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.