls when incorporating the phrase translation table as a bilingual constituent lexicon (30.1% for the set P' and 32.6% for the set EP'). This is simply because we restrict translation pairs within the phrase translation table by introducing the lower bounds p_0 of the translation probability. Consequently, we achieve the precisions to be around 80∼90% and satisfy the requirement of the procedure of manual judgement on accepting / ignoring the candidates. The estimated number of technical term translation pairs to be acquired is more than 1,500 for the evaluation set P' and is more than 1,700 for EP'. In total, for the set EP, we can acquire more than 3,600 novel technical term translation pairs per 1,000 patent families. Note that, in this procedure, acceptance rate of the manual judgement is over 95%, which is reasonably high.

7 Conclusion

This paper proposed to generate bilingual lexicon for technical terms not only from the parallel patent sentences extracted from patent families, but also from the remaining parts of patent families. The proposed method employed the compositional translation estimation technique utilizing the remaining parts as a comparable corpus for validating translation candidates. As the bilingual constituent lexicons in compositional translation, we used an existing bilingual lexicon as well as the phrase translation table trained with the parallel patent sentences extracted from the patent families. Finally, we showed that about 3,600 technical term translation pairs can be acquired from 1,000 patent families. Future works include applying an SMT technique straightforwardly to the task of technical term translation and comparing its performance with the compositional translation technique presented in this paper. We believe that the proposed framework of validating translation candidates is also effective with an SMT technique.

[8]Here, we suppose that we manually judge whether the translation candidates provided by the proposed method is correct or not and accept the correct ones while ignore the incorrect ones. We also assume that we can automatically or manually select Japanese technical terms (50%) from the whole set of compound nouns.

References

Fujii, A., M. Utiyama, M. Yamamoto, and T. Utsuro. 2008. Toward the evaluation of machine translation using patent information. In *Proc. 8th AMTA*, pages 97–106.

Fung, P. and L. Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 17th COLING and 36th ACL*, pages 414–420.

Huang, F., Y. Zhang, and S. Vogel. 2005. Mining key phrase translations from Web corpora. In *Proc. HLT/EMNLP*, pages 483–490.

Knight, K. and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Koehn, P. et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pages 177–180.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pages 127–133.

Liang, Bing, Takehito Utsuro, and Mikio Yamamoto. 2011. Identifying bilingual synonymous technical terms from phrase tables and parallel patent sentences. *Procedia - Social and Behavioral Sciences*, 27:50–60.

Lu, B. and B. K. Tsou. 2009. Towards bilingual term extraction in comparable patents. In *Proc. 23rd PACLIC*, pages 755–762.

Matsumoto, Y. and T. Utsuro. 2000. Lexical knowledge acquisition. In Dale, R., H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 24, pages 563–610. Marcel Dekker Inc.

Morishita, Y., T. Utsuro, and M. Yamamoto. 2008. Integrating a phrase-based SMT model and a bilingual lexicon for human in semi-automatic acquisition of technical term translation lexicon. In *Proc. 8th AMTA*, pages 153–162.

Och, F. J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Tonoike, M., M. Kida, T. Takagi, Y. Sasaki, T. Utsuro, and S. Sato. 2006. A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the web. In *Proc. 2nd Intl. Workshop on Web as Corpus*, pages 11–18.

Utiyama, M. and H. Isahara. 2007. A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pages 475–482.

Yasuda, K. and E. Sumita. 2013. Building a bilingual dictionary from a Japanese-Chinese patent corpus. In *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *LNCS*, pages 276–284. Springer.

Analysis of Parallel Structures in Patent Sentences, Focusing on the Head Words

Shoichi Yokoyama
Graduate School of Science and Engineering (Informatics), Yamagata University
4-3-16 Jonan, Yonezawa, Yamagata 992-8510, Japan
yokoyama@yz.yamagata-u.ac.jp

Abstract

One of the characteristics of patent sentences is long, complicated modifications. A modification is identified by the presence of a head word in the modifier. We extracted head words with a high occurrence frequency from about 1 million patent sentences. Based on the results, we constructed a modifier correcting system using these head words. About 60% of the errors could be modified with our system.

1 Introduction

In patent sentences, the "problem to be solved" and "solution" parts have complicated and long modificational phrases.

In order to analyze this complicated modification structure, we have investigated parallel conjunctions and parallel particles (Yokoyama 2005, 2007, 2009, 2011).

Here, we first considered the center of a modified noun as the head. Based on this, we constructed a system which corrects errors of modification. About 60% of the errors could be modified with our system. This paper is mainly based on the work of (Yokoyama 2012) and (Sakamoto 2012).

2 Parallel Structures in Patent Sentences

The majority of the "problem to be solved" and/or "solution" sentences in patents are often very long and complicated. These sentences sometimes have parallel structures with long modificational phrases.

We investigated parallel conjunctions and parallel particles to clarify the structure of the modifiers, and constructed a system which corrects errors of modification.

Here, we focused on the head words. First, we will describe the characteristics of each part of speech.

(a) Parallel conjunctions

Parallel conjunctions in Japanese include "mata wa" (or), "mosikuwa" (or), "oyobi" (and), and "narabini" (and).

Legal language places these conjunctions in a hierarchical structure to reduce and remove the ambiguity of law (Tajima 2006). However, following our investigations, no hierarchical structures were found in patent sentences (Yokoyama 2011).

(b) Parallel particles

In Japanese, parallel particles include "to" (and), "ya" (and), and "ka" (or). Patent sentences are often written in the form "A to B to no…" (the…of A and B). Our system can effectively correct relatively simple structures utilizing this form, but not more complicated ones (Yokoyama 2007).

(c) Head words

Head words are defined as the central nouns in modificational phrases such as "bu" (section) in the Japanese phrase "A-bu to, B-bu to…" (section A and section B …), and "sou" (layer) in the phrase "C-sou oyobi D-sou de wa …" (in layer C and layer D).

Fig. 1 shows an example abstract of a Japanese patent translated into English. In Fig. 1, the underlined and bold "section" words create parallel phrases in Japanese, but the underlined "section" words do not. In the Figure, these phrases are clearly separated by semicolons, but in Japanese they are connected by the coordinating conjunction "to" (and).

Here, we demonstrate that these head words clarify parallel structures.

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

(57)Abstract:
SOLUTION: The power line communication apparatus includes: an initial connection **section** for transmitting/receiving the device information between the other devices connected to the power line; a communicable party information creating **section** for creating the communicable party information of own, device based on the device information transmitted and received by the initial connection section; a communication-available party information transmitting/receiving **section** for transmitting and receiving the communication-available party information of own device, created by the communicable party information creating section between the other devices; a registration information creating **section** for creating the registration information where the other devices capable of communicating with own device are mapped; and a registration information transmitting/receiving section for transmitting and receiving the registration information created by the registration information creating section among other devices.

Fig.1 Patent language example (J2010-021954)

3 Materials and Methods

3.1 Materials

We used a patent information database made by AAMT (Asia-Pacific Association for Machine Translation)/Japio (Japan Patent Information Organization) Special Interest Group on patent translation (AAMT/Japio 2004). This database includes all patent applications filed in 2004, which consists of 339,716 patents, and 1,013,582 sentences.

3.2 Extraction and Classification of Head Words

All sentences are input into the common-use modification analyzer Cabocha (Cabocha 2012), and the analyzed results are output.

Fig. 2 shows an example of analysis by Cabocha (some parts omitted). In Fig. 2, the part of speech (for example, "noun, general") is shown translated into English, and the English translation of each word is in parentheses. Numbers like "0 2D" show the modification. Here, it shows that the phrase "tahoubu ni" (in the other section) modifies "aze-keisei souti" (ridge-forming device).

Tahoubu ni tutimori-souti to aze-keisei souti to
(filling device and ridge-forming device in other part)
*0 2D 1/2 0.644114
tahou (other) noun, general,*,*,*,*,
bu (part) noun, suffix, general,*,*,*,
ni particle, case particle, general, *,*,*,
*1 2D 2/3 0.000000
tuti (mud) noun, general, *, *,*,*,
mori (filling) noun, proper noun, general, *,*,
souti (device) noun, suru verb, *,*,*,*,
to (and) particle, parallel particle, *,*,*,*,
*2-1D 2/3 0.000000
aze (ridge) noun, general, *,*,*,*,
keisei (forming) noun, suru verb, *,*,*,*,
souti (device) noun, suru verb, *,*,*,*,
to (and) particle, parallel particle, *,*,*,*,
EOS

Fig. 2 Example analyzed by Cabocha

In Fig. 2, underlined "bu" (part) is a noun and suffix, and functions as a head word. The two underlined "souti" (device) words occurring in the second and third phrases can also be considered head nouns.

3.3 Investigation of Head Words with High Occurrence Frequency

We investigated 1 million sentences. To identify parallel phrases (Yamamoto 1996, Iwamoto 1993), we used coordinate particles and parallel conjunctions such as "to", "ya", "ka", ",", "katu", "oyobi", "mata", "narabini" (these translate into English as "and"), "aruiwa", "mosikuwa" (or), and "dake de (wa) naku" (not only...but also...). We ignored numbers and words written in original text using as the number of devices.

Table 1 Examples of head nouns

Word (Jap.)	Eng.	POS	Occur. Freq.
syudan	means	n. gr.	23,523
souti	device	n. v.	19,906
koutei	process	n. gr.	12,305
houhou	method	n. gr.	10,683
zyouhou	information	n. gr.	7,229
ki	radical	n. gr.	6,579
de-ta	data	n. gr.	4,671
buzai	component	n. gr.	4,534
suteppu	step	n. gr.	3,967
iti	location	n. v.	3,674

Table 1 gives some common head words occurring frequently in the text we searched. The columns show the head words, their English translation, part of speech, and occurrence frequency. We used the top 100 words with an occurrence frequency higher than 412 for our system. In Table 1, "n. gr." means "noun, general", and "n. v." means "noun, suru verb".

3.4 Investigation of Occurrence Frequency in a Specific Field

International patents are categorized by technical content, that is, using IPC (International Patent Classification). They are classified within hierarchies such as section, subsection, class, subclass, main group, and subgroup.

Sections are divided into 8 fields: A (human necessities), B (performing operations; transporting), C (chemistry; metallurgy), D (textiles; paper), E (fixed constructions), F (mechanical engineering; lighting; heating; weapons; blasting), G (physics), and H (electricity) (WIPO 2013).

Table 2 shows the high frequency words in Section C (chemistry; metallurgy) (27,969 patents, 76,517 sentences). There, nouns such as "atom" and "acid" (which do not occur very frequently in search results from all fields) have relatively high frequency. In Table 2, "n. adv." means nouns that can be adverbs.

Table 2 Examples of head nouns in Section C

Word (Jap.)	Eng.	POS	Occur. Freq.
ki	radical	n. gr.	3,729
ika	less than	n. adv.	1,941
koutei	process	n. gr.	1,844
zyusi	resin	n. gr.	1,654
houhou	method	n. gr.	1,613
gensi	atom	n. gr.	633
san	acid	n. gr.	580

4 Modification Correction System and Evaluation

4.1 Modification Correction System

Use of head nouns makes possible to deal with complicated telescopic modificational structures. We constructed a system to modify erroneous modification.

0 28D		mata, (and,)
1 2D		sono (its)
2 3D		tame no (for)
3 28D		seigyo reikyaku souti wa (control cooling device)
4 5D		atuen tyokugo no (just after rolling)
5 6D		kouhan no (steel plate)
6 7D		men ondo bunpu wo (surface temperature distribution)
7 8D		sokutei suru (measure)
8 15D	(8 17D)	ondosokutei [souti] wo (temperature measuring device)
9 11D		reikyakusui hedda- to (cool water header)
10 11D		kore ni (this)
11 12D		setuzoku sareta (connected)
12 13D		ramina-zyou no (laminar-formed)
13 14D		reikyakusui wo (cool water)
14 15D		kyoukyu suru (supply)
15 16D		nozuru to wo (nozzle)
16 17D		hukumu (including)
17 23D	(17 27D)	reikyaku <[souti]> to (cooling device)
18 19D		syotei no (designated)
19 20D		keisan puroguramu ni (computer program)
20 23D		sitagatte (following)
21 22D		kouhan no (steel plate)
22 23D		men ondo bunpu wo (surface temperature distribution)
23 25D		kin'ituka suru you ni (in order to standardize)
24 26D		reikyaku suiryou wo (cooling water volume)
25 27D		seigyo suru (control)
26 27D		reikyaku suiryou no (cooling water volume)
27 28D		seigyo <souti> to wo (controlled device)
28-1D		sonaeru (have)

Fig. 3 Example of correction by the system

Fig. 3 shows the correction result of the output of the system. The Japanese sentence we are using is the combination of every word in Fig. 3, and it is too complicated to translate it into English. Here, translation is only shown word by word.

In Fig. 3, the numbers show the phrase number, and numbers such as 2D, 3D show the phrase number modified. The correction shows parentheses such as (8 17D) and (17 27D). Head words relative to a modifier are shown by [ ], and heads relative to a modificand are shown by < >.

4.2 Evaluation

500 sentences randomly selected from patents in 2004, including parallel structures, are analyzed.

Table 3 shows the results. In Table 3, "C>C" shows that the analysis of modification is correct in the original Cabocha system, and the analysis by our system is also correct. "C>E" shows that the original analysis is correct, but our analysis is wrong. Conversely, "E>C" shows that the analysis of the original system is wrong, but our analysis can modify the result. "E>E" shows that the modification does not work.

We were able to modify 97 (58.4%) of 166 (97 + 69) erroneous sentences.

Table 3 Results of correction

	C>C	C>E	E>C	E>E	Total
Sent.	318	16	97	69	500
%	63.6	3.2	19.4	13.8	100

5. Concluding Remarks

In this paper, we described our experiment to correct erroneous analysis of modification. However, the correction was not as successful as expected; one reason is that the number of head words was restricted to words with a high occurrence frequency.

Next we plan to increase the number of words. We are also planning to use a thesaurus, and focus on numbers and symbols just after head words.

Acknowledgements

We thank Japio and the committee members for supporting this research and supplying the patent database.

References

AAMT/Japio, 2004. AAMT (Asia-Pacific Association for Machine Translation)/Japio (Japan Patent Information Organization) Special Interest Group on Patent Translation Patent Database.

Cabocha, 2012. Nara Advanced Institute of Science and Technology http://code.google.com/p/cabocha/

Hideaki Iwamoto, Kaoru Nagano, Hidetoshi Nagai, Teigo Nakamura, and Hirosato Nomur. 2004a, 1993: An Analysis of Coordinate Structures and its Applicatio to A Controlled Linguistic Model for Law Sentences, *Special Interest Group of Natural Language Processing in Information Processing Society Japan* (in Japanese) NL98-3

Nobutake Tajima, 2006. Basic Knowledge about Law Words (Third Edition) (in Japanese), Gyosei, Tokyo.

Kazuma Sakamoto, 2012. Analysis of Parallel Structures of Patent Sentences Focusing on the Head Words, *Bachelor Thesis of Department of Informatics, Faculty of Engineering, Yamagata University* (in Japanese).

WIPO (World Intellectual Patent Organization), 2013. IPC (international Patent Classification) http://web2.wipo.int/ipcpub/#refresh=page

Hiroomi Yamamura, Akira Suganuma, Kazuo Ushijima, 1996. A Simple Method to Analyze Coordinate Structures in a Japanese Document, and its Application to a Writing Tool, *Proceedings of 52nd Annual Meetings of Information Processing Society Japan* (in Japanese) 2J-1, pp.3-277-278.

Shoichi Yokoyama and Yuya Kaneda, 2005. Classification of Modified Relationships in Japanese Patent Sentences, *Proceedings of Workshop on Patent Translation*, pp.16-20.

Shoichi Yokoyama and Shigehiro Kennendai, 2007. Error Correcting System for Analysis of Japanese Patent Sentences, *Proceedings of Second Workshop on Patent Translation*, pp.24-27.

Shoichi Yokoyama and Masumi Okuyama, 2009. Translation Disambiguation of Patent Sentence using Case Frames, *Proceedings of Third Workshop on Patent Translation.*

Shoichi Yokoyama and Yuichi Takano, 2011. Investigation for Translation Disambiguation of Verbs in Patent Sentences using Word Grouping, *Proceedings of Fourth Workshop on Patent Translation*, pp.60-63.

Shoichi Yokoyama, 2012. Analysis of Parallel Structures of Patent Sentences Focusing on the Head Words, *Japio Year Book 2012* (in Japanese) pp.250-253.

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 %.

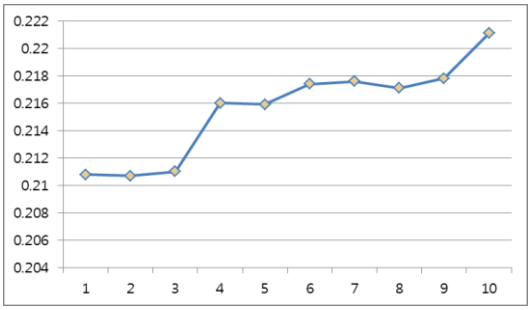

Figure 1: BLUE evaluation trend

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.

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.

Exploiting Multiple Resources for Japanese to English Patent Translation

Rahma Sellami
ANLP Research Group
Laboratoire MIRACL
University of Sfax, Tunisia
rahma.sellami@gmail.com

Fatiha Sadat
UQAM, 201 av. President Kennedy,
Montreal, QC, H3X 2Y3, Canada
sadat.fatiha@uqam.ca

Lamia Hadrich Belguith
ANLP Research Group
MIRACL Laboratory
University of Sfax, Tunisia
l.belguith@fsegs.rnu.tn

Abstract

This paper describes the development of a Japanese to English translation system using multiple resources and NTCIR-10 Patent translation collection. The MT system is based on different training data, the Wiktionary as a bilingual dictionary and Moses decoder. Due to the lack of parallel data on the patent domain, additional training data of the general domain was extracted from Wikipedia. Experiments using NTCIR-10 Patent translation data collection showed an improvement of the BLEU score when using a 5-grams language model and when adding the data extracted from Wikipedia but no improvement when adding the Wiktionary.

1 Introduction

Currently, there are five major patent offices in the world: Japan, Korea, China, Europe and the United States. These offices manage a huge amount of documents describing the patented inventions. There is a clear need to exchange the information related to such inventions, either for carrying out the legal tasks characteristics of the patent office, or for building systems that are able to search, access and translate patents content and make it available to the international community (Chechev et al., 2012). However, this task can be difficult to undertake through human translation when these documents are written in several languages; either due the outsize of the databases or the update frequency of the documents. For these reasons, the domain of patent machine translation is lately attracting the attention of researchers.

For any Statistical Machine Translation (SMT), the size of the parallel corpus used for training is a major factor in its performance. In order to improve the quality of Japanese-English machine translation of patent documents we propose to increase the size of the parallel patent corpus by adding parallel text extracted from Wikipedia and a dictionary of bilingual terminology extracted from the Wiktionary.

This paper describes the SMT system developed for the translation of Japanese to English patent documents, using NTCIR-10 data collection[1]. We try to improve the quality of translation of the Japanese to English patent documents by exploiting multilingual resources such as Wikipedia and Wiktionary.

This paper is organized as follows: In Section 2, we describe the patent documents and in Section 3 we describe some recent works for this domain. The approach used to develop the SMT system is described in Section 4. Section 5 and 6 give an overview of the proposed approach to improve the translation of the Japanese patent documents into English. Section 7 gives an overview of the experiment results. Section 8 concludes the present paper and discusses the possible future directions.

2 The Patent Domain

Patent documents are juridical documents, which are typically more structured than general documents, and they have their own special characteristics (Ma and Matsoukas, 2011). They also contain information about their publication, authorship and classification. Being an official document, the structure giving the terms of the patent is quite fixed. Every patent has a title, a description, an abstract with the most relevant information and series of claims.

A claim is a single sentence composed mainly of two parts: an introductory phrase and the body of the claim usually linked by a conjunction. It is in the body of the claim where there is the

[1] http://ntcir.nii.ac.jp/PatentMT-2/

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

specific legal description of the exact invention. Therefore, claims are written in a specific style and use a very specific vocabulary of the patent domain (Espana-Bonet et al, 2011).

Some of the patent document characteristics make MT easier, e.g., the presence of well-structured sentences and less ambiguity of word meanings. On the other hand, some characteristics become challenges for MT, e.g., long and complicated sentence structures, technical terminology and new terms that are originally defined by patent applicants. Compared to the newswire Japanese text data, the Japanese patent text has some specific characteristics such as legalese, technical terminology and long sentences. Also, patent text includes significantly more special strings that are not written in Japanese characters, such as English words, patent numbers, mathematical expressions and abbreviation names for materials.

3 Related works

The topic of patent translation is lately attracting the attention of researchers. Furthermore, a high number of patents is always registered and needs a form of translation into several languages. For these reasons, important efforts are being made in the last years to automate patent translation between different language pairs.

Researchers have explored various strategies to improve patent MT quality and have shown promising results, such as using the combined SMT with rule-based MT (Ehara, 2007; Wang, 2009; Jin, 2010).

In order to obtain a large coverage without losing quality in the translation, Espana-Bonet et al. (2011) proposed a combination between a grammar-based multilingual translation system and a specialized SMT system.

Enache et al. (2012) presented a hybrid translation system specifically designed to deal with patent translation. Indeed, the patent language follows a formal style adequate to be analysed with a grammar, but at the same time uses a rich and particular vocabulary adequate to be gathered statistically.

Ceausu et al. (2011) presented a number of methods for adapting SMT to the patent domain. They proposed some patent-specific preprocessing to resolve the problem of long sentences and references to elements in figure.

Ma et al. (2011) made changes to the SMT training procedure in order to better handle the special characteristics of patent data. He demonstrates that the re-training of the LM with patent text and the use of more features to the MT system improved the BLEU scores (Papineni et al., 2001) significantly.

Komachi et al. (2008) proposed a semi-supervised approach to acquire domain specific translation knowledge from the collection of Wikipedia. He has extracted a bilingual lexicon based on article tiles related by inter-language link and then applied the graph theoretic algorithm, regularized Laplacian, to find the most relevant translation pairs to the Patent domain.

4 MT System Basic Description

Our approach on statistical machine translation for Japanese and English pairs of languages is described as follows. First, a pre-processing step is performed on the source language, in order to convert raw texts into a format suitable for both training and decoding models.

Given the lack of word delimiters in written Japanese, word segmentation is generally considered a crucial first step in processing Japanese texts. For instance, the sequence ここでは、第 2 のコンタクトホール 61 内に (i.e., here, in the contacth all 61 of the second,) will have a proper segmentation as follows:|ここ |で|は|、 | 第 2 | の | コンタクト | ホール | 6 1 | 内| に | . We used Mecab tool (Kudo, 2002) to segment the Japanese texts. English pre-processing simply included down-casing and separating punctuation from words.

The common practice of extracting bilingual phrases from the parallel data usually consists of three steps: first, words in bilingual sentence pairs are aligned using automatic word alignment tools, such as GIZA++ (Och and Ney, 2003), in both directions; second, word alignment links are refined using heuristics, such as Grow-Diagonal-Final (GDF) method; third, bilingual phrases are extracted from the parallel data based on the refined word alignments with predefined constraints (Och et Ney, 2003). The 5-gram language models are implemented using the SRILM toolkit (Stolcke, 2002).

Decoding is the central phase in SMT, involving a search for the hypotheses t that have highest probabilities of being translations of the current source sentence s according to a model for P(t|s). Moses (Koehn et al., 2007), an open source toolkit for phrase-based SMT system, was used as a decoder.

These steps of building a translation system are considered as a common practice in the state-of-the-art of phrase-based SMT systems.

Once this is accomplished, a variant of Powell's algorithm is used to find weights that optimize BLEU score (Papineni et al., 2001) over these hypotheses, compared to reference translations. Weights of LM, phrase table and lexicalized reordering model scores were optimized on the development corpus thanks to the MERT algorithm (Och, 2003).

To build the patent MT system, we used the NTCIR-10 data for the Japanese-English subtask of patent MT evaluation. The data includes a parallel training corpus of approximately 3.2 millions of Japanese-English pairs of sentences, a development data set of 2,000 pairs of bilingual sentences in Japanese and English and a test data set of 2,300 pairs of patent sentences in Japanese. Furthermore, a set of 2,300 patent sentences in English is released at the end of the evaluations, to be considered as a reference set of the Japanese test sentences.

5 Parallel Corpora Extraction from Wikipedia

In most previous works on extraction of parallel sentences or phrases from comparable corpora, some coarse document-level similarity is used to determine which document pairs contain parallel data. For identifying similar web pages, Resnik and Smith (2003) compare the HTML structure. Munteanu and Marcu (2005) use publication date and vector-based similarity (after projecting words through a bilingual dictionary) to identify similar news articles.

Wikipedia is an online collaborative encyclopaedia available for a wide variety of languages. There are 24 language editions with at least 100,000 articles. Currently (May 2013), the English Wikipedia is the largest one with over then 4 millions articles. Whereas, Japanese Wikipedia contains approximately 862,000 articles[2].

Wikipedia contains annotated article alignments. Indeed, articles on the same topic in different languages are connected via "Inter-language" links, which are created by the articles' authors; we assume that the authors correctly positioned these links. This is an extremely valuable resource when extracting parallel data, as the document alignment is already provided. (Sellami et al., 2012) uses "Inter-language" for bilingual lexicon extraction from Wikipedia.

Wikipedia's markup contains other useful indicators for parallel sentence extraction. The several hyperlinks found in articles have previously been used as a valuable source of information. (Adafre et DeRijke, 2006) use matching hyperlinks to identify similar sentences. Two links match if the articles they refer to are connected by an "Inter-language" link.

Also, files, images and videos in Wikipedia are often stored in a central source across different languages; this allows the identification of captions, which are most of the time parallel (Smith et al., 2010). Figure 1 shows an example of captions in English and Japanese languages for an image extracted from Wikipedia. According to this example, the English phrase "The Sphinx against the Pyramid of Khafre" and the Japanese phrase "ギザの大スフィンクスとカフラー王のピラミッド" are considered as parallel.

We downloaded both English and Japanese XML Wikipedia dump and used the XML markup to extract pairs of titles connected by an "inter-language" link and pairs of captions that refer to the same file. Thus, parallel corpora extracted from Wikipedia contained 451,255 parallel phrases, with 422,425 phrases as pairs of titles and 28,830 phrases as pairs of captions.

English tokenization simply consists of separating punctuation from words. To segment the Japanese texts we used Mecab tool (Kudo, 2002). These extracted parallel corpora from Wikipedia were used with the initial NTCIR-10 Patent MT parallel corpora. Thus, a training of two language models and two translation models was completed using these extracted parallel corpora from Wikipedia and the initial NTCIR-10 Patent MT parallel corpora. The resulting system is called Patent+Wikipedia.

6 Hybrid MT System

Wiktionary[3] is a free-content, multilingual, web-based and freely available dictionary. It is considered as a lexical companion for Wikipedia. The size of the Wiktionary consists of approximately 16.5 million entries in 170 language editions[4].

[2] http://stats.wikimedia.org/EN/Sitemap.htm

[3] http://en.wiktionary.org/wiki/Wiktionary:Main_Page

[4] http://meta.wikimedia.org/wiki/Wiktionary

Figure 1. Captions for an image in English and Japanese languages

Each Wiktionary edition contains additional entries for foreign language terms. Therefore, each language edition contains a multilingual dictionary with a substantial amount of entries in different languages.

The English edition, the largest one, covers 3,691,575 entries on April 2013 while; the Japanese Wiktionary contains 110,463 entries on April 2013.

Entries in Wiktionary are accompanied with a wide range of lexical and semantic information such as part of speech, word sense, gloss, etymology, pronunciation, declension, examples, sample quotations, translations, collocations, derived terms, and usage notes.

In the current research, we have used the English Wiktionary since it contains more entries than the Japanese one; thus, we extracted all English terms having translations in Japanese. In total, we have extracted 1,528,475 pairs of English-Japanese terms. This process was based on the XML version of the English Wiktionary available online [5] and created by Sajous et al. (2010). Each entry in the XML Wiktionary contains an English term and its translations, gloss, POS, etc.

In our experiments, we have considered the alternative translation of a term to be likely equal. We can envisage attributing a score for each alternative using a disambiguation technique based on a statistical probability, which will consider the context of a term in the training corpus and the semantic information proposed by the Wiktionary.

7 Experiments and Results

We used the described tools in Section 4 in order to develop a basic SMT system for Japanese to English translation in the Patent domain.

We re-implemented our system described in Sadat et al. (2013); we implemented a 5-gram language model instead of a 3-gram language model. Our results were improved compared to the previous BLEU and NIST scores (Doddington, 2002), of 21.8 and 7.07, respectively (Sadat et al., 2013). Table 1 shows the formal run evaluation results for the Japanese-to-English translation, in terms of BLEU and NIST scores and the rates of Out-Of-Vocabulary words (OOV). By comparing the output of the three systems we found that the "Patent+Wikipedia" and "Patent+Wikipedia+Wiktionary" systems produced less unknown words than the basic "Patent" system. This is due to the large vocabulary introduced when adding Wikipedia and Wiktionary corpora to the parallel Patent corpora. Furthermore, our results show that the combination of the patent parallel corpora and the parallel data extracted from Wikipedia improved the BLEU score and decreased the OOV rate. Whereas, when we added the data extracted from Wiktionary to the patent parallel corpora and the parallel data extracted from Wikipedia, the BLEU score was decreased.

Wikipedia and Wiktionary are general domain corpora and do not contain the specific terminology and the legal vocabulary used in Patent documents. A domain adaptation method applied on the data of Wikipedia and the Wiktionary and the patent domain should improve the translation accuracy. Ceaușfu et al. (2011) compare different domain adaptation methods to different subject matters in patent translation and observe small gains over the baseline.

Some examples on a sentence of the test file are shown in Table 2. One can compare with the reference and realise the difference in terms of adequacy and fluency. These examples demonstrate that our translation is not very fluent but comprehensible and even we can consider it as very close to the reference.

[5] http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html

	BLEU	NIST	OOV
Patent	24.1	6.42	0.56
Patent+Wikipedia	24.67	6.557	0.39
Patent+Wikipedia+Wiktionary	24.5	6.551	0.39

Table 1. Results on the Japanese-to-English Patent MT Tasks

Input: 図 3 のフローチャートでステップ S6 に到達すると、 図 5 の「行程判別の可否判定」が起動される。
Patent: The flowchart of FIG. 3 reach the step S6. As shown in FIG. 5,0 stroke of the suitability determination is initiated.
Patent+Wikipedia: FIG. 3 is a flowchart of the step S6. As shown in FIG. 5, suitability determination whether the stroke is started.
Patent+Wikipedia+Wiktionary: FIG. 3 is a flow chart showing a step 6. As shown in FIG. 5, stroke determining suitability determination is initiated.
Reference: If step S6 is reached in the flowchart of FIG. 3. "Stroke determination propriety determination" as shown in FIG. 5 is launched.

Table 2. Examples of translations from Japanese to English with the references

8 Conclusion

In this paper, we have reported the results of our approach of Japanese-English MT using NTCIR-10 data collection for Patent MT. We have extracted parallel data from the Wiktionary and Wikipedia and have introduced a hybrid MT system using these multiple training data.

Evaluations using the basic Japanese-to-English statistical translation system could generate adequate and quite fluent translated sentences. The MT system using a 5 grams language model showed better results in terms of BLEU score compared to the MT system using a 3-grams language model. Also, introducing parallel data from Wikipedia improved BLEU score and decrease the OOV rate. However, introducing the Wiktionary did not show any improvement of the translation quality nor on the BLEU score. These evaluations were conducted without domain adaptation. In the future, we plan to pursue our research on domain adaptation for SMT using Wikipedia training data and possibly a combined statistical and rule-based MT system. Furthermore, we aim to participate in the future evaluations on Patent MT for Japanese-English language pairs.

References

Adafre S. F and De Rijke M. 2006. Finding similar sentences across multiple languages in wikipedia. *In Proceedings of EACL*, pages 62–69.

Ceausu, A., J. Tinsley, A. Way, J. Zhang, and P. Sheridan.2011. Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project. *In Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 2011).*

Chechev Milen, González Bermúdez Meritxell , MàrquezVillodre Lluísand España Bonet Cristina. 2012. The patents retrieval prototype in the MOLTO project. *In proceedings of the 21st International conference companion on World Wide Web". Lyon: ACM Press. Association for Computing Machinery*, p. 231-234.

Doddington G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

Enache Ramona, Espana-Bonet Cristina, RantaAarne and Marquez Lluıs. 2012. A Hybrid System for Patent Translation. *In Proceedings of the 16th EAMT Conference*, 28-30 May 2012, Trento, Italy.

Espana-Bonet C., Enache R., Slaski A., RantaA.,Marquez L., and Gonzalez M.. 2011. Patent Translation within the MOLTO project. *In Proceedings of the 4th Workshop on Patent Translation, MT Summit XIII*, pages 70{78, Xiamen, China.

Jin Yaohong. 2010. A hybrid-strategy method combing semantic analysis with rule-based MT for patent machine translation. *In Proceedings of 2010 International Conference on NLP-KE*.

Mamoru Komachi, Masaaki Nagata and Yuji Matsumoto. 2008. NAIST-NTT System Description for Patent Translation Task at NTCIR-7, *NTCIR-7*.

Koehn P., Shen W., Federico M., Bertoldi N., Callison-Burch C., Cowan B., Dyer C., Hoang H., Bojar O., Zens R., Constantin A., Herbst E., Moran C., and Birch A.. 2007. Moses: Open source toolkit for statistical machine translation. *In Proceedings of the ACL 2007 Interactive Presentation Sessions, Prague.*

Ma Jeff and Spyros Matsoukas. 2011. Building a Statistical Machine Translation System for Translating Patent Documents. *In Proceedings of the 4th Workshop on Patent Translation*, MT Summit XIII, Xiamen, China, pages 79-85.

Munteanu D. S. and Marcu D.. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Lin*guistics,31 (4):477–504.

Och Franz Josef. 2003. Minimum error rate training in statistical machine translation. *In 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

Och Franz Josef and Ney Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational linguistics*29 (1), 19-51.

Papineni K., Roukos S., Ward T., and Zhu W.. 2001. *Bleu: a Method for Automatic Evaluation of Machine Translation*. Technical Report RC22176 (W0109-022), IBM Research Division, Yorktown Heights, NY.

Resnik P. and Smith N. A. 2003. The web as a parallel corpus. *ComputationalLinguistics*, 29(3):349–380.

Sadat Fatiha and Fu Zhe. 2013. UQAM's System Description for the NTCIR-10 Japanese and English Patent MT Evaluation Tasks. *NTCIR-10*, Japon.

Sajous F., Navarro E. and Gaume B. 2011. Enrichissement de lexiques sémantiques approvisionnés par les foules: le système WISIGOTH appliqué à Wiktionary. *TAL*, 52(1), pp 11-35.

Sellami Rahma, Sadat Fatiha and Hadrich Belguith Lamia. 2012. Exploiting Wikipedia as a Knowledge Base for the Extraction of Linguistic resources: Application on Arabic-French Comparable Corpora and Bilingual Lexicons.*In Proceedings of the CAASL4 Workshop at AMTA 2012 (Fourth Workshop on Computational Approaches to Arabic Script-based Languages)*, San Diego, CA.

Smith J. R., Quirk, C. and Toutanova, K. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. *In Proceedings of NAACL-HLT 2010*, pp. 403-411.

Stolcke A.. 2002. Srilm-An Extensible Language Modeling Toolkit.*In Proceedings Of the International Conference on Spoken Language Processing.*

Kudo Taku Y. M.. 2002. Japanese dependency analysis using cascaded chunking. *In Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69.

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*, 11 September 2007, Copenhagen, Denmark, pages 13–18.

Wang Dan. 2009. Chinese to English automatic patent machine translation at SIPO. *World Patent Information*, Volume 31, Issue 2, pp. 137-139.