Machine Translation Summit XV
http://www.amtaweb.org/mt-summit-xv

PSLT 2015:
The 6th Workshop on Patent and Scientific Literature Translation

Organizers:
Hiroyuki Kaji (Shizuoka University, Japan)
Katsuhito Sudoh (NTT, Japan)
Proceedings of

The Sixth Workshop on Patent and Scientific Literature Translation (PSLT6)

Hiroyuki Kaji, Katsuhito Sudoh & Takashi Tsunakawa, Eds.
Introduction

The Workshop on Patent and Scientific Literature Translation (PSLT) succeeds the series of Workshops on Patent Translation that have been held biennially as parts of Machine Translation Summits X to XIV. The previous workshops have focused on the translation of patent information, which is one of the major application areas of machine translation. Scientific articles are also important and represent a particular target for machine translation because they are similar to patent documents in that they use many technical terms and require relatively strict writing styles. Therefore, we have decided to expand the scope of the workshop to include the translation of scientific literature such as patent documents, scientific articles, and technical reports.

This year’s workshop features five invited talks from various standpoints: two by users (Bruno Pouliquen, World Intellectual Property Organization, and Kei Kato, Japan Patent Office), one by a developer (John Tinsley, Iconic Translation Machines Ltd.), and two by researchers (Stefan Riezler, Heidelberg University, and Toshiaki Nakazawa, Japan Science and Technology Agency). They cover various topics: both patent translation and scientific paper translation, both European and Asian language translation, and both practical use of MT systems and new challenges for advanced MT systems. Thus, the invited talks as a whole provide a comprehensive overview of state-of-the-art patent and scientific literature translation. The workshop also accepted six contributed papers that deal with interesting topics including the domain adaptation of statistical machine translation, statistical post-editing, bilingual technical term acquisition from comparable corpora, and cross-lingual patent wikification. We have organized these five invited talks and six contributed papers into four sessions: MT in Patent Organizations, Effective Use of Patent MT, Challenges for Advanced Patent MT, and Beyond Patent Translation. We hope that the workshop will contribute to advances in machine translation and its application to patent and scientific literature.

We express our sincere appreciation to the invited speakers and the authors of the contributed papers as well as the Program Committee members and the MT Summit XV Workshop Chair.

Co-chairs of the Program Committee of the 6th Workshop on Patent and Scientific Literature Translation

Hiroyuki Kaji
Katsuhito Sudoh
Program Committee Co-Chairs:
  Hiroyuki Kaji (Shizuoka University, Japan)
  Katsuhito Sudoh (NTT, Japan)

Program Committee Members:
  Key-Sun Choi (KAIST, Korea)
  Hiroshi Echizen-ya (Hokkai-Gakuen University, Japan)
  Terumasa Ehara (Ehara NLP Research Laboratory, Japan)
  Isao Goto (NHK, Japan)
  Kinji Hanawa (Japan Patent Information Organization, Japan)
  Takayuki Hayakawa (Japan Patent Information Organization, Japan)
  Munpyo Hong (Sungkyunkwan University, Korea)
  Eduard Hovy (Carnegie Mellon University, USA)
  Kenji Inamura (National Institute of Information and Communications Technology, Japan)
  Hideki Isozaki (Okayama Prefectural University, Japan)
  Hiroaki Kawai (Japan Patent Information Organization, Japan)
  Philipp Koehn (Johns Hopkins University, USA)
  Akira Kumano (Toshiba Solutions Corporation, Japan)
  Sadao Kurohashi (Kyoto University, Japan)
  Jong-Hyeok Lee (Pohang University of Science and Technology, Korea)
  Bente Maegaard (University of Copenhagen, Denmark)
  Toshimichi Moriya (Japan Patent Information Organization, Japan)
  Toshiaki Nakazawa (Japan Science and Technology Agency, Japan)
  Takashi Ninomiya (Ehime University, Japan)
  Tadaaki Oshio (Japan Patent Information Organization, Japan)
  Svetlana Sheremetyeva (Lanaconsult, Denmark)
  Sayori Shimohata (Oki Electric Industry Co., Ltd., Japan)
  Jun-ichi Tsujii (Artificial Intelligence Research Center, AIST, Japan)
  Takashi Tsunakawa (Shizuoka University, Japan)
  Takehito Utsuro (University of Tsukuba, Japan)
  Andy Way (Dublin City University, Ireland)
  Shoichi Yokoyama (Yamagata University, Japan)
  Jiajun Zhang (Chinese Academy of Sciences, China)

Publications Chair:
  Takashi Tsunakawa (Shizuoka University, Japan)
Contents

I. Extended abstracts of invited talks

1 Full-text Patent Translation at WIPO: Scalability, Quality and Usability
Bruno Pouliquen

9 Initiatives of the Japan Patent Office on Machine Translation
Kei Kato

19 Improving Translator Productivity with MT: A Patent Translation Case Study
John Tinsley

22 Response-Based Learning for Patent Translation
Stefan Riezler

30 Promoting Science and Technology Exchange using Machine Translation
Toshiaki Nakazawa

II. Contributed papers

33 Simplify Sentence Structure for Improving Human Post-editing Efficiency on
Xiaona Ren, Yongpeng Wei, Rile Hu

44 Enhancing Function Word Translation with Syntax-Based Statistical Post-Editing
John Richardson, Toshiaki Nakazawa, Sadao Kurohashi

Hongzheng Li, Kai Zhao, Renfen Hu, Yun Zhu, Yaohong Jin

68 Collecting Bilingual Technical Terms from Patent Families of Character-Segmented
Chinese Sentences and Morpheme-Segmented Japanese Sentences
Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, Mikio Yamamoto
Resampling Approach for Instance-based Domain Adaptation from Patent Domain to Newspaper Domain in Statistical Machine Translation

Keisuke Noguchi, Takashi Ninomiya

Towards Cross-lingual Patent Wikification

Takashi Tsunakawa, Hiroyuki Kaji
Full-text Patent translation at WIPO: scalability, quality and usability

Bruno Pouliquen
Bruno.Pouliquen@wipo.int
World Intellectual Property Organization, Global Databases Service
34, chemin des Colombettes, CH-1211 Geneva 20, Switzerland

Abstract
WIPO has access to a huge amount of patent application texts in different languages, therefore, we have built a machine translation tool (called WIPO translate) trained on big parallel data. We focus on offering quality machine translation to the general public, WIPO translate is fully integrated on our search engine PATENTSCOPE\(^1\). We have recently experimented aligning patent full-texts (description and claims) and our tool can now be trained on billions of words. Automatic evaluation metrics show an improvement over publicly available translation sites for the translation of patent texts (e.g. Google Translate, Microsoft translate). We have developed specific user interfaces, which are fully integrated, in our search engine PATENTSCOPE with reasonable translation speed. WIPO translate has now reached maturity in providing translation with competitive scalability, quality and usability.

1. Introduction
WIPO has 5 years’ experience in providing quality machine translation on its search engine PATENTSCOPE. Originally trained exclusively on patent titles and abstracts, we have now experimented using descriptions and claims (full text) to train our statistical machine translation tool (called WIPO translate), based on the open source toolkit Moses. Machine translation of patent texts is now integrated in PATENTSCOPE, despite the issues of scalability (translation models trained on billions of words), quality (our automatic evaluation shows an improvement over publicly available translation sites: e.g. Google Translate) and usability (it is fully integrated in our search engine PATENTSCOPE, with a translation speed of less than 2 seconds per sentence).

2. Background
The World Intellectual Property Organization (WIPO) provides access to about 50 million patent applications on its search engine PATENTSCOPE. It includes the international Patent Cooperation Treaty (PCT\(^2\)) applications, but also documents of participating national and regional patent offices.

All the PCT applications must be filed in one of the following languages: Arabic, German, English, Spanish, French, Russian, Japanese, Korean, Portuguese or Chinese, then the title and abstract are translated into English and/or French. The description and the claims remain available only in the original filling language.

\(^1\) [http://patentscope.wipo.int](http://patentscope.wipo.int)

\(^2\) [http://www.wipo.int/pct](http://www.wipo.int/pct)
The other applications (from regional or national patent offices) are usually published only in the original language (sometimes the title and the abstract are translated in one other language, rarely the claims and almost never the description).

This creates a huge need for users to read patent applications in a language they do not master. WIPO has investigated the use of machine translation to offer users a tool to help them understanding the content of any patent application.

WIPO previously investigated the use of machine translation to offer users the possibility to translate the title and abstract of patent applications (see section 2.1). However, this first solution was not suitable for the description and claims.

As a temporary solution, PATENTSCOPE offered the possibility to translate description and claims using public translation engines: Google translate, Microsoft/Bing translate and, recently, Baidu translate.

A recent study shows that, every day, about 5 million words are automatically translated using one of the machine translation tools available on PATENTSCOPE (Google translate/Microsoft translate). It clearly indicates the need for such tools in the intellectual property domain. This does not come as a real surprise as, of the total of about 32 million descriptions available on PATENTSCOPE, (as of September 2015), only 11.5 million (36%) are written in English (see Table 1)

<table>
<thead>
<tr>
<th>Language</th>
<th># applications (millions)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>11.5</td>
<td>36.16%</td>
</tr>
<tr>
<td>Japanese</td>
<td>8.2</td>
<td>25.79%</td>
</tr>
<tr>
<td>Chinese</td>
<td>4.5</td>
<td>14.15%</td>
</tr>
<tr>
<td>German</td>
<td>2.8</td>
<td>8.81%</td>
</tr>
<tr>
<td>Korean</td>
<td>2.7</td>
<td>8.49%</td>
</tr>
<tr>
<td>Spanish</td>
<td>1.1</td>
<td>3.46%</td>
</tr>
<tr>
<td>Russian</td>
<td>0.7</td>
<td>2.20%</td>
</tr>
<tr>
<td>French</td>
<td>0.3</td>
<td>0.94%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0.2</td>
<td>0.63%</td>
</tr>
<tr>
<td>total</td>
<td>32.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Number of patent applications in PATENTSCOPE having their description available in a specific language

2.1. Statistical machine translation

The statistical machine translation (SMT) approach “learns” its translation model using parallel sentences and then combines it with the target language model learned on monolingual texts. This fully automatic approach is suitable for the patent domain as we can automatically build such parallel corpora.

WIPO translate is based on the open source Moses3 (Koehn et al 2006). We have built a set of tools to pre-process the texts and to offer practical interfaces to Moses. WIPO tools include specific natural language processes: decompounder (German, Korean), pre-reordering (German, Japanese), prefix splitting (Arabic), Tokenizer (Chinese, Japanese)…

---

3 www.statmt.org/moses
The first version of WIPO Translate (Pouliquen et al. 2011) was trained on 180 million words (8.7 million English and French segments). The corpus has been released for free for research purposes (“COPPA corpus” Pouliquen & Mazenc, 20114).

2.2. Domain aware machine translation

We make use of the IPC classification5 to categorize any application into 32 domains (medicine, data, engineering, chemical …); this domain information is then encoded as a “factor” in Moses so that the phrase table can give “priority” to in-domain phrases. WIPO translate can then translate differently the same sentence in any of the 32 domains. Similarly, we decided to include the fact that a sentence belongs to the description or to the claim as a factor in our phrase table. Various experiments are still going on to better include this context information in the translation process.

3. Description of the tools

3.1. Text alignment

We need texts aligned at the sentence level in order to train Moses models. This is straightforward for titles when an application has a title in two languages. We need to apply better techniques for abstracts (e.g.: one sentence in English could be translated as two in French); the WIPO home-made sentence aligner has been developed for this purpose.

For descriptions and claims, the problem is different: the same invention can be submitted in different offices in different languages, but the description may be slightly different and the claims are often re-written according to the office and the protection needed by the applicant (see for example Täger 2011). Links between applications of the same invention are stored in a “priority list”, but the parallel corpus one can extract from this information is rather a “comparable” corpus than a real parallel corpus. WIPO adapted its sentence aligner tool to better filter descriptions and claims. WIPO aligner relies on bilingual dictionaries to automatically align sentences of noisy comparable corpus. It can discard non-aligned set of sentences or discard noisy texts (where the texts are not any more a translation of each other). Similarly, we heavily filter the claims and do not try to align them when the number of claims is different (e.g.: if a Chinese Patent Office application contains 11 claims then the corresponding US Patent Office application must also contain 11 claims) or when the intra-claims references are different (e.g.: if the third Chinese claim refers to the first claim then the third US claim must also refer to the first claim).

3.2. Technical infrastructure

The web application is distributed via a software load balancer to two servers. Each server calls a set of “WIPO translate servers” which in turn call a set of “Moses Engine servers”. This architecture allows for a robust and scalable set up where we can build Moses translation server farms (adding translation servers when new language pairs – or more engines for an existing pair - need to be added). It also avoids any single point of failure: if a web application or a translate server fails, the application can continue working.

Confidentiality is of high important for PATENTSCOPE users (a private company may not want its translation requests to be observed by another company, e.g. Google or Mi-

4 The COPPA corpus is available at: http://www.wipo.int/patentscope/en/data/#coppa. Note that the Version 2 (to be released in 2016) will include more applications and more languages.
5 http://www.wipo.int/classifications/ipc
crosoft) therefore all our servers work in https mode. This ensures users that no information (including IP address) is ever disclosed.

It should be noted that WIPO translate includes automatic monitoring and alerting tools to ensure a good customer service close to 24/7 (with minimal administration work).

### 3.3. Handling large data: scalability

The translation models, even when they are trained on big data, must be of a “reasonable” size. WIPO uses a set of filters, pruning processes and binarization tools in order to keep the model size to a minimum, without sacrificing much quality. See Table 2 for an example of size reduction: pruning and binarization manages to reduce a 342 Gb model to 15.2 Gb (4.4% of the original size).

<table>
<thead>
<tr>
<th>Phrase table</th>
<th>Reordering model</th>
<th>Language model (5 grams)</th>
<th>Total size</th>
</tr>
</thead>
<tbody>
<tr>
<td># rows (in million)</td>
<td>Size</td>
<td># rows (in million)</td>
<td>Size</td>
</tr>
<tr>
<td>Basic</td>
<td>806</td>
<td>100.0G</td>
<td>806</td>
</tr>
<tr>
<td>Pruned</td>
<td>551</td>
<td>69.0G</td>
<td>551</td>
</tr>
<tr>
<td>Binarized</td>
<td>6.4G</td>
<td>4.2G</td>
<td>4.6G</td>
</tr>
</tbody>
</table>

Table 2: size reduction (Chinese into English model)

WIPO translate must offer translation of any text in a “reasonable” time. We are trying to parallelize the decoding process and avoid time-consuming analyzers. We can now translate a full page within few seconds.

### 3.4. Quality

We use various automatic metrics (BLEU, METEOR, RIBES) to compare different versions of WIPO Translate, but also to compare WIPO translate to other engines output. We conducted an evaluation of the translation of patent application texts (1000 randomly selected sentences from newly published patent applications) the same text was submitted to WIPO translate and Google translate (see Table 2 for the results). Note that we use only title and abstracts, but, for Chinese, we conducted an evaluation on claims and descriptions.

<table>
<thead>
<tr>
<th>From language into English</th>
<th>WIPO translate</th>
<th>Google translate</th>
</tr>
</thead>
<tbody>
<tr>
<td>German title&amp;abstract</td>
<td>46.11</td>
<td>37.94</td>
</tr>
<tr>
<td>Spanish title&amp;abstract</td>
<td>36.00</td>
<td>33.07</td>
</tr>
<tr>
<td>French title&amp;abstract</td>
<td>46.97</td>
<td>41.72</td>
</tr>
<tr>
<td>Russian title&amp;abstract</td>
<td>28.88</td>
<td>17.76</td>
</tr>
<tr>
<td>Korean title&amp;abstract</td>
<td>22.09</td>
<td>19.85</td>
</tr>
<tr>
<td>Japanese title&amp;abstract</td>
<td>22.10</td>
<td>21.27</td>
</tr>
<tr>
<td>Chinese title&amp;abstract</td>
<td>26.37</td>
<td>21.80</td>
</tr>
<tr>
<td>Chinese claims</td>
<td>28.68</td>
<td>21.89</td>
</tr>
<tr>
<td>Chinese descriptions</td>
<td>38.03</td>
<td>32.40</td>
</tr>
</tbody>
</table>

Table 3: Comparison between WIPO translate and other engines (BLEU scores)
3.5. Usability

WIPO translate aims at offering users access to automatic translation of patent texts, therefore we try to give easy access to translation tools to PATENTSCOPE users.

3.5.1 Cross lingual Information Retrieval (CLIR)

The CLIR allows users to search a term or a phrase and its variants in English, French, German, Spanish, Portuguese, Japanese, Russian, Chinese, Korean, Italian, Swedish or Dutch by entering the term/s in one of those languages in the search box. The system will suggest variants and translate the term(s), therefore allowing users to search PATENTSCOPE for documents disclosed in a language that they do not master. This system has also been automatically trained using Moses on titles and abstracts.

![Figure 1: Example of CLIR query ("razor" automatically expanded to various languages)](image)

3.5.2 Translating short texts: web interface

Any user can access WIPO translate from a simple web interface, publically available at https://patentscope.wipo.int/translate, this allows for the translation of any given text.
3.5.3 Translating long documents: WIPO translate widget

Translation of patent application descriptions created a new challenge, as the texts are usually long (on average 6,400 words, but sometimes up to 150,000 words). Launching the translation of the full document would require many CPUs for each language pair.

It should also be noted that pre-translating all the texts is not really an option as our models are evolving over time and PATENTSCOPE has an enormous corpus of texts that would require weeks to be translated.

Therefore WIPO has developed its own “widget” (JQuery program running on the client side) which is used to translate only the sentences the user is currently reading on the screen. This technique allows for a better share of the server load among users of PATENTSCOPE. It has been decided, as a first step, to offer only translation of Chinese-English (both directions), but more languages will be added in the future.

---

6 The total number of characters of English text in PATENTSCOPE is about 1 trillion (to give an order of idea, English Wikipedia is 49 billion characters)
This widget can be smoothly included on any page, it is currently available on the description page, but also for the claims page, the bibliographic page and the result list.

### 3.6. Our software in other contexts

It is of note that the software has been successfully installed in other UN organizations. It has been trained on their own data and is now running with different models, different usage scenarios and different quality. In contrast to WIPO Translate on PATENTSCOPE (where the tool is used for assimilation: i.e. to offer users a “gist” translation) the tool in other contexts is used mainly for disclosure, that is to say as a “translation accelerator”, offering translators a first translation to refine.

WIPO translate software (called “TAPTA”) can handle the internal parallel documents of an organization and can then offer translators a quality machine translation reproducing their own internal jargon. It has been installed at the United Nations since 2011 (see details in Pouliquen et al., 2013) and recently at the International Maritime Organization (IMO) (see Pouliquen et al. 2015). The tool is used internally in WIPO PCT (for the translation of titles and abstracts), in the Madrid sector (to translate goods and services) and in the WIPO Language Division (to translate official documents). In addition, it has been installed at the Food and Agriculture Organization (FAO) and the International Telecommunication Union (ITU). Early prototypes have been installed at the World Trade Organization (WTO), the International Labour Organization (ILO) and the International Social Security Association (ISSA).

### 4. Conclusion and future work

WIPO translate has now been running daily for five years, and we have smoothly incorporated the “WIPO translate widget” and the translation of description and claims. Despite the challenges of scalability, quality and usability, WIPO translate has reached maturity and is now a reliable system.
The WIPO translate widget has been available online since the beginning of September 2015, we can already see a major increase in the number of words translated every day\(^7\), even if we currently offer only Chinese-English translation for descriptions and claims.

We hope to extend full-text document translation to new languages in the near future (Japanese, German etc.) provided that we get enough computer power. A preliminary evaluation shows that we can get big models using Japanese (estimation: 3 Billion words), German (~ 1 Billion words) or Russian (~ 200 Million words). We also plan to offer Arabic and Portuguese translation in the future (our first goal is to cover the 10 official languages of the PCT).

We are currently investigating the following topics:

- Translating through pivot language to offer any language combination (e.g. German-Japanese)
- Various techniques for a better domain adaptation (e.g. use different language models for descriptions/claims/IPC domains etc.).
- Use collected post editions to add quality estimation metric for each translated sentence
- Use transliteration techniques to “translate” applicant names across different scripts: Arabic, Latin, Cyrillic, Hangul (Korean alphabet), Chinese, Kanji+Hiragana+Katakana (Japanese) etc…

Acknowledgement

Many thanks to Marcin Junczys-Dowmunt for his indispensable contributions to this work, and to Christophe Mazenc who built the CLIR component and for managing this project.

References


\(^7\) In September 2015, WIPO translate receives an average of 200,000 words/day while, before the introduction of the translation of description and claims, the rate was 50,000 words/day. The translation of descriptions (English from/to Chinese) accounts for half of the requests.
Initiatives of the Japan Patent Office

on Machine Translation

Kei Kato
Patent Information Policy Planning Office, Policy Planning and Coordination Department, Japan Patent Office
3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan
kato-kei@jpo.go.jp

Abstract
Recently, each Intellectual Property (IP) Office has faced common big challenges: promoting the international work sharing in the examination process and developing an environment to access foreign patent documents written in languages other than its native language. As a means of resolving such challenges, the Japan Patent Office (JPO) has actively utilized machine translation. Currently, the JPO has widely provided the foreign general users with the Japanese-English machine translated information on its patent examination results through the “One Portal Dossier” allowing them to retrieve dossier information for applications filed with the IP5 Offices (Japan, the U.S., Europe, China, and the Republic of Korea). Also, the JPO launched a new system in January 2015, utilizing the Chinese-Japanese machine translation dictionary including more than 2 million words. This system enables users to search in the Japanese language more than 12 million machine translated patent and utility model documents of the Chinese and Korean languages.

1. Introduction
The number of patent applications filed worldwide greatly increased to 2.57 million in 2013, compared with 1.58 million in 2004. Especially, the number of patent applications filed in China has dramatically increased and grew to 32.1% of the total applications in 2013 from 8.3% in 2004. From a viewpoint of searching patent information, this situation shows that the needs for searching foreign documents including those of the Chinese languages have increased. That is, as the examiners judge the novelty of inventions, inventive step, etc. based on the result of prior art search in the patent examination, it is necessary to search prior art documents including these foreign patent documents precisely as well as efficiently in order to grant a stable right recognized in the world. Actually, as the rate at which patent documents of the Chinese languages are cited as prior art in the examination has gradually accelerated at each Intellectual Property (IP) Office, it has become a challenge common to all IP Offices to improve an access environment such as understanding and searching foreign patent documents written in languages other than its native language.

On the other hand, it is not easy to understand and to exhaustively search foreign patent documents around the world. Needless to say, while translation enables us to understand them in our native languages, translating quite a large number of documents only by human has some limitations in terms of costs and resources, as the number of patent applications filed worldwide has increased year by year. Under this situation, the expectation of im-
The improvement of the environment for searching foreign documents by utilizing machine translation is getting higher.

Furthermore, it is hoped that machine translation will be utilized in order to disseminate patent information to other IP Offices. Currently, while intellectual property related activities are getting globalized, applications for a patent of one invention are more often filed in several countries or regions of the world. In case applications for a patent of the same invention are filed in several countries or regions, examiners in each IP Office need to share examination information (dossier information) owned by each IP Office in order to eliminate inefficiencies of conducting duplicate examinations or searches. But if each IP Office provides its examination information in its native language, it is difficult for other IP Offices to utilize such examination information. In order to promote the international work sharing, it is important to disseminate not only information in its native language but also one translated into other languages such as English.

In view of this situation where attention is getting focused on utilizing machine translation in the area of patent information, the Japan Patent Office (JPO) has taken measures to actually utilize machine translation. In this paper, I will introduce a search system for the Chinese and Korean documents utilizing machine translation and a service to provide examination information which the JPO has provided.

2. Improvement of the Environment for Searching Patent Documents of the Chinese and Korean languages

2.1. The Chinese and Korean Gazette Translation and Search System (CKGS)

As shown in Figure 1, in 1998 the documents to be understood and searched in the Japanese language accounted for 55% of the total patent documents issued in the world. But recently as the number of patent documents of the Chinese language has rapidly increased, documents to be understood only in the Chinese and Korean languages have accounted for 65% of the total. In January 2015, the JPO launched a new service of the “Chinese and Korean Gazette Translation and Search System” (CKGS) which enables users to search in the Japanese language patent and utility model documents of the Chinese and Korean languages in order to enhance the convenience of the environment for searching them.

A rule-based translation method is adopted for this system so that important technical terms can be translated and searched properly. The system stores patent and utility model documents machine translated from Chinese and Korean to Japanese, so that Japanese key words can be used for the full-text search of these documents. As of the end of March 2015, the system can be used to search about 12 million Chinese and Korean patent and utility model documents and it is planned to increase the number of such searchable documents gradually in the future.

The quality of texts machine translated by the system plays an important role if the CKGS is to be used for practical applications. To this end, the JPO has been conducting projects since Fiscal 2012 to develop a specialized dictionary for patent terms which will be used for the system’s machine translation, as its initiatives to improve machine translation quality (JPO, 2013, 2014a, 2015). Specifically, corresponding Japanese and Chinese sentences are extracted from the patent family of a same invention applied to Japan and China in order to prepare Chinese-Japanese translation corpuses as well as to develop a specialized dictionary. Currently, about 153 million Chinese-Japanese translation corpuses and the specialized user dictionary containing more than 2.2 million words have been developed. The user dictionary is installed to the CKGS to improve the system’s translation quality.
Furthermore, measures are taken to enable the system to handle new technical terms as they appear. Specifically, based on information concerning unknown words which are detected as those not registered in the machine translation dictionary during the system’s machine translation and that concerning incorrect translations and other errors reported by the system’s users, efforts are made to update the dictionary, make additional registrations to the system’s translation memory, and tune the translation engine parameters. With these efforts, the dictionary and the translation engine will be improved so that new technical terms can be handled as they appear.

In addition, through the Internet, the CKGS has become available not only to examiners but also to general users. As described above, continued efforts are being made to improve the system’s machine translations so that practical application environment for uses to search Chinese and Korean documents can be provided, and the system’s users appreciate it very much.

Figure 1. Rapid increase in Chinese patent documents. Note: Patent (incl. utility model) documents issued worldwide are categorized by language and duplicated data are eliminated. Regarding patent documents for the same invention which have been filed and published at multiple Offices, those documents published in Japanese are counted as JP. In case of no Japanese publications, such documents are counted first as US (English), secondly as EP (English, French, German), thirdly as KR (Korean) and last as CN (Chinese), if each language is applicable.
2.2. Quality Evaluation of Machine Translations from Chinese to Japanese and Those from Korean to Japanese

For realizing machine translations to be widely accepted and used, the quality of such translations plays a very important role. As a prerequisite to deciding whether to introduce such machine translations, the quality of such translations must be evaluated properly. A “Survey on How to Evaluate Quality of Patent Document Machine Translations” was conducted in FY 2013 to investigate how the quality evaluation method of machine translated patent documents should be (JPO, 2014b). Based on the survey’s results, “Quality Evaluation Procedures for Patent Document Machine Translations” was developed in Fiscal 2014 as a guideline to evaluate the quality of machine translation results properly (JPO, 2014c).

In order to verify the results of the JPO’s efforts to improve the system’s machine translation quality in accordance with the “Quality Evaluation Procedures for Patent Document Machine Translations,” machine translations provided by the Chinese and Korean Gazette Translation and Search System were evaluated in the end of 2014 with regard to how accurately the system could translate technical terms (or the system’s translation accuracy of technical terms) and how well the system could communicate original contents (or the system’s original contents communication level).

With regard to the system’s translation accuracy of technical terms, the system’s machine translations of Chinese documents into Japanese were surveyed. A total of 196 technical terms were selected from various fields, and evaluated and classified into the following 4 grades by human evaluators.

- **A (Properly Translated Word):** Compared with one translated by a human, it is a word translated to a technically same or similar meaning and generally used.
- **B (Acceptably Translated Word):** It is not a translation word generally used as a technical word, but its meaning is almost correct.
With regard to the system’s original contents communication level, the system’s machine translations from Chinese and Korean to Japanese were surveyed. 100 machine translated sentences of each of the languages were selected from various fields, and evaluated and classified into the following 5 grades by human evaluators.

- 5: All contents of its important information are communicated correctly (100%).
- 4: Almost all contents of its important information are communicated correctly (80% or more).
- 3: No less than half of the contents of its important information are communicated correctly (50% or more).
- 2: Several contents of its important information are communicated correctly (20% or more).
- 1: Its translations cannot be understood, or almost no contents of its important information are communicated correctly (less than 20%).

In order to verify how effective it was to install a Chinese-Japanese dictionary developed by JPO to the system, the system’s machine translations from Chinese to Japanese were evaluated before and after that dictionary with approximately 1 million words was installed to it for both surveys on the system’s translation accuracy of technical terms and on the system’s original contents communication level. In addition, for the purpose of comparison, the translated sentence using statistic machine translation (SMT) with 100 million Chinese-Japanese translation corpuses were also evaluated in the same manner. Figure 3 shows the evaluation results of the system’s original contents communication level through machine translations from Chinese and Korean to Japanese. While the system’s average original contents communication level for Korean-Japanese translations is above 4 points, one for Chinese-Japanese translations is merely above 2 points. In addition, in all the technical fields, the system’s original contents communication levels for Korean-Japanese translations are higher than those for Chinese-Japanese translations. It was concluded that this was due to the fact that the grammatical characteristics of Korean were closer to those of Japanese than those of Chinese, and the results indicate that continued efforts must be made to improve the system’s translation accuracy of Chinese documents. Furthermore, the system’s translation accuracy of either Chinese or Korean documents varies from one field to another, and measures must also be taken for individual fields. With regard to this, by the end of FY 2015, the JPO plans to measure the system’s translation accuracy of Chinese documents in individual fields, and focus on developing the Chinese-Japanese dictionary for the technical fields whose translation accuracy is low.

Figures 4 and 5 show how accurately the system can translate Chinese documents’ technical terms and how well the system can communicate their original contents, before and after the Chinese-Japanese dictionary developed by JPO was installed, respectively. That dictionary being added to the system’s translation engine, the system’s translation accuracy of technical terms increased by about 4%, and the system’s original contents communication level by about 13%. Specifically, the translation accuracy of technical terms by the system’s machine translation based on a rule-based method is higher than one by the SMT.
On the other hand, the original contents communication level by the SMT is evaluated to be higher. Specialized technical terms are frequently used as key words to search patent documents, and the translation accuracy of such technical words plays an important role for

Figure 3. Original contents communication levels through machine translations from Chinese and Korean to Japanese in various fields. The figures show average evaluation points for respective fields averaging points of individual machine translations in the respective fields which are evaluated according to the 5 ranks of 1 to 5 points. “A” represents 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.

On the other hand, the original contents communication level by the SMT is evaluated to be higher. Specialized technical terms are frequently used as key words to search patent documents, and the translation accuracy of such technical words plays an important role for

Figure 4. Evaluation of translation accuracy of technical terms in Chinese-Japanese machine translations: The above figures represent the sums of evaluation A (Properly Translated Word) and B (Acceptably Translated Word) ratios averaged over the 196 words.

Figure 5. Evaluation of original contents communication Level: The above figures represent 5 level evaluations of 1 to 5 points on machine translations averaged over the 100 sentences.
such searches. However, original contents communication level is also important to understand documents’ contents, and thus efforts must be made to improve the system’s translation quality.

3. Providing Information on Examination in English

3.1. Advanced Industrial Property Network (AIPN)/One Portal Dossier (OPD)

The number of yearly patent applications in the world increased 1.6 times in 10 years from 2004 to 2013, the 80%, which showed 2.08 million applications in 2013, were filed with five intellectual property offices in Japan, U.S., Europe, China and the Republic of Korea (or the IP5 Offices). Out of these, not a few applications for single inventions were filed in multiple countries and regions.

In order to facilitate examination work sharing among examiners in respective IP offices, the JPO has been providing foreign IP offices with its information on patent applications and examinations (dossier information) translated from Japanese to English by machine translation since October 2004 through a network called the Advanced Industrial Property Network (AIPN). The AIPN enables individual IP offices’ examiners to cross check the individual IP offices’ information on patent applications and examinations (dossier information) mutually so that if a same invention is filed to multiple countries and regions, duplicate searches and examinations on the same invention can be excluded. The AIPN enables Japan’s dossier information to be understood in English, and individual IP offices’ examination results to be shared among the individual IP offices.

Furthermore, an IT service called the “One Portal Dossier” (OPD) was started in 2013

![Figure 6. Conceptual diagram of AIPN/OPD.](image-url)
for the IP5 offices’ examiners, which collects the IP5 offices’ information on examinations together, including the JPO’s information on examinations translated by the AIPN, and provides such information in an easy-to-see format. Currently, this OPD’s service is enhanced, and the JPO’s information on examinations is machine translated and provided to general users together with other IP5 offices’ dossier information. In 2014 alone, through this service, the JPO’s examiners browsed about 160,000 cases in the dossier information of the EPO, the USTPO, the SIPO and the KIPO, and the examiners of the EPO, the USTPO, the SIPO and the KIPO browsed about 210,000 cases in the dossier information of the JPO. Such dossier information helps to make examinations to be more efficient.

The JPO’s information on examinations, which is provided through the AIPN/OPD, contains texts machine translated by a translation engine based on a rule based method. In general, machine translations from Japanese to English tend to show low translation accuracy when compared with those from English to Japanese. Notwithstanding, the JPO is constantly enhancing its machine translation dictionary to improve its service’s translation accuracy. Specifically, the JPO collects untranslated words (unknown words) and registers their translations to its dictionary, as well as analyze feedbacks about mistranslations in the AIPN from foreign IP offices and incorporate their corrections into its dictionary. Thereby, the JPO is making efforts to enhance its dictionary.

### 3.2. Quality Evaluation of Japanese-English Machine Translations

In order to verify the translation accuracy of machine translations from Japanese to English which the JPO provides, in 2011 the JPO conducted a “Survey on Translation Accuracy Evaluation for Providing English Texts of Machine Translations of Information Related to Patent Examinations” (JPO, 2011). This survey evaluated 511 sentences extracted from patent gazettes and reasons for refusal in the fields of electricity, physics and chemistry. Two types of translation methods, that is, the JPO’s machine translation system (AIPN) and commercially available Japanese-English translation software for patent, were applied to the sentences to obtain their machine translations, and the 2 types of machine translation methods were evaluated automatically as well as by human evaluators. The human evaluators evaluated the individual machine translations from the view point of whether the machine translations correctly reflect their originals’ meanings, and assigned 0 to 4 points to each of them according to a 5 grade evaluation scheme. Three kinds of evaluation methods, that is, BLEU, NIS and IMPACT were used for the automatic evaluation.

Figure 7 shows the results of the human evaluators’ evaluation. While the AIPN’s average point is 2.5 points, the commercial product’s is 2.0 points. In addition, the results indicate that the AIPN’s points are higher than the commercial product’s in all the technical fields. Results obtained by the automatic evaluation methods (BLEU, NIS and IMPACT) show trends almost similar to one obtained by the human evaluator’s evaluation (Figures 8 to 10). Because of the results mentioned above, it can be concluded that the JPO’s efforts to improve the accuracy of its Japanese-English machine translations have made some progress. However, it is noticed that the translation accuracy varies from one technical field to another. In addition, different evaluation methods produced different results, and a method to evaluate translation accuracy must be reviewed and examined continually. The JPO is still continuing to make its efforts to improve the quality of its Japanese-English machine translations even after the 2011 survey, and also plans to conduct another survey on the quality evaluation of its Japanese-English machine translations including the analysis of evaluation methods in FY 2015 in order to verify the results of this effort.
4. Conclusion

Demands for machine translations of patent information are increasing year by year from the viewpoint of patent information searches and dissemination. While machine translation is a developing technology and is being improved constantly, its accuracy has not become satisfactory yet, and one of its important issues is how to improve its accuracy. To improve machine translation quality, a solution by translation logic may be effective to solve difficulties in analysis due to languages’ grammatical characteristics. The JPO is evaluating the quality of Chinese-Japanese and English-Japanese machine translations in Fiscal 2015, and henceforth it may be necessary in the future, for example, to consider whether to adopt a statistic machine translation engine instead of a rule-based machine translation method based on quality evaluation results and the fact that the translation accuracy of statistic machine
translation is improving in recent years. In addition, to improve translation quality, methods
and criteria to evaluate individual machine translation schemes’ qualities are also important
because they are preconditions to evaluating translation quality. With regard to this, it is a
common issue among the IP5 offices how to improve machine translation quality, and the
offices plan to have further discussions on making methods and criteria for machine
translation quality evaluations common among the IP5 offices.

References


Improving translator productivity with MT: a patent translation case study

John Tinsley  john@iconictranslation.com
Iconic Translation Machines Ltd. Invent DCU, Glasnevin, Dublin 9, Ireland

Abstract
When evaluating the suitability of MT for post-editing, there are a lot of variables to consider that could have an impact on its overall effectiveness, including: the languages in question, the documents being translated, the translator workflow, and not least the individual translators themselves. Add to that the various ways we can carry out evaluation - automatic measures, subjective assessments of fluency and adequacy, etc. - and we have got a lot of data on our hands. Despite all of these data points, we are ultimately trying to answer a simple question: is the MT useful and to what extent? In this article, we introduce Iconic Translation Machines and describe a case study on a large-scale post-editing evaluation involving more than 20 translators working on Chinese to English patent translation. We discuss how various evaluations were carried out – from initial MT engine development to translator productivity – and discuss the implications of these findings on the real-world application of MT.

1. Introduction
Iconic Translation Machines (Iconic) is a machine translation software and service provider that specialises in domain-adapted MT. We focus on developing MT engines for specialist content types that are particularly challenging for translation and require more than a pure data-driven approach. Our flagship solution, IPTranslator, was an MT service adapted specifically for patent and intellectual property-related content, and in addition to this, Iconic now develops domain-adapted engines in the areas of finance, life-sciences, and e-commerce.

2. Machine Translation with Subject Matter Expertise
Iconic’s approach to MT extends on the classic data-driven approach of statistical MT by incorporating aspects of syntax-based MT and rule-based MT. These approaches are combined with domain-specific processes that have been developed to address particular stylistic conventions of different content types, such as extremely long sentences with complex alpha-number sequences in patent data. Distinct processes, addressing the language, domain, and style, are combined together in Iconic’s Ensemble Architecture™ which selects the most effective combination of processes for a particular input type at runtime.

3. RWS: A Case Study
RWS Group is a world-leading language service provider that specialises in patent translation. Iconic worked with RWS to customise a domain-adapted MT engine for Chinese to English
translation. The goal of using MT at RWS was to improve the productivity of translators through post-editing and this was taken into account during the development and automatic evaluation of the engines.

Iconic used its existing baseline for Chinese—English and supplemented it with additional in-domain data suited to RWS content. This adaptation achieved significant improvements in BLEU and TER scores as shown in Figure 1.

The TER (translation edit rate) scores give a good indication that the MT output is of sufficient quality to facilitate faster post-editing. In the next section, we describe the evaluations we carried out in order to validate and quantify this in a practical, real-world scenario with professional translators.

3.1. Evaluation Setup: mitigating the variables

In order to quantify the increase in productivity of translators post-editing MT output as opposed to translating from scratch, we carried out an evaluation using the TAUS Dynamic Quality Framework (DQF) ¹. Using this tool, translators post-edit and translate (from scratch) alternating segments in a given test document and the amount of time spent on each segment is measured. The total time spent post-editing vs. translating from scratch is calculated, allowing us to calculate the percentage increase in speed between the two tasks (with the assumption that post-editing will be faster).

There are a number of variables in such an evaluation that could have an impact on the veracity of the results. These variables, and the steps we have taken in our evaluation setup in order to mitigate their impact, are show below:

- **Translator attitude towards MT**: Translators may have a certain bias as relates to MT. We used 24 translators for this evaluation to reduce potential noise from outlying results.
- **Lack of familiarity with the task**: Translators were provided with written instructions, an instructional video, and the opportunity to test the tool prior to beginning the task.
- **Ability of the translator**: Some translators may adapt faster to the task of post-editing than others. We used translators with varying levels of experience to reduce this effect.
- **Difficulty of the test set**: A particularly challenging test set for MT/translation could produce skewed results. We used four different test sets to avoid this.

Each translator translated more than 200 segments each and the findings are presented in the next section.

¹ https://evaluate.taus.net/evaluate/dqf/dynamic-quality-framework
3.2. Results: a resounding success

The main finding from the practical evaluations was that 83% of the translators were faster when post-editing MT output for Chinese to English translation. Figure 2 below shows that almost 40% of translators were more than 30% faster, while only 4 translators were slower while post-editing.

![Figure 2 Range of improvement across translators](image)

Post-editing was shown to be effective for both experienced and inexperienced translators, and consistent improvements were seen across all test sets.

4. Conclusions

The concept of a one-size-fits-all solution for MT is far-fetched. In order to achieve good quality results, especially for difficult languages and domains, engines must be highly tuned with a number of syntactic and content-specific processes, on top of a baseline SMT architecture.

We have demonstrated the effectiveness of this approach in improving machine translation quality for Chinese to English patent translation. The quality and usability of the MT output was validated in translation evaluations which showed a significant increase in the productivity of translators who were post-editing the MT output in a practical translation scenario.
Response-Based Learning for Patent Translation

Stefan Riezler
riezler@cl.uni-heidelberg.de
Computational Linguistics & IWR
Heidelberg University, 69120 Heidelberg, Germany

Abstract
In response-based structured prediction, instead of a gold-standard structure, the learner is given a response to a predicted structure from which a supervision signal for structured learning is extracted. Applied to statistical machine translation (SMT), different types of environments such as a downstream application, a professional translator, or an SMT user, may respond to predicted translations with a ranking, a correction, or an acceptance/rejection decision, respectively. We present algorithms and experiments that show that learning from responses alleviates the supervision problem and allows a direct optimization of SMT for tasks such as cross-lingual patent prior art retrieval, or translation of technical patent documents.

1 Introduction
Response-based learning describes a range of statistical learning methods that replace the full-information supervised learning scenario by extracting supervision signals from the response of an extrinsic environment to a predicted translation. Learning proceeds by “trying out” or “grounding” translations in a task that is external to translation itself, receiving a response from interacting in this task, and converting this response into a supervision signal for updating model parameters. We focus on response-based structured prediction that operates according to the following learning protocol:

1. Environment generates input structure $x_t$
2. Learner predicts output structure $y_t$
3. Environment generates response signal $r_t$
4. Learner uses pair $(x_t, r_t)$ to update its prediction rule

Clearly, the key advantage of response-based learning is to alleviate the supervision problem by a repeatable generate-and-test procedure where feedback is obtained from the environment. Such feedback is in general easier and less costly to obtain than full supervision information. Furthermore, learning from task-based responses has the effect of grounding the learning process in an extrinsic environment.

In this paper, we will describe three different types of response signals that are elicited by grounding SMT into three different environments, namely ranking responses elicited in cross-lingual information retrieval (Section 2), correction responses elicited in form of post-edits by professional translators (Section 3), and acceptance/rejection responses elicited in personalized SMT (Section 4).

2 Grounding SMT in Cross-Lingual Patent Retrieval
The industry standard in cross-lingual information retrieval (CLIR) is to use established translation models for context-aware translation of query strings, effectively reducing the problem
of CLIR to a pipeline of direct translation (DT) and monolingual retrieval (Chin et al., 2008). Only recently, research approaches to probabilistic structured queries (PSQ) have been presented, that is, they include (weighted) translation alternatives into the query structure to allow a more generalized term matching (Ture et al., 2012a,b). In both DT and PSQ, the integration of SMT remains agnostic about its use for CLIR, and is instead optimized to match fluent, human reference translations. In contrast, retrieval systems often use bag-of-word representations, stopword filtering, and stemming techniques during document scoring, and queries are rarely fluent, grammatical natural language queries (Downey et al., 2008).

Attempts to inform the SMT system about its use for retrieval by optimizing its parameters towards a retrieval objective have been presented in the form of re-ranking (Nikoulina et al., 2012) or ranking (Sokolov et al., 2014).

The most direct integration of SMT and CLIR has been presented by Hieber and Riezler (2015) in an approach to Bag-of-Words Forced Decoding (BOW-FD) where IR features for words in the bag-of-words representation of documents force the SMT decoder to prefer relevant documents with high probability. The crucial steps of the response-based online learning protocol are instantiated in this approach as follows:

**Prediction:** Full translation hypergraph.

**Response:** BM25 scores of partial translation hypotheses.

**Learning:** Jointly optimize SMT and IR feature weights by direct ranking optimization on relevant documents.

The key advantage of this approach clearly is the exploitation of the full translation search space, which is made possible by decomposable IR features that relate partial translation hypotheses to documents in the retrieval collection. This in turn allows to use the SMT decoder directly in retrieval. The key modeling idea is to factorize the cross-lingual search problem of finding document $d_e$ given query $q_f$ as follows:

$$P(d_e|q_f) = \sum_h P(h|q_f) \times P(d_e|h, q_f)$$

Learning is done by a joint optimization of a score function for linear models $F_{smt}$ and $F_{ir}$ where

$$\text{score}(q_f, d_e) = \max_h e^{F_{smt}(h, q_e, q_f)} + F_{ir}(h, d_e).$$

The parameters of this objective are optimized for ranking for given relevance ranked documents s.t.

$$\text{rank}(d^+, q) > \text{rank}(d^-, q) \iff \text{score}(d^+, q) > \text{score}(d^-, q).$$

Table 1 presents evaluation results for large-scale patent CLIR. It shows a clear advantage for BOW-FD over the aforementioned DT and PSQ approaches. Data used were Japanese-

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>NDCG</th>
<th>PRES</th>
<th>MAP</th>
<th>NDCG</th>
<th>PRES</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>.3678</td>
<td>.5691</td>
<td>.7219</td>
<td>.2554</td>
<td>.5397</td>
<td>.5680</td>
</tr>
<tr>
<td>PSQ</td>
<td>.3642</td>
<td>.5671</td>
<td>.7165</td>
<td>.2659</td>
<td>.5508</td>
<td>.5851</td>
</tr>
<tr>
<td>BOW-FD</td>
<td>*3919</td>
<td>*5963</td>
<td>*7528</td>
<td>*.2883</td>
<td>*.5819</td>
<td>*.6251</td>
</tr>
</tbody>
</table>

Table 1: Retrieval results of baseline systems and BOW-FD on large-scale CLIR task. * denotes significant difference compared to both baselines.
Table 2: BLEU scores over patent test documents and mean difference and standard deviation from baseline (in small font size).

<table>
<thead>
<tr>
<th>Patents</th>
<th>BLEU</th>
<th>$\Delta [\sigma]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>30.26</td>
<td></td>
</tr>
<tr>
<td>rerank</td>
<td>32.54</td>
<td>$+2.28 [\pm 1.47]$</td>
</tr>
<tr>
<td>tm+lm</td>
<td>33.24</td>
<td>$+2.98 [\pm 2.03]$</td>
</tr>
<tr>
<td>tm+lm+rerank</td>
<td>34.02</td>
<td>$+3.76 [\pm 2.08]$</td>
</tr>
</tbody>
</table>

English Patent CLIR data\(^1\) and German-English Wikipedia CLIR data\(^2\). Both datasets were extracted automatically by using the citation graph in patents and Wikipedia to extract ranked relevance links.

For more information on the data, model, and learning of BOW-FD, see Hieber and Riezler (2015).

3 Learning SMT from Translator Post-Edits

Recent research in computer-assisted translation (CAT) has shown that post-editing of machine translations by professional translators leads to improved productivity of translators, and to improved quality of final translations (Koehn, 2009; Garcia, 2011; Green et al., 2013). This scenario can be turned on its head by focusing on the human post-editor supporting the SMT system, leading to a mutually beneficial cycle of human-assisted machine translation where the SMT system performs online learning from human post-edits. The goal is to improve translation consistency for a given document, and to offer the user the experience of a system that immediately learns from corrections (Wäschle et al., 2013; Denkowski et al., 2014; Green et al., 2014).

Patent data are especially well-suited for an application of this scenario because of the high repetitivity of patent documents. Wäschle et al. (2013) and Bertoldi et al. (2014) recently presented an application of human-assisted SMT to patent data and data from legal and IT domains.

The crucial steps of the response-based online learning protocol are instantiated in online learning from post-edits as follows:

**Prediction:** Most probable sentence translation.
**Response:** User post-edit.
**Learning:** Dynamically extend phrase table and language model; update feature weights by reranking.

The key difference between online learning from post-edits instead of from reference translations is the dynamic extension of the phrase table. Wäschle et al. (2013) and Bertoldi et al. (2014) use a constrained search technique (Cettolo et al., 2010) that optimizes the coverage of both source and target sentences. It produces exactly one phrase segmentation and alignment, and allows gaps such that some source and target words may be uncovered. It differs in this respect from forced decoding which produces an alignment only when the target is fully reachable with the given models.

Assume the following phrase segmentation and alignment:

---

\(^1\)111k train + 1,088k test, available under www.cl.uni-heidelberg.de/boostclir
\(^2\)245k train + 1,455k test, available under www.cl.uni-heidelberg.de/wikiclir
From this, three types of phrase pairs can be collected: (i) new phrase pairs by aligning unambiguous gaps (Technical Offer → Offerta Tecnica); (ii) known phrase pairs already present in the given model (Annex → Allegato and to the → all'); (iii) full phrase pairs consisting of the complete source sentence and its user translation (Annex to the Technical Offer → Allegato all’ Offerta Tecnica). Only phrases that contain at least one content word are considered.

The weight update for the extended phrase table is done by updating relative frequency features (collected into a cache during online learning) using a perceptron update on a training example $(x_t, y_t)$ with feature representation $f(x_t, y)$ in case the prediction $\hat{y} = \text{arg max}_y w^\top f(x_t, y)$ does not match the target $y_t$

$$w = w + f(x_t, y_t) - f(x_t, \hat{y}).$$

Table 2 shows experimental results for German-English patent data sampled from title, abstract and description sections from the PatTR corpus. Significant gains in BLEU score can be obtained by reranking using a perceptron, or by updating phrase table (tm) and language model (lm). In combination, nearly 4 BLEU points improvement are achieved.

For more information on data, model, and learning, see Wäschle et al. (2013) and Bertoldi et al. (2014).

4 Towards Personalized SMT: Learning from Partial User Feedback

While the above described post-editing scenario is a big step towards high-quality human-assisted SMT, the high cost and effort of post-editing may dampen the excitement about this scenario. The question to ask is whether production-quality post-edits from professional translators are really necessary for human-assisted SMT, or whether weaker feedback from less specialized users might be sufficient for learning.

Sokolov et al. (2014) recently addressed this question from a coactive learning perspective that provides a formal notion of feedback strength. They present a convergence analysis of online structured prediction algorithms that learn from feedback consisting of slight improvements over predicted translations, instead of optimal feedback consisting of full post-edits or gold-standard translations. A simulation experiment on news data confirmed this theoretical finding. This research implies that well-known online structured predictors can be used for learning from weak feedback, without changes to the algorithms. What is open for change is the strength of feedback, allowing “light” post-edits from non-professionals. An application of this scenario to online learning for patent translation is a desideratum for future work.

We refer the reader to Sokolov et al. (2014) for more information on the theory and proof-of-concept experiments in the coactive learning framework.

Another attack at learning from weak feedback can be taken from the direction of bandit learning. Learning from bandit feedback describes an online learning scenario, where on each of a sequence of rounds, a learning algorithm makes a prediction, and receives partial information in terms of feedback to a single predicted point. In difference to the full information
Algorithm 1 Bandit Structured Prediction

1: Input: sequence of learning rates $\gamma_t$
2: Initialize $w_0$
3: for $t = 0, \ldots, T$ do
4: Observe $x_t$
5: Calculate $E_{p_{w_t}(y'|x_t)}[\phi(x_t, y')]$
6: Sample $\tilde{y}_t \sim p_{w_t}(y'|x_t)$
7: Obtain feedback $\Delta(\tilde{y}_t)$
8: Update $w_{t+1} = w_t - \gamma_t \Delta(\tilde{y}_t)(\phi(x_t, \tilde{y}_t) - E_{p_{w}(y'|x_t)}[\phi(x_t, y')])$

Algorithm Structured Dueling Bandits

1: Input: $\gamma, \delta, w_0$
2: for $t = 0, \ldots, T$ do
3: Observe $x_t$
4: Sample unit vector $u_t$ uniformly
5: Set $w'_t = w_t + \delta u_t$
6: Compare $\Delta(\tilde{y}_{w_t}(x_t))$ to $\Delta(\tilde{y}_{w'_t}(x_t))$
7: if $w'_t$ wins then
8: $w_{t+1} = w_t + \gamma u_t$
9: else
10: $w_{t+1} = w_t$

supervised scenario, the learner does not know what the correct prediction looks like, nor what would have happened if it had predicted differently. Applied to SMT, this means that the learning algorithm only has access to a $1 - \text{BLEU}$ loss evaluation of a predicted translation instead of obtaining a gold standard reference translation. Clearly, a one-shot user quality estimate of the predicted translation is easier and faster to obtain than light or full post-edits of predicted translations, or than reference translations generated from scratch. This framework can be seen as a first step towards personalized machine translation where a given large SMT system is adapted to a user solely by single-point user feedback to predicted structure.

The crucial steps of the response-based online learning protocol are instantiated in this approach as follows:

**Prediction:** Exploration/exploitation sampling of sentence translation.

**Response:** User feedback on loss value of sampled translation.

**Learning:** Stochastic update using unbiased estimate of gradient.

Sokolov et al. (2015) presented two algorithms for online structured prediction in SMT from bandit feedback that implement these ideas as follows. Algorithm 1 optimizes an expected $1 - \text{BLEU}$ loss criterion (Och (2003), Smith and Eisner (2006), He and Deng (2012), Auli et al. (2014), Wuebker et al. (2015), *inter alia*) by performing simultaneous exploration/exploitation by sampling translations from a Gibbs model (line 6), and using the obtained user feedback (line 7) to perform an update in the negative direction of the instantaneous gradient (line 8).

The second algorithm extends Yue and Joachims (2009)’s dueling bandits algorithm to a Structured Dueling Bandits algorithm. It compares a current weight vector $w_t$ with a neighboring point $w'_t$ along a direction $u_t$, performing exploration (controlled by $\delta$, line 5) by probing random directions, and exploitation (controlled by $\gamma$, line 8) by taking a step into the winning direction. The comparison step in line 6 is adapted to structured prediction from the original algorithm by comparing the quality of $w_t$ and $w'_t$ via an evaluation of the losses $\Delta(\tilde{y}_{w_t}(x_t))$ and $\Delta(\tilde{y}_{w'_t}(x_t))$. 
Table 3: Corpus BLEU on test set for SMT domain adaptation from Europarl to NewsCommentary by k-best reranking.

\[ \Delta(\hat{y}_{w'}(x_t)) \] of the structured arms corresponding to predicting the most probable translation under \( w_t \) and \( w'_t \), respectively.

Sokolov et al. (2015) present an evaluation that follows the standard of simulating bandit feedback by evaluating task loss functions against gold standard structures without revealing them to the learner. Here the setup is a reranking approach to SMT domain adaptation where the k-best list of an out-of-domain model is re-ranked (without re-decoding) based on bandit feedback from in-domain data. This can also be seen as a simulation of personalized machine translation where a given large SMT system is adapted to a user solely by single-point user feedback to predicted structures.

The out-of-domain baseline SMT model is trained on 1.6 million parallel Europarl data and includes the English side of Europarl and in-domain NewsCommentary in the language model. The full-information in-domain SMT model gives an upper bound by MERT tuning the out-of-domain model on in-domain development data. Learning under bandit feedback started at the learned weights of the out-of-domain median model. It uses the parallel NewsCommentary data to simulate bandit feedback, by evaluating the sampled translation against the gold standard reference using as loss function \( \Delta \) a smoothed per-sentence \( 1 - \text{BLEU} \) (by flooring zero n-gram counts to 0.01).

Table 3 shows the final results that were obtained by online-to-batch conversion where the model trained for 100 epochs on in-domain training data is evaluated on a separate in-domain test set. Results for Bandit Structured Prediction and Dueling Bandits are very close, however, both are significant improvements over the out-of-domain SMT model that even includes an in-domain language model. The range of possible improvements is given by the difference of the BLEU score of the in-domain model and the BLEU score of the out-of-domain model – nearly 3 BLEU points. Bandit learning can improve the out-of-domain baseline by 1.26 BLEU points (Bandit Structured Prediction) and by 1.52 BLEU points (Dueling Bandits). Clearly, a comparison between Bandit Structured Prediction and Dueling Bandits is skewed towards the latter approach that has access to two-point feedback instead of one-point feedback as in the former case. It has been shown that querying the loss function at two points leads to convergence results that closely resemble bounds for the full information case (Agarwal et al., 2010), however, such feedback is clearly twice as expensive and, depending on the application, might not be elicitable from users.

We refer the reader to Sokolov et al. (2015) for more information on data, model, and experiments.

## 5 Conclusion

We presented a comprehensive perspective on machine learning approaches that attempt to replace the full information supervised scenario by a setup in which supervision signals are extracted from responses of an extrinsic environment to system predictions. The discussed types of responses ranged from rankings deduced from performance of translations in extrinsic tasks such as cross-lingual retrieval, to improvements of structures by post-edits of professional translators, to partial feedback consisting of mere assessments of the quality of the predicted translation. We showed the efficacy of response-based learning in several simulation experi-
ments. Clearly, improvements over traditional full-information structured prediction cannot be expected from learning from such weaker types of feedback. Instead, the goal is to investigate learning situations in which full information is not available. Moreover, task-based feedback might be even preferable to independently created gold standard structures if the ultimate goal is improved performance of a translation-related extrinsic task.

In future work, we would like to apply response-based learning from weak feedback to real-life interactive scenarios. The new challenges of future work will be an investigation of response signals that can be elicited from humans efficiently and reliably, and still are informative enough for learning in SMT.

Acknowledgements

This research was supported in part by DFG grant RI-2221/1-2 “Weakly Supervised Learning of Cross-Lingual Systems”.

References


Promoting Science and Technology Exchange using Machine Translation

Toshiaki Nakazawa
nakazawa@pa.jst.jp
Japan Science and Technology Agency (JST), 5-3, Yonbancho, Chiyoda-ku, Tokyo 102-8666
Japan

1 Introduction

There are plenty of useful scientific and technical documents which are written in languages other than English, and are referenced domestically. Accessing these domestic documents in other countries is very important in order to know what has been accomplished and what is needed next in the science and technology fields. However, we need to surmount the language barrier to directly access these valuable documents. One obvious way to achieve this is using machine translation systems to translate foreign documents into the users’ language. Even after the long history of developing machine translation systems among East Asian languages, there is still no practical system. We have launched a project to develop practical machine translation technology for promoting science and technology exchange. As the starting point, we aim at developing Chinese ↔ Japanese practical machine translation system. In this talk, I will introduce the background, goals and status of the project. Also, I will give you the summary of the 2nd Workshop on Asian Translation (WAT2015)\(^1\) where Chinese ↔ Japanese scientific paper translation subtasks has been carried out.

2 Background

Figure 1 shows the number of scientific papers in the world which are written in “English”. We can presume that the number of papers written in each language has the similar proportion to this graph. You can see that the number of papers from China is rapidly growing in recent years, which means we have a large number of “Chinese” papers.

Some of them may include important and useful information but are never published in English. We, not Chinese native speakers, cannot get any information from such papers as they are, and throw them to the MT engines to translate into our mother tongue. Chinese-to-English and Japanese-to-English MT systems have been developed for years and the quality are sufficient enough for gisting the paper. However there is little MT system between Asian languages such as Chinese-to-Japanese and vice versa which is good enough for gisting. Therefore, we have launched a project to develop practical machine translation technology between East Asian languages.

3 Goals

We have 3 goals in the project and they are summarized in Figure 2.

\(^1\)http://lotus.kuee.kyoto-u.ac.jp/WAT/
1. Language Resource Construction
   Most of recent corpus-based machine translation systems require parallel corpus where translation rules are acquired. In addition, parallel dictionary is necessary to cover technical terms because the number of technical terms keeps growing.

2. Sentence Analyzers (especially for Chinese)
   Both Chinese and Japanese require the word segmentation technology because they do not have white spaces between words. The word segmentation of Chinese is more difficult than that of Japanese because Chinese sentences are basically composed of only Chinese characters. Also, Chinese and Japanese have different language characteristics, so the high-level sentence analysis such as parsing is important to achieve high translation quality.

3. MT Engine Development
   Our MT engine should be able to use the rich information from the deep sentence analysis. We choose dependency-to-dependency example-based machine translation method in our project.

4 Workshop on Asian Translation
   The Workshop on Asian Translation (WAT) is a new open evaluation campaign focusing on Asian languages organized by NICT\(^2\) and JST. The 2nd WAT has 6 subtasks: English ↔ Japanese scientific paper translation, Chinese ↔ Japanese scientific paper translation, Chinese → Japanese patent translation and Korean → Japanese patent translation. The distinctions of WAT are:

---
\(^2\)National Institute of Information and Communications Technology
Figure 2: The goals of the JST Chinese-Japanese MT project.

- Open innovation platform
  The test data is fixed and open, so you can repeat evaluations on the same data and confirm changes in translation accuracy over time. WAT has no deadline for the automatic translation quality evaluation (continuous evaluation), so you can submit translation results at any time.

- Domain and language pairs
  WAT is the world’s first workshop that uses scientific papers as a domain and Japanese-Chinese as a language pair. In the future, we will add more Asian languages, such as Korean, Vietnamese, Indonesian, Thai, Myanmar and so on.

- Context-aware evaluation
  The test data of WAT is prepared using the paragraph as a unit, while almost all other evaluation campaigns use the sentence as a unit. We would like to consider how to realize context-aware evaluation in WAT.

- Evaluation method
  Evaluation will be done by both automatic and human evaluation. For human evaluation, WAT will use crowdsourcing, which is low cost and allows multiple evaluations.

  I will introduce the results and some insights acquired from this year’s workshop (WAT2015).

Xiaona Ren
renxiaona@lingosail.com
Yongpeng Wei
weiyongpeng@lingosail.com
Rile Hu
hurile@lingosail.com
Beijing Lingosail Tech Co., Ltd, Qingyun Dangdai Building, No.43, West Road of North 3rd Ring Road, Haidian District, Beijing, P. R. China

Abstract
In this paper, we propose a new approach to improve human post-editing efficiency by simplifying the sentence structure on Chinese-to-English patent machine translation (PMT). To simplify the structure of a patent sentence, we use a recognizer to recognize the max noun phrases (MNPs) \(^1\) in a Chinese sentence before translating the sentence. The MNPs are replaced with their head words in the sentence, which makes the sentence structure simpler to be translated. Therefore, the task of translating a complicated sentence is transformed into two subtasks: one is the translation task of MNPs and the other is the translation task of the simplified sentence. And then, the translation results of two subtasks are combined to get the final translation result of the input Chinese sentence. This method outperforms NTCIR-10 official baseline by approximate 2 BLEU points. Moreover, the translation results are beneficial for human post-editing, which can save human post-editing time and improve the quality of translation.

1 Introduction
Patent information is important for communities all around the world, and there is a significant practical need for translations in order to understand patent information written in foreign languages and to apply for patents in foreign countries (Goto et al., 2013). PMT is a significant and important task.

Patent document belongs the category of scientific literature. Patent constitute one of the challenging domains for machine translation because patent sentences can be quite long and contain complex structures, and the presentation of sentence is rigorous and contains a lot of professional terms. Moreover, the structure of Chinese sentence is complex, and the word order is rather arbitrary. Therefore it is difficult to translate a sentence in Chinese patent with general phrase-based statistical machine translation (SMT) method.

In this paper, we propose a novel method for Chinese-to-English PMT, which uses a MNP recognizer to recognize the MNPs in Chinese sentence for simplifying the sentence structure. That is to say, the task of translating a long Chinese sentence can be divided into two subtasks, one subtask is the translation of MNP, and the other subtask is the translation of simplified

\(^1\)Maximal noun phrase is the noun phrase which is not contained by any other noun phrases.
Figure 1: The translation process of a example sentence.

sentence. And then the translation results of two subtasks can be combine to get the final translation result of input Chinese sentence. Our method can be illustrated by the following example sentence as shown in Figure 1. “管线处理器的简化实例的功能方块图” and “根据本文所述技术的条件指令处理” are the MNPs of the example sentence, which be replaced with their core heads to simplify the sentence structure. Therefore, the simplified sentence is “图1是功能方块图，所述管线处理器可实施条件指令处理”。 And then combine the translation results of MNPs and the simplified sentence to get the final translation of the input sentence. Different from the previous works, our approach has two advantages. One is that our approach can improve the translation performance of long sentences on Chinese-to-English PMT, and the other is that the translation result is convenient to be post-edited by translators.

The rest of the paper is organized as follows: Section 2 describes the related works, and Section 3 introduces our translation system. We will present the evaluation results in Section 4. Finally, we conclude this paper and discuss future work in Section 5.

2 Related Works

Most recent researches on Chinese-to-English PMT are based on phrase-based translation model (Fujita and Carpuat, 2013; Zhao et al., 2013b), and some people (Zhao et al., 2013a) use a combined output from two types of grammars supported in their SMT engine (Zheng, 2007), with two different word segmentations. Huang et al. (2013) investigated a sentence-level language model adaptation approach to take advantage of the characteristics of patent documents, and developed SMT system based on the string-to-dependency translation model (Shen et al., 2010) for a variety of languages and contributed substantially to the improvement of translation quality.

One of the characteristics of patent sentences is long, complicated modification. A modification is identified by the presence of a head word. Yokoyama (2013) constructed a modifier correcting system using head words extracted from about 1 million patent sentences, which corrected 60% of the errors. However, our system used a MNP recognizer to recognize the MNPs of patent sentence for simplifying the structure of patent sentence, which can improve the performance of PMT and the efficiency of human post-editing.

Post-editing (PE) is not a new topic in MT-related research, it was studied quite eagerly already in the 1980s especially in Europe where international organizations had a need to share information rapidly in many languages. The initial motivation was to develop effective strate-
gies and methodologies of human PE in order to make the most efficient use of MT output (Tatsu-
sumi, 2010). Sometimes only accuracy is needed, but sometimes stylistic refinement is re-
quired (McElhaney and Vasconcellos, 1988; Austermuhl, 2001; Allen, 2001; TAUS, 2010). Ebara (2007) reports on their system which is specifically developed to work on English to
Japanese patent translation. Based on the experiment and evaluation, he concluded that the
rule-based part of the system is good at handling structural transfer, and the statistical part of
the system is good at lexical transfer of technical terms. Our approach is inspired by a similar
idea used in post-edit process. We can simplify the structure of patent sentence with a MNP
recognizer, which is advantage for translators to handle the structure of sentence. And we use
the method of statistical machine translation, which is advantage for translators to translate the
patent terms.

3 Overview of the translation system

As illustrated in Figure 2, our translation system consists of the following five steps:

- Recognize the MNPs of a input Chinese sentence $i$ using a MNP recognizer;

- Replace the MNPs with their head words, and simplify the input sentence $i$, and generate
  the simplified sentence $s$;

- Use the translation model $m$ for translating each MNP;

- Use the translation model $m$ for translating the simplified sentence $s$;

- Combine the translation of MNPs and the simplified sentence $s$, and generate the final
  translation result $o$ of the input sentence $i$. 

Figure 2: The framework of our machine translation system.
3.1 MNP Recognizer

In this work, we use a statistical method to recognize MNPs of a sentence, which is similar to the related work (Ren et al., 2010). We use the same feature template, but not include the post-process step. The Chinese Treebank (CTB) 5.1 is used in this experiment, and is split into three partitions for training, developing and testing, respectively, following its conventional split in most previous works in the field, as shown in Table 1.

<table>
<thead>
<tr>
<th>Sections</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>others 18,104</td>
</tr>
<tr>
<td>Dev</td>
<td>271-300 350</td>
</tr>
<tr>
<td>Test</td>
<td>301-325 348</td>
</tr>
</tbody>
</table>

Table 1: The split of CTB 5.1 for experiments.

We conducted three experiments to show the performance of the MNP recognizer. Figure 3 shows the flow charts of three experiments. The first experiment is to evaluate the performance of the MNP recognizer based on the CTB 5.1 test corpus that word segmentation and part-of-speech tagging are full correct (test_set1). In the second experiment, we used a Chinese word segmentation tool (i.e. Institute of Computing Technology, Chinese Lexical Analysis System, namely, ICT-CLAS) to implement word segmentation and the Stanford POS tagger to tag CTB 5.1 test corpus (test_set2). The third experiment is to evaluate the performance using the same method as the second experiment just the test sentences are different. In the third experiment, we used the NTCIR-10 Chinese-English bilingual patent text corpus (test_set3) as our experimental test set. The results of three comparative experiments are shown in Table 2.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>82.59%</td>
<td>84.09%</td>
<td>83.33%</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>71.43%</td>
<td>72.29%</td>
<td>71.86%</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>61.18%</td>
<td>63.41%</td>
<td>62.28%</td>
</tr>
</tbody>
</table>

Table 2: The results of three experiments.

However, as can be seen from Table 2, the performance of MNP recognizer is relatively poor on patent domain. One reason is that the word segmentation and POS tagging errors affect the MNP recognition, and the other important reason is that the model for MNP recognition was
trained by Chinese Treebank rather than the patent corpus. Because there is not the Treebank of patent corpus, and manual tagging corpus takes a lot of time and hard work. Although the performance of MNP recognition is relatively poor on patent domain, the MNP recognizer still is favorable for simplifying the sentence structure for machine translation.

3.2 MNP Analysis

In this part, we analysis the influence of using a MNP recognizer on the machine translation. The boundaries of MNPs may be recognized wrong, but not all of them are harmful for translation, which can be seen in the following example.

Input Chinese Sentence: 为了更清楚地说明本发明[实施例或现有技术中的技术方案].
Output English Sentence: to more clearly illustrate the present invention [embodiments or in the prior art technical solution].

In the example, the phrases in the brackets are the recognized MNP of the input Chinese sentence and the corresponding English translation result in output English sentence. We can see that the left boundary of Chinese MNP is wrong. The correct MNP of input Chinese sentence should be recognized as the phrase “本发明实施例或现有技术中的技术方案”. But, the translation result is acceptable for the post-editing translators, which does not affect the post-editing translators to understand the sentence structure. Therefore, we conducted the following experiments to analysis in detail the influence of the MNP recognizer.

We randomly extract 300 sentences from patent test set, and use our MNP recognizer to recognize MNPs from them. There are 305 MNPs be recognized, which contain 192 MNPs are recognized to be right, 113 MNPs are recognized to be wrong. And then we define the MNPs that are beneficial for translators to post-edit translation as good MNPs, and otherwise as bad MNPs. In the Table 3 we present the results of our experiments.

<table>
<thead>
<tr>
<th></th>
<th>Good</th>
<th>Bad</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right MNPs (192)</td>
<td>188</td>
<td>4</td>
<td>97.92%</td>
</tr>
<tr>
<td>Wrong MNPs (113)</td>
<td>44</td>
<td>69</td>
<td>38.94%</td>
</tr>
</tbody>
</table>

Table 3: The experimental results of MNP analysis.

From the experimental results, we can see that 97.92 percent of right MNPs are convenient to simplify the structure of sentence and post-edit the machine translation results. Moreover, 38.94 percent of wrong MNPs also are convenient for translators to understand the structure of sentence and post-edit the translation results. Therefore, the method of simplifying sentence structure using a MNP recognizer is beneficial for translators to understand the overall structure of sentence and improve the human post-editing efficiency.

3.3 Translation System

Our machine translation system is a phrase-based system based on two translation models. Like many other MT systems, we use two phrase-based SMT models with some factored translation features for the MNP and simplified sentence translation tasks. However, the two translation models were trained by standard methods. The only difference between them is that the distortion limit parameter value of translating simplified sentence is set as 6 while the distortion limit parameter value of translating MNP is set as 3, which is the best value proved by our experiments.

3.4 Post-Editing Process

Depalma and Kelly (2009) state that even when the MT output needs human Post-editing (PE), it is generally faster and cheaper than human translation, and when the cost is the same, MT
plus PE achieves faster turnaround. Also, some studies have shown that the quality of the final product of MT plus PE can in some cases exceed the quality of human translation (Fiederer and O’Brien, 2009; Koehn, 2009), which is especially obvious in the domain of patents. Therefore, it may further justify the increasing employment of this workflow.

In this paper, we develop a platform for Chinese-English machine translation. Firstly, the imputed Chinese sentence can be translated to English by our MT method, and then be corrected into right English translation by translators with the post-editing interface of our platform. This workflow of post-editing includes some basic operation, such as add, delete, remove, modify, etc. And we highlight the MNPs of sentence for translators more convenient to translate. The machine translation system is available on our web site, and the interface is shown as the following Figure 4.

4 Experiments

The experiment process is divided into two steps. The first step is to evaluate the performance of our translation system, which recognizes the MNPs with a MNP recognizer for simplifying the sentence structure. The second step is to evaluate whether the translation of our translation system is more advantageous for post-editing translators than other statistical translation systems.

4.1 Data Sets

We used the experimental corpus from the NTCIR-10 Patent MT Chinese-to-English task, which contains 1 million bilingual sentences pairs for training and 2,000 sentence pairs for development data, and monolingual patent corpus in English covering a span of 13 years (1993-2005).

To measure the overall performance of our method (i.e. translation system and post-editing strategy), we used two metrics: automatic evaluation scores (sentence-level) and human evaluation scores. We then applied some currently available automatic evaluation methods BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) metric to evaluate the performance of our translation system. BLEU and NIST scores were calculated using NIST’s mteval-v13a.pl\textsuperscript{2}. The

\textsuperscript{2}http://www.itl.nist.gov/iad/mig/tools/
human evaluation method is used to evaluate the efficiency of post-editing, which include three indicators: degrees of fit and fluent, the number of operation steps and time. The detail information of evaluation will be shown on subsection 4.3.

4.2 Automatic Evaluation

We trained standard phrase-based SMT system (baseline) on training set and tuned model parameters on development set using Moses (Koehn et al., 2007). A trigram language model (LM) was trained on the English side of the training set by SRILM Toolkit for both systems. GIZA++ (Och and Ney, 2003) was used to obtain symmetric word alignment model. We used a Chinese word segmentation tool (ICT-CLAS) to implement word segmentation. The automatic evaluation results including the baseline system and our system are shown in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Our system</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>33.22</td>
<td>33.84 (+0.62)</td>
</tr>
<tr>
<td>NIST</td>
<td>8.2389</td>
<td>8.2763 (+0.0374)</td>
</tr>
</tbody>
</table>

Table 4: The automatic evaluation results of two systems.

From the results in Table 4, it can be seen that although our system achieves slight improvements over the baseline system, there is still much to be done to improve it. However, our system is more beneficial to translate complex sentences. We randomly extract 1500 sentences from the test set (Auto_test_set1) as a new test set (Auto_test_set2) that contains at least 70 Chinese words. Table 5 is shown the automatic evaluation results based on Auto_test_set2.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Our system</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>32.76</td>
<td>34.46 (+1.7)</td>
</tr>
<tr>
<td>NIST</td>
<td>8.2233</td>
<td>8.4171 (+0.1938)</td>
</tr>
</tbody>
</table>

Table 5: The automatic evaluation results of two systems based on Auto_test_set2.

Obviously, the experimental results reveal that our method is superior to the baseline system based on the new test set. It can be seen that the accuracy rate of MNP recognizer is not high, but our method improved the performance by 1.7 BLEU points and 0.1685 NIST points. Therefore, our translation system is more beneficial to translate long sentences.

4.3 Human Evaluation

In this subsection, we use manual evaluation method to evaluate the improvement of efficiency for post-editing translators. The manual evaluation method contains the following three metrics:

1. Semantic accuracy and fluency (SA and SF), which follows the general evaluation criterion of machine translation system (King, 1996).

2. Translation error rate (TER), which is the ratio of the number of edits incurred to the total number of words in the reference text (Przybocki et al., 2006; Snover et al., 2006).

3. Time, which is the individual productivity in words per hour based on the results of machine translation system (Plitt and Masselot, 2010).

We used 300 sentences with average 230 characters from set2 as the test set of this subsection. Three translators for patent participated in the human evaluation work. Each translator
evaluated the semantic accuracy and fluency of the translation of baseline system and our system. The average values and the allocation of scores of semantic accuracy and fluency are shown in Table 6.

TER is an error metric for machine translation that measures the number of edits required to change a system output into one of the references. We define the edit distance similar to the reference (Przybocki et al., 2006), which is the number of insertions, deletions, and substitutions that are required in order to make a system translation equivalent in meaning to that of a reference translation, using understandable English. We also use publicly available software\(^3\) developed by Snover (Snover et al., 2005) to calculate edit distance. We divided the test sentences into three groups, and there were a total of six test participants. Each two participants are the same group, and they translate the same test sentences based on two different machine translation results. Figure 5 shows the individual productivity of each test participant in words per hour, and the translation error rates of each participant are shown in Figure 6.

From the above two Figures and Table 6, we can see that the results of manual evaluation are consistent with the results of automatic evaluation as shown in Table 5. Our translation

---

\(^3\)http://www.cs.umd.edu/~snover/tercom/

---

### Table 6: The human evaluation results of semantic accuracy and fluency.

<table>
<thead>
<tr>
<th></th>
<th>Allocation of scores</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>2.98</td>
<td>6</td>
<td>75</td>
<td>132</td>
<td>82</td>
<td>5</td>
</tr>
<tr>
<td>Our system</td>
<td>3.16</td>
<td>8</td>
<td>89</td>
<td>149</td>
<td>52</td>
<td>2</td>
</tr>
<tr>
<td><strong>SF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>2.78</td>
<td>3</td>
<td>52</td>
<td>126</td>
<td>113</td>
<td>6</td>
</tr>
<tr>
<td>Our system</td>
<td>2.95</td>
<td>7</td>
<td>60</td>
<td>145</td>
<td>87</td>
<td>1</td>
</tr>
</tbody>
</table>

---

**Figure 5: Individual productivity in words per hour.**
system performs somewhat better than the baseline system, though there is still much to be explored and improved. Figure 5 shows that post-editing time is shorter based on the translation provided by our system than baseline system, and we can see from the Figure 6 the translation error rates are relatively lower based on the translation provided by our system than baseline system. Therefore, our translation system is more suitable for translators to post-editing.

5 Conclusions and Future Works

The experiments show the efficiency of the proposed method of simplifying sentence structure in the field of PMT, and this proposed method is more efficient to improve the speed and quality of translation for post-editing translators.

In future work, we would like to explore further performance improvement of our translation system. Accurately, there is a big space for our system to improve the performance of translation. On the one hand, we can improve the accuracy of MNP recognizer by increasing the size of the training corpus. We build an extended patent training set automatically using a Chinese word segmentation tool (ICT-CLAS), Stanford POS tagger and our MNP recognizer, and then correct the high frequency of MNPs. The extended model is more suitable for the MNP recognition in the domain of patents. On the other hand, the training corpus for training MNP translation model also can be extended, which be added 1000 pairs of bilingual MNPs. We make a preliminary experiment using the extended training corpus to train a new MNP translation model. The preliminary experiments indicate that the method achieves some improvements in the BLEU score on the same test data.

References


Enhancing Function Word Translation with Syntax-Based Statistical Post-Editing

John Richardson  
john@nlp.ist.i.kyoto-u.ac.jp  
Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

Toshiaki Nakazawa  
nakazawa@pa.jst.jp  
Japan Science and Technology Agency, Kawaguchi-shi, Saitama 332-0012, Japan

Sadao Kurohashi  
kuro@i.kyoto-u.ac.jp  
Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

Abstract

The generation of precise and comprehensible translations is still a challenge in the patent and scientific domain. In particular, function words are often poorly translated in standard machine translation systems, particularly across language pairs with greatly differing syntax. In this paper we exploit the target-side structure in tree-to-tree machine translation to post-edit function words automatically using a tree-based function word language model. We show that a significant improvement in human evaluation can be achieved with our proposed method.

1 Introduction

A high level of machine translation fluency is sought after in all subject domains. Translations with high adequacy however are especially important in patent and scientific translation, where it is particularly necessary to preserve the meaning of the input sentence in the generated translation.

Error analysis of state-of-the-art machine translation systems has shown that poorly translated function words are a major cause of loss in translation comprehensibility. For example, negation and passive structures can completely reverse their meaning when missing the correct function words, and it is important for understanding to select appropriate prepositions. We have also found that lack of (or incorrectly placed) relative pronouns has a large effect on preserving sentence meaning, and that badly formed punctuation impedes understanding.

Surprisingly few studies have been made specifically on improving function word translation for statistical machine translation systems, despite this having been looked at in rule-based systems (Arnold and Sadler, 1991). While we were unable to find any previous work on function word statistical post-editing, function words have been used to generate translation rules (Wu et al., 2011). The most similar approach to our method of editing function words used structural templates and was proposed for SMT (Menezes and Quirk, 2008). Statistical post-editing of MT output in a more general sense (Simard et al., 2007) and learning post-editing rules based on common errors (Elming, 2006; Huang et al., 2010) have shown promising results. The majority of statistical post-editing methods work directly with string output, however a syntactically motivated approach has been tried for post-editing verb-noun valency (Rosa et al., 2013).
Figure 1: String vs Tree Output: The intended meaning of the translation is often unclear from string output. In this case we cannot tell easily that ‘translate documents’ is a relative clause (missing the relative pronoun ‘which’ or ‘that’) and that ‘the paper’ is a prepositional phrase (missing the preposition ‘in’) rather than the direct object of ‘described’.

We believe that the intended meaning of a sentence is often unclear from flat MT output. For example, in Figure 1, the intended meaning is much clearer from the dependency tree representation. Based on this observation, we attempt to exploit the target structure of the output of a dependency tree-to-tree machine translation system in order to understand better the cause of the function word errors and therefore correct them more effectively.

2 Syntax-Based Post-Editing

Our proposed model starts with the dependency tree output of a tree-to-tree machine translation system. From this we analyze the position of function words and attempt to modify them with a tree-based function word language model.

We assume a set of function words $F$, a subset of the entire target-side vocabulary. We also define an empty token ‘<none>’ which represents the lack of a function word. A root node and leaf nodes can be added to the tree to allow insertion of function words as the sentence root and leaves respectively.

A dependency tree can be decomposed into token–head pairs $(t, t')$. We derive a simple language model $P(f \mid t, t')$ approximating the probability of function word $f \in F$ being inserted between $t$ and $t'$. The model is estimated over the training data by counting the occurrence of $(f, t, t')$ tuples where $f$ is a function word appearing between $t$ and $t'$. Note that to make this definition well-defined, we strictly require that function words have only one child. The probability $P(f \mid t, t')$ is then calculated as:

$$P(f \mid t, t') = \sum_{g \in F \cup \langle \text{none} \rangle} \frac{\text{count}(f, t, t')}{\text{count}(g, t, t')}$$

In our experiments we include part-of-speech tags inside tokens to reduce homonym ambiguity (e.g. use ‘set-NN’ instead of ‘set’). We also split $P(f \mid t, t')$ into two cases, $P_{left}(f \mid t, t')$ and $P_{right}(f \mid t, t')$, to consider the difference between $t$ being a left or
right descendent of \( t' \). We will write \( P_s \) to refer to whichever of \( P_{\text{left}} \) or \( P_{\text{right}} \) applies in each case.

2.1 Operations

For a token–head pair \((t, t')\), word insertion is performed when \( P_s(f | t, t') > P_s(<\text{none}> | t, t') \) for some function word \( f \). We choose the function word with the highest probability if there are multiple candidates. Replacement of function word \( t \) is performed similarly if \( P_s(\text{child}(t) | f, t') > P_s(\text{child}(t) | t, t') \) for some other function word \( f \). Similarly we choose the best \( f \) if there are multiple candidates. Deletion can be performed using the same method as for replacement by adding the function word ‘<none>’ to \( F \). The full algorithm for post-editing a tree \( \text{Tree} \) is shown in Algorithm 1.

**Algorithm 1** Post-Edit Tree

1: procedure POSTEDIT(Tree)
2: loop:
3: # Traverse tree from left-to-right
4: for \((t, t') \in \text{Tree}\) do
5: if \( t \in F \) then
6: \( \text{child} \leftarrow \text{GetUniqueChild}(t) \)
7: # Find the best function word to replace \( t \)
8: \( \text{max}_f, \text{max}_p \leftarrow t, P_s(t | \text{child}, t') \)
9: for \( f \in F \cup \{<\text{none}>\} \) do
10: if \( P_s(f | \text{child}, t') > \text{max}_p \) then
11: \( \text{max}_f, \text{max}_p \leftarrow f, P_s(f | \text{child}, t') \)
12: end if
13: end for
14: if \( \text{max}_f \neq t \) then
15: # Replace \( t \) with \( \text{max}_f \) and restart for entire tree
16: Tree.Replace(\( \text{max}_f, \text{child}, t' \))
17: goto loop
18: end if
19: else
20: \( \text{max}_f, \text{max}_p \leftarrow t, P_s(<\text{none}> | t, t') \)
21: # Find the best function word to insert
22: for \( f \in F \) do
23: if \( P_s(f | t, t') > \text{max}_p \) then
24: \( \text{max}_f, \text{max}_p \leftarrow f, P_s(f | t, t') \)
25: end if
26: end for
27: if \( \text{max}_f \neq <\text{none}> \) then
28: # Add function word \( \text{max}_f \) and restart for entire tree
29: Tree.Add(\( \text{max}_f, t, t' \))
30: goto loop
31: end if
32: end if
33: end for
34: end procedure
2.2 Filtering Replacements/Deletions with Word Alignments

In the majority of cases we found it counter-productive to replace or delete function words corresponding directly to non-trivial source words in the input sentence. For example, in a Chinese–English translation task, consider the two translations:

- 听/音乐 (listen/music) → listen to music
- 下面/100/米 (below/100/m) → 100m below

In the first sentence, the function word ‘to’ in the English translation has no corresponding word in the Chinese input and therefore its existence is based only on the target language model. In contrast, the preposition ‘below’ in the second sentence directly corresponds to ‘下面 (below)’ in the input and care should be taken not to delete it (or change it to a completely different preposition such as ‘above’).

We therefore propose restricting replacement/deletion to function words that are aligned to trivial or ambiguous source-side words (function words without concrete meaning, whitespace, punctuation). This allows us to change for instance the unaligned ‘to’ in ‘listen to’ but not ‘below’ with an input alignment. The source–target word alignments are stored in the translation examples used by the baseline SMT system and kept track of during decoding.

3 Experiments

3.1 Data and Settings

We performed translation experiments on the Asian Scientific Paper Excerpt Corpus (ASPEC)\(^1\) for Japanese–English translation. The data was split into 3 million training sentences, 1790 development sentences and 1812 test sentences.

We defined English function words as those tokens with POS tags of functional types such as determinants and prepositions, and treated Japanese particles as function words for the purposes of alignment-based filtering. The primary post-editing model was trained on the training fold of the ASPEC data. Since our model only requires monolingual data, for comparison we also trained a separate model on a larger (30M sentences) in-house monolingual corpus (Mono) of technical/scientific documents.

For the baseline SMT system we used KyotoEBMT (Richardson et al., 2014), a state-of-the-art dependency tree-to-tree translation system that can keep track of the input–output word alignments. Post-editing was performed on the top-1 translation produced by the tree-to-tree baseline system.

Japanese segmentation and parsing were performed with Juman and KNP (Kawahara and Kurohashi, 2006). For English we used NLParser (Charniak and Johnson, 2005), converted to dependency parses with an in-house tool. Alignment was performed with Nile (Riesa et al., 2011) and an in-house alignment tool. We used a 5-gram language model with modified Kneser-Ney smoothing built with KenLM (Heafield, 2011).

3.2 Evaluation

Human evaluation was conducted to evaluate directly the change in translation quality of function words. We found that automatic evaluation metrics such as BLEU (Papineni et al., 2002) were not sufficiently sensitive to changes (the change rate is relatively low for post-editing tasks) and did not accurately measure the function word accuracy.

In human evaluation we asked two native speakers of the target language (English) with knowledge of the source language (Japanese) to decide if the system output was

\(^1\)http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
better, worse, or neutral compared to the baseline. A random sample of 20 edited sentences were selected for each experiment and the identity of the systems was hidden from the raters. The Fleiss’ kappa inter-annotator agreement (Fleiss et al., 1981) for wins/losses was 0.663, and when including neutral results this was reduced to 0.285.

3.3 Tuning and Test Experiments
We first performed a preliminary tuning experiment on the development fold of ASPEC to investigate the effect of model parameters. The results in Table 1 show for each row the model settings, the number of wins (+), losses (–) and neutral (?) results compared to the baseline, and the change rate (CR) over the entire development set.

The first three settings (‘OnlyIns’, ‘OnlyRep’, ‘OnlyDel’) show the effects of allowing only insertions, replacements and deletions respectively without using source–target alignments (see Section 2.2). We can see that the quality for deletions is lower than insertions and replacements, and error analysis showed that the major cause was deletion of function words aligned to content words in the input.

We reran the experiments using the alignment-based filtering (‘AlignA’ and ‘AlignB’) and found the results improved. While possible to achieve a higher change rate by allowing all three operations, we could only achieve a slight increase in accuracy by disallowing replacements (the setting ‘AlignB’). The difference was mainly due to alignment errors, which caused more serious problems for replacement as they were able to alter sentence meaning more severely.

The best settings in the tuning experiment (‘AlignB’) were used to conduct the final evaluation on the unseen test data from ASPEC. We also compared models trained on the ASPEC training fold and on our larger monolingual corpus. Table 2 shows the final evaluation results. The results on the test set show significant improvement on win/loss sentences at p < 0.05. There was no clear improvement gained by increasing the size of model training corpus, however the change rate could be improved by using more data.

<table>
<thead>
<tr>
<th>Insert</th>
<th>Replace</th>
<th>Delete</th>
<th>Align</th>
<th>+</th>
<th>-</th>
<th>?</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>OnlyIns</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>OnlyRep</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>11</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>OnlyDel</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>7</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>AlignA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>11</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>AlignB</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>11</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1: Results of tuning experiments on development set.

<table>
<thead>
<tr>
<th>Insert</th>
<th>Replace</th>
<th>Delete</th>
<th>Align</th>
<th>+</th>
<th>-</th>
<th>?</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASPEC</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>12</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Mono</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>11</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Both</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>23</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2: Final evaluation results on unseen data.

4 Error Analysis and Conclusion
The experimental results show that in general our proposed method is effective at improving the comprehensibility of translations by correctly editing function words. Ta-
Table 3: Examples of improved translations after deleting and replacing incorrect function words.

<table>
<thead>
<tr>
<th>Input</th>
<th>基準値の設定に基づくりん酸の測定（モリブデン法）では、 (...)。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Especially, in the measurement of phosphate by simple colorimeter (molybdenum method), (...).</td>
</tr>
<tr>
<td>Proposed</td>
<td>Especially, the measurement of phosphate by simple colorimeter (molybdenum method), (...).</td>
</tr>
</tbody>
</table>

Table 4: Examples of worsened translations. The first example shows a case where an important function word is lost, and this example was fixed by using the source–target alignments. The second example shows an error caused by model sparsity.

<table>
<thead>
<tr>
<th>Input</th>
<th>(…) 小型個体 (…) の水揚げ量を (…) 15%以下に抑えることが勧告された。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>(...) it was recommended that (...) suppress fish catch of small individuals (...) to 0,15%.</td>
</tr>
<tr>
<td>Proposed</td>
<td>(...) it was recommended that (...) suppress fish catch of small individuals (...) 0,15%.</td>
</tr>
</tbody>
</table>

We found that using source–target alignments was effective in avoiding errors such as the first example in Table 4, however there remained some trickier cases where the alignment information was not sufficient, for example when function words were null or incorrectly aligned. The remainder errors were primarily caused by incorrect parsing and sparsity issues. The second example in Table 4 shows such a sparsity error, which could perhaps be fixed by normalizing numerical values.

In this paper we have shown that target-side syntax can be used effectively to improve the quality of scientific domain machine translation through the automatic post-editing of function words. We have presented an algorithm for inserting/deleting/Replacing function words based on a simple tree-based language model and demonstrated the effectiveness of using source–target alignments to improve accuracy. In the future we plan to extend the model to provide more robustness against parsing/alignment errors and experiment with other language pairs.

Acknowledgments

We would like to thank the reviewers for their detailed comments and suggestions. We are also grateful to Raj Dabre for his assistance in conducting the human evaluation.
References


Hongzheng Li  lihongzheng@mail.bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xniejekou Wai St., Haidian District, Beijing, 100875, China

Kai Zhao  765136570@qq.com
Institute of Chinese Information Processing, Beijing Normal University, 19, Xniejekou Wai St., Haidian District, Beijing, 100875, China

Renfen Hu  irishere@mail.bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xniejekou Wai St., Haidian District, Beijing, 100875, China

Yun Zhu  zhuyun@bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xniejekou Wai St., Haidian District, Beijing, 100875, China

Yaohong Jin  jinyaohong@bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xniejekou Wai St., Haidian District, Beijing, 100875, China

Abstract

This paper presents a novel hybrid system, which combines rule-based machine translation (RBMT) with phrase-based statistical machine translation (SMT), to translate Chinese patent texts into English. The hybrid architecture is basically guided by the RBMT engine which processes source language parsing and transformation, generating proper syntactic trees for the target language. In the generation stage, the SMT subsystem then provides lexical translation according to the defined structures and generates final translation. According to our empirical evaluation, the hybrid approach outperforms each individual system across a varied set of automatic translation evaluation metrics, verifying the effectiveness of the proposed method.

1. Introduction

As one of important applications in Natural Language Processing (NLP), Machine translation (MT) has developed several paradigms in past decades, basically including rule-based MT (RBMT) and statistic-based MT (SMT). Both the two approaches have strengths and weaknesses.

RBMT systems tend to produce better translations and deal with long distance dependencies, agreement and constituent reordering in a more principled way (Gorka et al., 2014), since they perform the analysis, transfer and generation steps based on syntactic principles. However, they usually have problems in word translation selection preferences, which usually have negative impacts on the translation quality. Also, in cases in which the input sentence has an unexpected syntactic structure, the parser may fail and the quality of the translation will decrease dramatically.
Contrary to RBMT, SMT models are more robust and usually better in fluent lexical selection since they exploit explicit probabilistic language models trained on very large corpora (Xuan et al., 2012). On the downside, SMT has difficulties in dealing with requirements of linguistic knowledge, such as syntactic functions and long distance word reordering, especially in the translation between distant language pairs such as Japanese and English (Isozaki et al., 2010), which may generate translations with improper even worse structures. While SMT has been recognized as the main stream approach of translation, RBMT has tended to be more effective for limited subject domains than SMT (List, 2012). As a result, hybrid MT (HMT) models have become increasingly popular in recent years, aiming to improve final translation effects and qualities. An typical example that can reflect the the increasing interest in hybrid approaches to MT is the workshop on hybrid approaches to translation (Hytra), which was first held at the EACL2012 conference, since then, it was continuously held in 2013 and 2014, in this year, it took part in conjunction with the ACL2015 conference held in Beijing.

It is well known that MT can be applied to various domains. With continued growth in the number of patent applications and the need of exchanging related information, patent domain MT has become one new application of MT, and attracted worldwide attentions of researchers and governments. In this article, we present a novel hybrid translation combination architecture that takes advantage of RBMT and phrase-based SMT to translate Chinese patent texts into English. As juridical and official documents, Chinese patent documents are usually featured by formal fixed expressions, and much longer sentences with more complex syntactic structures, compared with SMT, rule-based method is more suitable to describe the structures more precisely. Thus, our HMT system is constructed based on the RBMT system (Zhu and Jin, 2012). The main idea is that the RBMT guides main steps in performing source language parsing and transfer, generating proper transferred and reordered syntactic trees for the output, and the SMT system “Moses” (Koehn et al. 2007) then helps the lexical selection by providing more alternative translations according to the trees for target language generation. The final decoding also accounts for fluency by using language models. Since the structures of the translation are already decided by the RBMT subsystem, decoding of SMT will be more fast and efficient in turn.

We performed some experiments on the HMT system with several automatic evaluation metrics to test its performance. After comparing the HMT with individual RBMT and SMT systems, as well as Google online translation, the HMT outperformed all individual MT systems and gained much improvements in evaluation metrics, indicating the hybrid approach is indeed beneficial and effective for translation qualities.

The rest of the article is organized as follows. Section 2 overviews the related literatures on MT system combination and hybridization. Section 3 presents the individual systems and the architecture of system combination in detail. Section 4 describes the experimental work carried out with the hybrid architecture and discusses the obtained results. Finally, Section 6 concludes the work.

2. Related Work

This section mainly includes two parts: the first part overviews syntactic reordering in MT and the second part will discuss some previous work on MT system combination.

2.1. Syntactic Reordering

In SMT, reordering positions of chunks in source languages to generate proper and acceptable translation has been a hot issue, and syntactic reordering is effective in improving the performance of MT. Xia and McCord (2004) proposed an approach for French-English
translation by automatically extracting rewrite patterns after parsing the source and target sides of the training corpus. Collins et al., (2005) described a method for reordering German clauses in German-English translation. Some lexicalized reordering models (Tillman, 2004; Galley and Manning, 2008; Cherry et al., 2012) were employed to predict reordering by taking advantage of lexical information. Different with lexicalized models, a hierarchical phrase-based translation model (Chiang, 2007; Nguyen and Vogel, 2013) based on synchronous grammar was also used in reordering the chunks.

For Chinese-English MT, Wang et al., (2007) described a set of syntactic reordering rules that exploited systematic differences between Chinese and English word order and introduced a reordering approach. Zhang et al., (2007) described a sourceside reordering method based on syntactic chunks for phrase-based statistical machine translation. The source language sentences were first shallow parsed. Then, reordering rules were automatically learned from source-side chunks and word alignments. During translation, the rules were used to generate a reordering lattice for each sentence. Cao et al., (2014) proposed a novel lexicalized reordering model which is built directly on synchronous rules. For each target phrase contained in a rule, they calculated its orientation probability conditioned on the rule. Based on a set of dependency-based pre-ordering rules, Cai et al., (2014) presented a dependency-based pre-ordering approach for C-E MT, improved the BLEU score by 1.61 on the NIST 2006 evaluation data.

2.2. MT System Combination

System combination has been shown to improve classification performance in various tasks in the field of NLP (Rosti et al., 2007). Frederking and Nirenburg (1994) first applied system combination to MT. They integrated outputs of three different translation system (knowledge-based MT, example-based MT and a lexical transfer system) with Chart Walk Algorithm, then performed post-editing processing on the integrated outputs to generate final translation results. Bangalor et al. (2001) introduced recognizer output voting error (ROVER) (Fiscus, 1997) into MT, using a multiple string alignment (MSA) approach to align the hypotheses together, their experiments proved that integrated output was better than single system translation. Since then, system combination has aroused more attention around the world.

Confusion networks is one of the common methods used in combination strategies, which try to combine fragments from a number of different systems and use consensus network decoding to search for the best output from a list of n-best translations (Bangalore et al., 2001; Matusov et al., 2006; Chen et al., 2008; Ayan et al., 2008).

In most hybrid systems, the statistical components are usually selected as basic skeletons and in charge of the translation, correspondingly, the companion system provides complementary information. On the other hand, hybrid architectures where the RBMT system leads the translation and the SMT system provides complementary information to adjust the output from the RBMT, has been less explored. Such systems are applied to relative small domains (Simard et al., 2007), the output tends to be grammatical, and the main effect of the combination is an increase in lexical selection quality (Dugast et al., 2007). Following are some typical works led by RBMT in patent domain.

Jin (2010) proposed a hybrid approach which combined semantic analysis with rule-based method to translate Chinese patent to English. Alexandru et al., (2011) conducted some experiments on English–French patent domain adaptation of the MT systems used in the PLuTO project, both manual and automatic evaluations showed a slight preference for the hybrid system over the two individual baseline engines. Enache et al., (2012) also presented a system for English-French patent translation on the basis of large scale corpora with statistic method. Sheremetyeva (2013) discussed a Russian-English patent MT system which integrated
hybrid and rule-based components for several complementary levels of output. There also exists some hybrid systems participating the patent evaluation workshop of the NTCIR conference held in Japan (Isao et al., 2013).

Unlike many previous works using SMT system as the basic skeleton and adapting confusion networks, our hybrid system, oriented for patent domain, is constructed based on the RBMT system and does not involve a confusion network. In the RBMT, we build a considerable knowledge base and manually write rules to help the system analyse and reorder the source sentences according to the grammatical expressions in target language, and the SMT is responsible for the target words selection. As a result, the hybrid system can guarantee both proper syntactic structures and lexical selection qualities that are consistent with target language. The hybrid system will be clarified in detail in following sections.

3. System Architecture

Before presenting the system combination architecture, we need to first introduce the individual RBMT and SMT.

3.1. RBMT

The RBMT engine is based on the traditional translation model which is mainly divided in three steps: (i) analysis of the source language into syntactic-semantic tree structures, (ii) transfer and transformation from source language to target language, and (iii) generation of the target language. It is well known that rule-based approach is featured by the knowledge base and rules which can describe linguistic information. In the system, we have built a considerable knowledge base with more than 50,000 words which cover most patent texts. In the knowledge base, the words are annotated with various syntactic and semantic information. We also manually wrote numerous formal targeted rules to help the engine process the sentences in each step. These rules provide a hierarchical parsing and reordering access to deal with various structures and chunks in source sentences. By using the information both in the knowledge base and the rules, the MT system will finish the processing of three steps. We will describe each of the stages in the following.

Analysis of Source Language

As mentioned, sentences in patent texts are usually much longer. A sentence (S) ended with a full stop may include several sub-sentences (marked as SS) and chunks separated by punctuations (marked as SST, most are commas, colons and semicolons also included). That is, \( S = SS_1, SS_2, \ldots, SS_n \). Considering the expression features of patent documents, a parser is specially developed for the patent texts and integrated into the translation engine, aiming to regards a whole long sentence as the basic processing unit.

The analysis is conducted in three syntactic levels: first, the sentence is separated into several SS according to the punctuations; then, parsing each SS into chunks served as direct componts of SS, including identifying the subject, predicate verb, object and adverbial etc.; last, further analyse the chunks into terminals (leaf nodes on syntactic trees). Thus, S is the root of sentence, and it has several SS nodes and separators SST, then each SS node is composed of several chunks such as NP, VP, and adverbial phrases (ADVP) etc, further, chunks are composed of terminal nodes.

In the first level, the main purpose is to divide the complex long sentence into several SS mainly by commas. But not all commas can separate the sentence, because some of them may follow by phrases. In our research, we determine that, for commas following NP and ADVPs, they cannot separate the sentence, and they will be marked as DBT.
In the subsentence level, the system parses the subsentences to get the syntactic components. Parsing rules basically include rules for indentifying predicate verb, ADVP, special NP with long modifiers, and prepositional phrases (PPs) introduced by unique prepositions (such as “把BA” “被BEI” “将JIANG”, etc.) in Chinese. We want to mainly discuss about identification of predicate verbs, which play more important role in parsing.

As Chinese lacks necessary lexical changes, when several verbs appear in the same sentence, it is usually more difficult to identify the proper core verb. During the beginning stage of the identification process, we first write some rules according to the context information to exclude some verbs that cannot be selected as core verb, next, we then design various high and low weights for remaining possible verbs, the weights represent the possibilities that verbs serve as predicate. When matching kinds of rules, the verbs will be added with corresponding weights, as a result, after comparing the weights, the verbs with the highest weight will be selected as final predicate verb.

The final stage is chunk-level parsing. In this level, the system will continue to analyse the non-terminal chunks into leaf nodes. Since some chunks may contain complex and nested structures, the system needs to exploit the rules in a circular manner and perform the syntactic analysis hierarchically until each node is parsed.

After finishing parsing, the system will generate a syntactic tree for each source sentence, and each node on the tree possesses several marks and symbols representing various syntactic and semantic information.

Here is an example sentence in patent texts, including two subsentences, which followed by the syntactic tree.

E.g.1: Source sentence: 在上述结构中,单电池由突起部支撑,因此可以提高耐振动性。

Target sentence: In above structure, the single cell is supported by the protrusions, therefore the vibration resistance can be improved.

![Syntactic Tree of the Example Sentence](image)

Figure 1. Syntactic Tree of the Example Sentence

Fig.2 shows the form of the tree of the example sentence in our MT system. It can be seen from the vertical tree, there exists some orange circles under each subsentences, and each of them represents the direct component chunk of the subsentence, symbols in the angle brackets “< >” indicate syntactic information of the chunks and punctuations, and when click the “+” button in front of some circles, the chunks will extend and show the detailed information of each terminal under the chunk.
Transfer from Source Language to Target Language

When transforming Chinese sentences into English, it is necessary to reorder and transfer chunks and words to guarantee fluent and proper expressions. Corresponding to the analysis phases above, the transformation process also includes three levels in a top-down order: ① transformation of relationships between subsentences, ② structural reordering of chunks in subsentences and ③ reordering inside the chunks. Basic operations in the reordering include add or delete nodes, chunks position adjustment etc. In the following, we will introduce the three stages with some rules and examples.

Transformation between subsentences mainly refers to transferring the source sentences into expressions commonly used in target language according to the semantic relationships between subsentences.

Rule1: \{SS1&CHN[在于,包括]&END%\}+ CHN[，]+ SS2 \rightarrow SS1+ that + SS2

The rule means that, if the Chinese characters(CHN) such as “在于(lie in), 包括(include)” appear in the end of the first SS1, and followed by the comma and another SS2, then the comma will be replaced by the word “that” when transferred into English.

E.g.2: Source sentence: 本发明的特征在于,它可以调节输出装置的参数。

Target sentence: The feature of this invention lies in that it can adjust the parameter of the output device.

In the example, the source sentence includes two subsentences, in which the second one is actually the object of the first one. Considering that, when transformed into English, it is better to transfer the two subsentences into a single sentence by using object clause and replacing the comma with the connecting word “that”.

Generally, Chunk-level reordering and transformation inside the chunks play more important roles in generating grammatical target language. Which mainly includes following types:

Changing form, tense and voice etc. of core verbs. As for form, for example, if some modal verbs appear before the core verbs, the verbs should be transformed in the form of prototype. As for tense, simple present is considered as default tense in most cases, but if some
words such as “已, 已经”(have already)” appear before the core verbs, then they need to be changed into the perfect tense. As for voice, the default voice is active voice, but verbs should be changed into passive voice if they are followed by words representing passive voice such as the typical preposition “被(BEI)”. On the other hand, for some sentences without subjects, the predicates may also be changed into passive voice, and the objects will serve as subject in English (\(VP+NP \rightarrow NP+VP\) (passive voice)).

**Transforming the ADVPs introduced by prepositions.** Such transformation includes two aspects: (1) reordering positions of adverbials. For those located between subject and predicate verb in Chinese, they need to be reordered to the end of the sentence in English. If some parallel adverbials appear in the same sentence, it is better to reorder them in reverse order. (2) Transformation of long-distance fixed structures. In some adverbial chunks, “当……时”(when……) and “在……中”(in……), for example, the left and right boundary words are usually collations and appear together, which can be directly replaced with corresponding words in English. We have wrote rules to cover the fixed collations as much as possible.

Rule2: \(NP+ADVP+VP \rightarrow NP+VP+ADVP\)

Rule3: \(NP+ADVP1+ADVP2+VP \rightarrow NP+VP+ADVP2+ADVP1\)

Rule4: \((0)CHN[当]+(f)CHN(时)] \rightarrow DELETE (0)+DELETE (f)+ADD[when]\)

E.g.3: Source sentence: [NP本发明][ADVP1在实验中][ADVP2通过一种有效的方法][VP提高产品的性能]。

Target sentence: [NP This invention][VP improves the performance of the products][ADVP2 by an effective method][ADVP1 in the experiment].

**Reordering special prepositional phrases (PPs) in Chinese.** Some prepositions, such as “把(BA)” “将(JIANG)” and “被(BEI)” etc., are unique in Chinese and lack corresponding translation in English. PPs composed of such prepositions and NPs always appear in front of VP, when transfer them into English, these prepositions should be deleted, and NPs behind the prepositions must be reordered to proper positions, usually after the VP.

Rule5: \(NP1 + prep. + NP2 + VP \rightarrow NP1 + VP + NP2\)

E.g.4: Source sentence:这些计数器对这些数据输入/输出装置的数量进行计数。

Target sentence: These counters count the number of these data input/output devices.

<table>
<thead>
<tr>
<th>Before syntactic reordering</th>
<th>After syntactic reordering</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP1</td>
<td>NP1</td>
</tr>
<tr>
<td>这些计数器(These counters)</td>
<td>这些计数器(These counters)</td>
</tr>
<tr>
<td>PP</td>
<td>PP</td>
</tr>
<tr>
<td>对(DUI)</td>
<td>进行计数 (count)</td>
</tr>
<tr>
<td>NP2</td>
<td>NP2</td>
</tr>
<tr>
<td>这些数据输入/输出装置的数量</td>
<td>这些数据输入/输出装置的数量</td>
</tr>
<tr>
<td>(the number of these data input/output devices)</td>
<td>(the number of these data input/output devices)</td>
</tr>
<tr>
<td>VP</td>
<td>VP</td>
</tr>
<tr>
<td>进行计数 (count)</td>
<td>进行计数 (count)</td>
</tr>
</tbody>
</table>

Figure 3. Original and Reordered Trees of Example 4

Rule6: \(NP1 + prep. + NP2 + VP + NP3 \rightarrow NP1 + VP + NP2 + NP3\)

E.g.5: Source sentence:第二通信模块将第二表示数据发送到计算机系统。

Target sentence: The second communication module sends the second indicating data to the computer system.
Structural reordering inside chunks, especially in NPs. The most common structure of NPs in Chinese is “modifiers + 的(DE) + head NP”. In which the modifiers can include NP, VP, quantifier phrase (QP), determiner phrase (DP), adjective phrase (ADJP) or even relative clauses. The placement of QP, DP, and ADJP modifiers is somewhat similar to English that these phrases typically occur before the nouns they modify, and they need not reordering.

For NP1+DE+NP2, although it is analogous to the English possessive structure of “NP1’s NP2” and does not require reordering, the Chinese possessive structure “NP1 DE NP2” can express more sophisticated relationships, additionally, the “NP2 of NP1” expression is more general and can replace “NP1’s NP2” in many cases, except for the case that the NP1 is a pronoun. Thus, the reordering rules will state that and map the following rule.

Rule7: NP1+DE+NP2 → NP2+DE+NP1

NPs modified by relative clauses (CP), with long distance structures, are quite different with those in English. For such NPs, we apply the rules to reposition the child CP after its sibling head NP under a parent NP.

Rule8: CP+DE + NP → NP+that+CP

E.g.6: Source sentence: 将在性能测试中产生的电能传输给车辆的装置。

Target sentence: The device that transmits the electrical energy produced in the performance test to the vehicle.

From the syntactic trees, it can be seen that the CP modifier is reordered to the position just after its sibling head NP, and the whole NP is transformed into a NP modified by an attributive clause. In the transformation, it is necessary to add an additional word “that” between the antecedent and the clause.

The example is also a nested chunk with multi-level structures, it clearly outlines the various types needed to be transformed, including ADVP, PP, NP and VP mentioned above. Reordering rules sequentially process the elements in a top-down order, and the rules will be exploited circularly.
Generation of Target Language

Generation can be decomposed into two steps. First, word selection. According to the reordered syntactic transformation tree, the MT engine selects target words for each node from the Chinese-English parallel translation dictionary. Second, morphological generation, which consists of generating the target surface forms from their associated morphological information.

3.2. SMT System

The phrase-based SMT baseline system Moses is built on the basis of freely available state-of-the-art tools: the GIZA++ toolkit (Och 2003) to estimate word alignments, the IRST Language Modelling toolkit (IRSTLM) (Federico et al., 2008) with modified Kneser-Ney smoothing (Chen and Goodman 1999) to guarantee more fluent target language outputs. And in the paper, we use the IRSTLM toolkit to train a 5-gram language model with the patent texts corpus. Last, as decoding is the central stage of SMT, the Moses decoder (Koehn et al. 2007) is employed to find the highest scoring sentence in the target language corresponding to given source sentence.

3.3. Hybrid System Architecture

Many previous works use SMT as basic skeleton of the hybrid system. In our work, considering the pros and cons of RBMT and SMT, as well as the special features of patent texts, we try to build a hybrid patent MT system guided by the RBMT. Just as mentioned before, RBMT usually performs better in dealing with long distance structure and reordering. In the system combination, the RBMT is responsible for parsing source language and generate grammatical syntactic reordering lattice for the target language by applying the knowledge base and the rules, the main task of SMT is to generate translation for each node according to the reordered tree determined by the RBMT. While the RBMT guarantees basic proper structures of target language, the SMT provide more lexical selection, as a result, the hybrid system is supposed to generate more fluent and acceptable translation.

In the hybrid system, after word segmentation, the RBMT first analyses Chinese sentences and transform positions of chunks according to the corresponding expressions in English by matching kinds of rules. Next, instead of generating all the translation for the words, the RBMT just generates partial translation for special words (most are functional words) in the sentences. On the other hand, the RBMT also adds some connecting words such as “that” to make the final translation more fluent. Let’s take a sentence for example.
4. Experiments

In this part, we conducted some evaluations on the hybrid system to test its performances, which were measured by several popular automatic evaluation metrics: WER (Nielsen et al., 2000), PER (Tillmann et al., 1997), TER (Snover et al., 2006), BLEU (Papineni et al., 2002), NIST (Doddington, 2002), GTM (Melamed et al., 2003), METEOR (Banerjee and Lavie., 2005) and ROUGH (Lin and Och, 2004). All measures were calculated with the Asiya toolkit for MT evaluation (Giménez and Márquez, 2010). We also compared the scores of the hybrid system with those of RMBT, SMT and Google online translation.

4.1. Experimental Setting

We exploited the training set of NTCIR-9 (Sakai and Joho, 2011), including 1 million bilingual Chinese-English patent sentence pairs, to train the Moses decoder. In the develop set of NTCIR-
which included 2000 sentence pairs, 1000 sentence pairs were randomly extracted as development set, and remaining 1000 pairs as test set. Following is the statistic data of the experiments.

<table>
<thead>
<tr>
<th>sentence pairs</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>1 million: 188MB (Chinese), 227MB (English)</td>
</tr>
<tr>
<td>Development Set</td>
<td>1000: 188KB (Chinese), 212KB (English)</td>
</tr>
<tr>
<td>Test Set</td>
<td>1000: 191KB (Chinese), 214KB (English)</td>
</tr>
</tbody>
</table>

Table 1. statistic data of the experiments

4.2. Experimental results

Table 2 gave the comparative evaluation scores of the four MT systems.

<table>
<thead>
<tr>
<th>WER</th>
<th>PER</th>
<th>TER</th>
<th>ROUGE</th>
<th>BLEU-4</th>
<th>NIST-5</th>
<th>METEOR</th>
<th>GTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT</td>
<td>93.08</td>
<td>71.45</td>
<td>87.55</td>
<td>40.19</td>
<td>11.44</td>
<td>4.45</td>
<td>19.03</td>
</tr>
<tr>
<td>SMT</td>
<td>61.33</td>
<td>32.62</td>
<td>53.44</td>
<td>25.30</td>
<td>7.57</td>
<td>3.47</td>
<td>22.58</td>
</tr>
<tr>
<td>HMT</td>
<td>59.41</td>
<td>30.63</td>
<td>51.68</td>
<td>56.46</td>
<td>31.22</td>
<td>7.84</td>
<td>35.16</td>
</tr>
<tr>
<td>Google</td>
<td>68.12</td>
<td>45.99</td>
<td>61.05</td>
<td>59.66</td>
<td>37.36</td>
<td>8.80</td>
<td>27.96</td>
</tr>
</tbody>
</table>

Table 2. Comparison of Several Systems

4.3. Analysis

As clearly shown in table 2, the hybrid system outweighted other two individual systems in all the evaluation metrics. The experimental data have proved that the proposed method performed well in improving the translation results significantly.

After analysing the results of each system, we can come to some conclusions. Take the BLEU-4 scores for example, compared with SMT and HMT, the score of RBMT is quite lower. Several reasons are supposed to account for the result: (1) some errors occurred in the word segmentation process and ambiguous structures resulted in improper or even wrong syntactic parsing, further affected the transformation and generation of target language. (2) some bugs of the system also made the final translation worse. (3) last, each node in Chinese usually has only one corresponding English translation in the bilingual dictionary, and the lexical selection is quite limited, so the candidate target words are more likely different with the reference translation.

The large difference in BLEU scores between RBMT(11.44) and SMT(29.30) has reflected a well known phenomenon of automatic lexical-matching evaluation metrics overestimating the quality of statistical systems on in-domain test sets. (Giménez and Márquez, 2007). Despite the difference, the hybrid system is still able to take advantage of the combination and consistently improve results over the individual SMT system.

Note that, although the score is low, the syntactic structures of many target sentences generated by the RBMT are good and grammatical, especially some complex structures with long distance. which indicates the rule-based approach is more easily to describe and consider the linguistic information. Following is a comparison of a sentence transformed by the RBMT and Google.

<table>
<thead>
<tr>
<th>Source sentence</th>
<th>在活塞缸52与基底构件48b之间设置第二弹簧58，而另外在第二弹簧58与基底构件58b之间设置一个或一组垫片60。</th>
</tr>
</thead>
</table>

Source sentence: 在活塞缸52与基底构件48b之间设置第二弹簧58，而另外在第二弹簧58与基底构件58b之间设置一个或一组垫片60。
A second spring 58 is positioned between the piston cylinder 52 and the base member 48b, and a shim 60 or series of shims 60 is further positioned between the second spring 58 and the base member 58b.

Second spring 58 is arranged between the piston jar 52 and the basement member 48b, the other one or one group gasket 60 is arranged between second spring 58 and the basement member 58b.

Disposed between the base member 52 and the piston cylinder 48b of the second spring 58, while the other set of one or a group of the spacer 60, between the second spring 58 and the base member 58b.

<table>
<thead>
<tr>
<th>Reference</th>
<th>A second spring 58 is positioned between the piston cylinder 52 and the base member 48b, and a shim 60 or series of shims 60 is further positioned between the second spring 58 and the base member 58b.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT</td>
<td>Second spring 58 is arranged between the piston jar 52 and the basement member 48b, the other one or one group gasket 60 is arranged between second spring 58 and the basement member 58b.</td>
</tr>
<tr>
<td>Google</td>
<td>Disposed between the base member 52 and the piston cylinder 48b of the second spring 58, while the other set of one or a group of the spacer 60, between the second spring 58 and the base member 58b.</td>
</tr>
</tbody>
</table>

Table 3. Comparison of an Example Translated by the RBMT and Google

Since both the two subsentences lack subjects, thus it is better to transfer the predicate verbs into passive voice, which will be more fluent and be consistent with English expression. The RBMT has transferred successfully and generate fluent translation. On the contrary, as for Google, the translation is quite bad and unacceptable. The orders of chunks are actually ungrammatical, and, what's worse, the core verb "设置(dispose)" in the second subsentence is even translated into a noun(set), resulting in the absence of predicate verb.

Compared with Google, the BLEU score of the HMT is lower than 5 percent. As a typical representative of SMT approaches, Google has always possessed powerful and state-of-the-art algorithms on NLP and language technology, we guess it’s natural for Google to gain a better score. However, scores of error rates (WER, PER and TER) and METEOR metrics were all better than Google. The main reason is that, the syntactic reordering structures of target language generated by the RBMT benefits the HMT to a large extent, on the contrary, Google tends to have much difficulties in reordering complex chunks, especially for those with long distances and dependency.

5. Conclusion

This paper presents a hybrid MT system, which combines an individual RBMT and phrase-based SMT system, for Chinese-English patent translation. Since RBMT is generally able to produce grammatically better translations, different with many previous system combinations mainly guided by SMT, our HMT is built based on the RBMT, its analysis and transfer modules are exploited to generate the backbone of the translation and provide syntactic reordering structures of target language. In the generation stage, statistical decoding of SMT-based translation generate translation for the source sentences, which can improve lexical selection and fluency of the final translation.

We conducted experiments to evaluate the system on patent texts with several evaluation metrics, the hybrid system outperformed both the RBMT and SMT, indicating the proposed approach is a good choice and efficient in improving final translation performance. The experimental results also confirmed that syntactic reordering provided by the RBMT is essential.

The work still sleaves some issues that deserve further research. In the future, we need to expand the size of patent corpus to design more proper processing rules and improve the RBMT system to produce better transferred and reordering structures. On the other hand, we would like to train the SMT models with larger data set and optimize the related parameters.
Acknowledgement

This work was supported by the National Hi-Tech Research and Development Program of China (2012AA011104), and we also want to express our sincere thanks to the anonymous reviewers of this article.

References


Collecting Bilingual Technical Terms from Patent Families of Character-Segmented Chinese Sentences and Morpheme-Segmented Japanese Sentences

Zi Long
Takehito Utsuro
Grad. Sch. Sys. & Inf. Eng., University of Tsukuba, Tsukuba, 305-8573, Japan

Tomoharu Mitsuhashi
Japan Patent Information Organization, 4-1-7, Tokyo, Koto-ku, Tokyo, 135-0016, Japan

Mikio Yamamoto
Grad. Sch. Sys. & Inf. Eng., University of Tsukuba, Tsukuba, 305-8573, Japan

Abstract
In manual translation of patent documents, a technical term bilingual lexicon is inevitable for a translator to efficiently translate patent documents. Dong et al. (2015) proposed a method of generating bilingual technical term lexicon from morpheme-segmented parallel patent sentences. The proposed method estimates Japanese-Chinese translation of technical terms using the phrase translation table of a statistical machine translation model. The procedure of generating bilingual technical term lexicon consists of the following four steps: (1) extracting Japanese technical terms from Japanese side of parallel patent sentences, (2) collecting all the sentences that contain the extracted Japanese term, (3) generating Chinese translation of the Japanese technical term referring to the phrase translation table of a statistical machine translation model, and (4) applying the Support Vector Machines (SVMs) to the task of identifying bilingual technical terms. In this paper, we segment the Chinese sentences into characters instead of segmenting them into morphemes as in Dong et al. (2015), and represent Japanese-Chinese patent families in terms of character-segmented Chinese sentences and morpheme-segmented Japanese sentences. Then, to those Japanese-Chinese patent families, we apply the framework (Dong et al., 2015) of identifying bilingual technical terms. As a result, we achieve the performance of over 90% precision with the condition of more than or equal to 60% recall.

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 natural language 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; Bouamor et al., 2013; Morin and Hazem,
Figure 1: The Procedure of Translation Estimation of a Technical Term using a Phrase Translation Table and a Parallel Sentence Pair

<table>
<thead>
<tr>
<th>Japanese Phrase</th>
<th>Chinese Phrase</th>
<th>Translation Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>冷延鋼板</td>
<td>冷轧钢板</td>
<td>0.77</td>
</tr>
<tr>
<td>冷延鋼板</td>
<td>冷轧钢带</td>
<td>0.02</td>
</tr>
<tr>
<td>冷延鋼板</td>
<td>冷钢带</td>
<td>0.01</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Among those efforts of acquiring bilingual lexicon from text, Dong et al. (2015) proposed to acquire Japanese-Chinese technical term translation lexicon from the phrase translation tables, which are trained by a phrase-based statistical machine translation (SMT) model with parallel sentences automatically extracted from patent families. One of the major advantages of the proposed approach is that the resource we utilize in this approach is Japanese-Chinese patent families, which continue to be published every year. The procedure of generating bilingual technical term lexicon consists of the following four steps: (1) extracting Japanese technical terms from Japanese side of parallel patent sentences, (2) collecting all the sentences that contain the extracted Japanese term, (3) generating Chinese translation of the Japanese technical term referring to the phrase translation table of a statistical machine translation model, and (4) applying the Support Vector Machines (SVMs) (Vapnik, 1998) to the task of identifying bilingual technical terms. In this paper, we segment the Chinese sentences into characters instead of segmenting them into morphemes as in (Dong et al., 2015), and represent Japanese-Chinese patent families in terms of character-segmented Chinese sentences and morpheme-segmented Japanese sentences. Then, to those Japanese-Chinese patent families, we apply the framework (Dong et al., 2015) of identifying bilingual technical terms. As a result, we achieve the performance of over 90% precision with the condition of more than or equal to 60% recall.
2 Japanese-Chinese Parallel Patent Documents

Japanese-Chinese parallel patent documents are collected from the Japanese patent documents published by the Japanese Patent Office (JPO) in 2004-2012 and the Chinese patent documents published by State Intellectual Property Office of the People’s Republic of China (SIPO) in 2005-2010. From them, we extract 312,492 patent families, and the method of Utiyama and Isahara (2007) is applied to the text of those patent families, and Japanese and Chinese sentences are aligned. In this paper, we use 3.6M parallel patent sentences with the highest scores of sentence alignment.

3 Phrase Translation Table of an SMT Model

As a toolkit of a phrase-based SMT model, we use Moses Koehn et al. (2007) and apply it to the whole 3.6M parallel patent sentences. Before applying Moses, Japanese sentences are segmented into a sequence of morphemes by the Japanese morphological analyzer MeCab with the morpheme lexicon IPAdic, while Chinese sentences are segmented by characters. In this procedure of Chinese segmentation, a consecutive sequence of numbers as well as a consecutive sequence of alphabetical characters are segmented into a token.

As the result of applying Moses, we have a phrase translation table in the direction of Japanese to Chinese translation, consisting of 268M translation pairs with 194M Japanese phrases with Japanese to Chinese phrase translation probabilities \( P(p_C \mid p_J) \) of translating a Japanese phrase \( p_J \) into a Chinese phrase \( p_C \). For each Japanese phrase, those multiple translation candidates in the phrase translation table are ranked in descending order of Japanese to Chinese phrase translation probabilities.

4 Translation Estimation of a Technical Term using a Phrase Translation Table and a Parallel Sentence Pair

Figure 1 shows the procedure of estimating Chinese translation of a Japanese technical term using a phrase translation table and a parallel sentence pair. The phrase translation table is first referred to when identifying a bilingual technical term pair, given a parallel sentence pair \( \langle S_J, S_C \rangle \) and a Japanese technical term \( t_J \). Given a parallel sentence pair \( \langle S_J, S_C \rangle \) containing a Japanese technical term \( t_J \), Chinese translation candidates collected from the phrase translation table are matched against the Chinese sentence \( S_C \) of the parallel sentence pair. Among those found in \( S_C \), \( \hat{t}_C \) with the largest translation probability \( P(t_C \mid t_J) \) is selected and the bilingual technical term pair \( \langle t_J, \hat{t}_C \rangle \) is identified.

5 Translation Estimation by SVM using Features extracted from Multiple Parallel Patent Sentences

5.1 Selecting Japanese Technical Terms for Evaluation

When selecting Japanese technical terms for evaluation, we first extract 1.2M noun phrases from the 3.6M parallel patent sentences. Next, we divide the set of all the Japanese noun phrases
<table>
<thead>
<tr>
<th>$j_{fe}$</th>
<th>$jf=1$</th>
<th>2 ≤ $j_{fe}$ ≤ 5</th>
<th>5 ≤ $j_{fe}$ ≤ 10</th>
<th>10 ≤ $j_{fe}$ ≤ 15</th>
<th>15 ≤ $j_{fe}$ ≤ 20</th>
<th>20 ≤ $j_{fe}$ ≤ 30</th>
<th>30 ≤ $j_{fe}$ ≤ 50</th>
<th>50 ≤ $j_{fe}$ ≤ 100</th>
<th>100 ≤ $j_{fe}$ ≤ 500</th>
<th>500 ≤ $j_{fe}$ ≤ 1,000</th>
<th>1,001 ≤ $j_{fe}$ ≤ 10,000</th>
<th>≥10,000</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ≤ $j_{ref}$ ≤ 50</td>
<td>37.62</td>
<td>26.26</td>
<td>12.19</td>
<td>20.33</td>
<td>21.57</td>
<td>11.20</td>
<td>24.40</td>
<td>19.49</td>
<td>25.70</td>
<td>22.76</td>
<td>36.121</td>
<td>8.09</td>
<td>1.28</td>
</tr>
<tr>
<td></td>
<td>59.7%</td>
<td>77.0%</td>
<td>63.2%</td>
<td>60.7%</td>
<td>56.8%</td>
<td>55.1%</td>
<td>50.0%</td>
<td>38.8%</td>
<td>35.4%</td>
<td>29.6%</td>
<td>29.8%</td>
<td>11.6%</td>
<td>3.6%</td>
</tr>
<tr>
<td>5 ≤ $j_{ref}$ ≤ 10</td>
<td>55.58</td>
<td>24.28</td>
<td>12.15</td>
<td>16.24</td>
<td>14.23</td>
<td>22.32</td>
<td>21.42</td>
<td>28.55</td>
<td>23.55</td>
<td>22.86</td>
<td>24.79</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-94.9%</td>
<td>-85.8%</td>
<td>-80.0%</td>
<td>-66.7%</td>
<td>-60.9%</td>
<td>-68.2%</td>
<td>-50.0%</td>
<td>-41.9%</td>
<td>-25.6%</td>
<td>-43.1%</td>
<td>-3.6%</td>
<td>272.525</td>
<td></td>
</tr>
<tr>
<td>10 ≤ $j_{ref}$ ≤ 15</td>
<td>41.43</td>
<td>16.16</td>
<td>14.16</td>
<td>16.18</td>
<td>9.15</td>
<td>15.25</td>
<td>12.19</td>
<td>23.39</td>
<td>14.51</td>
<td>3.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-95.4%</td>
<td>-100.0%</td>
<td>-87.5%</td>
<td>-87.5%</td>
<td>-88.9%</td>
<td>-60.0%</td>
<td>-63.2%</td>
<td>-59.0%</td>
<td>-45.2%</td>
<td>-5.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 ≤ $j_{ref}$ ≤ 20</td>
<td>38.40</td>
<td>13.14</td>
<td>9.9</td>
<td>6.7</td>
<td>4.8</td>
<td>8.8</td>
<td>10.15</td>
<td>7.10</td>
<td>6.17</td>
<td>1.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-95.0%</td>
<td>-100.0%</td>
<td>-85.8%</td>
<td>-66.7%</td>
<td>-100.0%</td>
<td>-66.7%</td>
<td>-70.0%</td>
<td>-35.3%</td>
<td>-16.7%</td>
<td>-77.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 ≤ $j_{ref}$ ≤ 25</td>
<td>31.51</td>
<td>7.7</td>
<td>8.8</td>
<td>1.2</td>
<td>7.8</td>
<td>4.6</td>
<td>8.9</td>
<td>7.11</td>
<td>1.4</td>
<td>74.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-50.0%</td>
<td>-87.5%</td>
<td>-66.7%</td>
<td>-88.9%</td>
<td>-63.7%</td>
<td>-25.1%</td>
<td>-86.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 ≤ $j_{ref}$ ≤ 30</td>
<td>32.32</td>
<td>6.8</td>
<td>3.4</td>
<td>6.10</td>
<td>5.7</td>
<td>6.9</td>
<td>2.8</td>
<td>1.5</td>
<td>61.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-100.0%</td>
<td>-75.0%</td>
<td>-75.0%</td>
<td>-60.0%</td>
<td>-71.5%</td>
<td>-66.7%</td>
<td>-25.1%</td>
<td>-20.1%</td>
<td>-73.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 ≤ $j_{ref}$ ≤ 50</td>
<td>25.26</td>
<td>15.17</td>
<td>7.9</td>
<td>7.11</td>
<td>8.10</td>
<td>5.5</td>
<td>3.13</td>
<td>68.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-96.2%</td>
<td>-83.3%</td>
<td>-77.8%</td>
<td>-63.7%</td>
<td>-80.0%</td>
<td>-60.0%</td>
<td>-23.1%</td>
<td>-74.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 ≤ $j_{ref}$ ≤ 100</td>
<td>24.25</td>
<td>21.21</td>
<td>7.9</td>
<td>8.8</td>
<td>6.6</td>
<td>4.14</td>
<td>76.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-96.0%</td>
<td>-100.0%</td>
<td>-77.8%</td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-84.4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 ≤ $j_{ref}$ ≤ 200</td>
<td>25.25</td>
<td>17.18</td>
<td>5.6</td>
<td>7.8</td>
<td>4.8</td>
<td>58.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-100.0%</td>
<td>-94.5%</td>
<td>-83.4%</td>
<td>-87.5%</td>
<td>-50.0%</td>
<td>-89.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 ≤ $j_{ref}$ ≤ 500</td>
<td>23.24</td>
<td>10.10</td>
<td>6.6</td>
<td>7.8</td>
<td>4.3</td>
<td>42.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-95.9%</td>
<td>-100.0%</td>
<td>-75.0%</td>
<td>-95.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500 ≤ $j_{ref}$ ≤ 1,000</td>
<td>31.31</td>
<td>5.5</td>
<td>1.1</td>
<td>37.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-100.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,001 ≤ $j_{ref}$ ≤ 10,000</td>
<td>24.24</td>
<td>6.7</td>
<td>30.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-100.0%</td>
<td>-83.8%</td>
<td>-96.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥10,000</td>
<td>10.10</td>
<td>10.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-100.0%</td>
<td>-100.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Rates of Positive Examples (# of positive examples / # of positive and negative examples) for Each Frequency Range of Japanese Technical Term Frequency ($j_{fe}$) and Japanese-Chinese Co-occurrence Frequency ($j_{ref}$).
Table 2: Numbers of Positive / Negative Examples in the Reference Set

<table>
<thead>
<tr>
<th></th>
<th>positive</th>
<th>negative</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,255</td>
<td>837</td>
<td>2,092</td>
</tr>
</tbody>
</table>

Table 3: Positive Examples for Low / Middle / High Frequency Range of Japanese Technical Terms Frequency \((jf)\) and Japanese-Chinese Co-occurrence Frequency \((jcf)\)

<table>
<thead>
<tr>
<th>Frequency Range</th>
<th>Positive Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low frequency range</td>
<td>&lt;電動スピーカー型発電装置&gt; (electric power generating apparatus)</td>
</tr>
<tr>
<td>Middle frequency range</td>
<td>&lt;スクリーン装置&gt; (screen device)</td>
</tr>
<tr>
<td>High frequency range</td>
<td>&lt;反応混合物&gt; (reactant mixture)</td>
</tr>
</tbody>
</table>

5.2 Developing a Reference Set of Bilingual Technical Terms

For each \(t_J\) of the 578 Japanese technical terms selected in the previous section, we first collect all the parallel sentence pairs \(\langle S_{iJ}, S_{iC} \rangle (i = 1, 2, \ldots, n)\) containing the given Japanese technical term \(t_J\). From each of those parallel sentence pairs \(\langle S_{iJ}, S_{iC} \rangle (i = 1, 2, \ldots, n)\), at most one bilingual technical term pair \(\langle t_J, \hat{t}_C \rangle\) is extracted according to the translation estimation procedure presented in Section 4. Thus, for an input Japanese technical term \(t_J\), we obtain zero or more Japanese-Chinese technical term translation pairs \(\langle t_J, \hat{t}_C \rangle (j = 1, 2, \ldots, m (\leq n))\). In total, for the 578 Japanese technical terms selected in the previous section, we obtain 2,092 candidates of Japanese-Chinese technical term translation pairs. For each of the 2,092 candidates of technical term translation pairs, we manually judge whether it is correct technical term translation pair or not. Finally, as shown in Table 2, we obtain 1,255 correct translation pairs as positive examples, and the remaining 837 erroneous ones as negative examples. We use the prefixes or ending with certain suffixes are not appropriate as Japanese technical terms and are excluded. Those which include symbols or numbers are also excluded.

7 Those 578 Japanese technical terms for evaluation are exactly the same as those used in (Dong et al., 2015).
8 The Japanese-Chinese phrase translation table that is applied in the procedure of translation estimation is trained with patent families consisting of character-segmented Chinese sentences and morpheme-segmented Japanese sentences, and is different from that used in (Dong et al., 2015). As a result, we obtained a set of candidates of Japanese-Chinese technical terms.
Table 4: Features for Identifying Bilingual Technical Terms by SVM

<table>
<thead>
<tr>
<th>class</th>
<th>feature</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>monolingual</td>
<td>( f_1 ): frequency of Japanese term</td>
<td>the ID (1~13) of the frequency range of the Japanese technical term</td>
</tr>
<tr>
<td></td>
<td>( f_2 ): frequency of Chinese term</td>
<td>the ID (1~13) of the frequency range of the Chinese technical term</td>
</tr>
<tr>
<td>bilingual</td>
<td>( f_3 ): translation probability</td>
<td>the translation probability ( P(t_C</td>
</tr>
<tr>
<td></td>
<td>( f_4 ): rank of Chinese translation candidates (descending order)</td>
<td>the rank of ( t_C ) with respect to the descending order of the conditional translation probability ( P(t_C</td>
</tr>
<tr>
<td></td>
<td>( f_5 ): co-occurrence frequency of bilingual technical term pairs</td>
<td>the ID (1~13) of the co-occurrence frequency range of the Japanese-Chinese technical term pairs</td>
</tr>
<tr>
<td></td>
<td>( f_6 ): difference of the frequency of Japanese technical term and the co-occurrence frequency of bilingual technical term pairs</td>
<td>returns 1 if the difference of the frequency of the Japanese technical term and the co-occurrence frequency of bilingual technical terms is less than or equal to the upper bound (we use 105 as this upper bound in this paper), while returns 0 otherwise.</td>
</tr>
<tr>
<td></td>
<td>( f_7 ): number of Chinese translation candidates</td>
<td>the number of Chinese translation candidates for the Japanese technical term ( t_J )</td>
</tr>
<tr>
<td></td>
<td>( f_8 ): rate of parallel sentences where phrase alignment is consistent with word alignments</td>
<td>( f_8 = \text{the number of parallel sentences where the phrase alignment is consistent with word alignments } \over \text{co-occurrence frequency of the Japanese-Chinese technical term pair} )</td>
</tr>
<tr>
<td></td>
<td>( f_9 ): translation probability of compositional translation generation</td>
<td>translation probability when generating the Chinese translation candidate compositionally from constituents of the Japanese technical term</td>
</tr>
</tbody>
</table>
Table 5: Evaluation Results (%)

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>60.0</td>
<td>100</td>
<td>75.0</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximum</td>
<td>93.9</td>
<td>59.0</td>
<td>72.5</td>
</tr>
<tr>
<td>precision</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximum</td>
<td>80.6</td>
<td>87.2</td>
<td>83.7</td>
</tr>
<tr>
<td>F-measure</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

set of those positive / negative examples as the reference set of Japanese-Chinese technical term translation pairs in the evaluation of this paper.

In Table 1, we also show the numbers of positive / negative examples for each pair of the 13 frequency ranges of Japanese technical term frequency ($j f$) and Japanese-Chinese co-occurrence frequency ($j c f$). Furthermore, Table 3 lists positive examples of Japanese-Chinese technical term translation pairs for each pair of low / middle / high frequency ranges of Japanese technical term frequency ($j f$) and Japanese-Chinese co-occurrence frequency ($j c f$).

5.3 Procedure of Applying SVM

In the training and testing of the classifier for identifying bilingual technical terms, we first divide the reference set of 2,092 bilingual technical terms into 10 subsets. Here, Japanese-Chinese bilingual technical term pairs which are generated from an identical Japanese term are collected into one subset, but are not separated into more than one subsets.

As a tool for learning SVMs, we use TinySVM9. As the kernel function, we use the polynomial (2nd order) kernel. In the training of SVMs, we use 8 subsets out of the whole 10 subsets. In the tuning of SVMs classifier, we regard the distance from the separating hyper-plane to each test instance as a confidence measure and tune the lower bound of the confidence with one of the remaining two subsets. We consider tuning instances satisfying the confidence measure over a certain lower bound only as positive samples. Here, we tune the lower bound in two ways: i.e, for maximizing precision while keeping recall more than or equal to 60%10, and for maximizing F-measure. In the testing, we test the trained classifier against another one of the remaining two subsets, where we return test instances satisfying the confidence measure over the lower bound only as positive samples. Finally, we repeat this procedure of training / tuning / testing 10 times, and average the 10 results of test performance.

5.4 Features

Table 4 lists all the features used for training and testing of SVMs for identifying Japanese-Chinese technical term translation pairs. Features are roughly divided into two types: those of the first type $f_1$ and $f_2$ are monolingual features, while those of the second type $f_2, \cdots, f_9$ are bilingual features which represent various characteristics of the input bilingual technical term pairs.

Chinese technical term translation pairs, which is different from that used in (Dong et al., 2015). While over 95% of those correct technical term translation pairs are the same as those used in (Dong et al., 2015), only about 55% of erroneous ones are the same as those used in (Dong et al., 2015).

9http://chasen.org/~taku/software/TinySVM

10 In the situation of a real application of the technique of compiling a bilingual technical term lexicon, it is recommended to prefer precision rather than F-measure as the evaluation criterion. This is mainly because those who are working on lexicon compilation just judge whether the output bilingual technical term pairs are correct or not and keep the positive ones while ignore the negative ones, instead of finding out the appropriate translations for all of the negative cases.
Among the monolingual features are the frequency of the Japanese term \( (f_1) \) and the frequency of the Chinese term \( (f_2) \), where their feature values are represented as IDs of the 13 frequency ranges.

Among the bilingual features are the translation probability \( (f_3) \), rank of the Chinese translation candidates \( (f_4) \), co-occurrence frequency of the bilingual technical term pairs \( (f_5) \), the difference of the frequency of Japanese technical term and the co-occurrence frequency of bilingual technical term pairs \( (f_6) \)\(^{11}\), the number of Chinese translation candidates \( (f_7) \), rate of parallel sentences where phrase alignment is consistent with word alignments \( (f_8) \), and the translation probability when generating the Chinese translation candidates compositionally from constituents of the Japanese technical term \( (f_9) \)\(^{12}\).

The following briefly describes why we employ those features introduced in this section. First, we observed that each term of a bilingual technical term pair tends to be a correct translation of each other when their frequencies are close to each other. Also, since we apply the polynomial (2nd order) kernel as the kernel function of SVMs, we simply introduce primitive features such as frequency of Japanese terms \( (f_1) \), frequency of Chinese terms \( (f_2) \), and co-occurrence frequency of bilingual technical term pairs \( (f_5) \), as well as the difference of those frequencies \( (f_6) \). In addition to that, they are correct translation of each other if they have a high translation probability and/or are ranked highly in the SMT phrase translation table. Thus, we use those information directly as the features of \( f_3 \) and \( f_4 \). Furthermore, we define a feature for the translation probability of compositional translation generation using the phrase translation table \( (f_9) \). We also employ the number of translation candidates as another feature \( (f_7) \), since a term tends to be a technical term if the number of its translation candidates is small. Finally, we employ the rate of parallel sentences where phrase alignment is consistent with word alignments as a feature \( (f_8) \), since this rate tends to be large in the case of correct translation pairs.

### 5.5 Evaluation Results

Table 5 shows the evaluation results for a baseline as well as for SVMs. As the baseline, we simply judge all of the input Japanese-Chinese technical term pairs as correct translation, which is exactly the same procedure as shown in Figure 1. In the tuning of the lower bound of the confidence measure, when maximizing precision, we achieve almost 94% precision while keeping recall almost 60% with the test data. When maximizing F-measure, we achieve almost 84% F-measure with around 80% precision and 87% recall.

Table 6 also shows the evaluation results for each pair of the 13 frequency ranges of Japanese technical term frequency \( (jf) \) and Japanese-Chinese co-occurrence frequency \( (jcf) \). As shown in the table, in most pairs of the 13 frequency ranges of Japanese technical term frequency and Japanese-Chinese co-occurrence frequency, we achieve around 90% or higher precision. It is also obvious by comparing those evaluation results with the rates of positive examples for pairs of the 13 frequency ranges of Japanese technical term frequency \( (jf) \) and Japanese-Chinese co-occurrence frequency \( (jcf) \) in Table 1 that, we have lower recalls for certain frequency range pairs when the rates of positive examples are lower for those frequency range pairs\(^{13}\).

---

\(^{11}\)The upper bound 105 shown in Table 4 is used following the result of a preliminary experiment.

\(^{12}\)Patent families are one of the largest parallel sentences resource which contain lots of technical term pairs, and their number grows year by year. As the result of using patent families as knowledge source for solving the task of extracting bilingual technical term pairs, some of the features studied in this paper, such as co-occurrence frequency of the bilingual technical term pairs \( (f_5) \) and the number of Chinese translation candidates \( (f_7) \) and so on, happen to be effective only in the case of using patent families as knowledge source.

\(^{13}\)It is also interesting to note that the higher the frequencies of the Japanese technical terms are, the more the variety of translation candidates is, the more the rates of negative examples are, and finally the lower the recall is. On the contrary, the higher the co-occurrence frequencies of the Japanese-Chinese technical term pairs are, the more reliably
Table 6: Evaluation Results (Precision / Recall / F-measure) (%) for Each Frequency Range of Japanese Technical Term Frequency ($j_f$) and Japanese-Chinese Co-occurrence Frequency ($j_{cf}$).

| $j_f$ | $2 \leq j_f \leq 5$ | $6 \leq j_f \leq 10$ | $11 \leq j_f \leq 15$ | $16 \leq j_f \leq 20$ | $21 \leq j_f \leq 30$ | $31 \leq j_f \leq 50$ | $51 \leq j_f \leq 100$ | $101 \leq j_f \leq 200$ | $201 \leq j_f \leq 500$ | $501 \leq j_f \leq 1,000$ | $1,001 \leq j_f \leq 10,000$ | $10,001 \leq j_f \leq 10,000$ | Total |
|------|----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|------|
|      | 100/27.1 | 40.0/0 | 53.0/0 | 3.9/0 | 4.2/0 | 0.0/0 | 0.0/0 | 0.0/0 | 0.0/0 | 0.0/0 | 0.0/0 | 0.0/0 | 89.3/7.9 | 22.6 |
| $2 \leq j_{cf} \leq 5$ | 97.9/83.7 | 71.8/69.3 | 63.2/60.9 | 53.8/51.4 | 62.9/58.6 | 75.0/69.3 | 85.8/87.2 | 75.0/75.0 | 85.0/84.5 | 57.2/52.1 | 70.0/67.7 | 77.0/72.6 | 91.4/85.0 | 50.6 |
| $6 \leq j_{cf} \leq 10$ | 100/95.2 | 97.5/96.8 | 92.4/92.9 | 92.5/92.9 | 80.0/84.4 | 75.0/75.0 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 98.5/96.8 | 76.2 |
| $11 \leq j_{cf} \leq 15$ | 100/97.4 | 98.7/96.2 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 91.5/89.4 | 87.8 |
| $16 \leq j_{cf} \leq 20$ | 100/98.4 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 94.6/93.3 | 93.9 |
| $21 \leq j_{cf} \leq 30$ | 100/98.5 | 75.0/75.0 | 66.7/66.7 | 75.0/75.0 | 85.8/85.8 | 80.0/80.0 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 88.1/96.8 | 92.2 |
| $31 \leq j_{cf} \leq 50$ | 100/98.0 | 87.5/87.5 | 92.4/92.4 | 75.0/75.0 | 80.0/80.0 | 80.0/80.0 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 88.3/88.3 | 88.3 |
| $51 \leq j_{cf} \leq 100$ | 100/95.9 | 97.6/97.6 | 92.4/92.4 | 80.0/80.0 | 86.7/86.7 | 86.7/86.7 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 95.9/92.9 | 94.3 |
| $101 \leq j_{cf} \leq 200$ | 100/98.0 | 92.4/92.4 | 80.0/80.0 | 86.7/86.7 | 86.7/86.7 | 86.7/86.7 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 96.5/96.5 | 91.2 |
| $201 \leq j_{cf} \leq 500$ | 100/98.0 | 92.4/92.4 | 80.0/80.0 | 86.7/86.7 | 86.7/86.7 | 86.7/86.7 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 98.7/98.7 | 98.7 |
| $501 \leq j_{cf} \leq 1,000$ | 100/98.0 | 92.4/92.4 | 80.0/80.0 | 86.7/86.7 | 86.7/86.7 | 86.7/86.7 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 99.7/99.7 | 99.7 |
| $1,001 \leq j_{cf} \leq 10,000$ | 100/98.0 | 92.4/92.4 | 80.0/80.0 | 86.7/86.7 | 86.7/86.7 | 86.7/86.7 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 99.7/99.7 | 99.7 |
| $10,001 \leq j_{cf} \leq 10,000$ | 100/98.0 | 92.4/92.4 | 80.0/80.0 | 86.7/86.7 | 86.7/86.7 | 86.7/86.7 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 100/100 | 99.7/99.7 | 99.7 |
| Total | 100/27.1 | 98.3/88.0 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 | 98.9/86.4 |
Table 7: Examples of Judgement by SVM

(a) Examples of Correct Judgement by SVM

<table>
<thead>
<tr>
<th>Japanese technical term for evaluation</th>
<th>Chinese translation candidate</th>
<th>feature $f_1$</th>
<th>feature $f_2$</th>
<th>feature $f_3$</th>
<th>feature $f_4$</th>
<th>feature $f_5$</th>
<th>feature $f_6$</th>
<th>feature $f_7$</th>
<th>feature $f_8$</th>
<th>reference judgement</th>
<th>judgement by SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>置換/基 (substituent)</td>
<td>取代基 (substituent)</td>
<td>1,000 ≤</td>
<td>10,000</td>
<td>0.8</td>
<td>1</td>
<td>1,000 ≤</td>
<td>10,000</td>
<td>6</td>
<td>0.12</td>
<td>correct translation</td>
<td>correct translation</td>
</tr>
<tr>
<td>気/液/分離/器 (gas-liquid separator)</td>
<td>気液/反应/器 (gas-liquid reactor)</td>
<td>1,000 ≤</td>
<td>10,000</td>
<td>0.0008</td>
<td>14</td>
<td>jcf=1</td>
<td>14</td>
<td>0</td>
<td>translation error</td>
<td>translation error</td>
<td></td>
</tr>
</tbody>
</table>

(b) Examples of Erroneous Judgement by SVM

<table>
<thead>
<tr>
<th>Japanese technical term for evaluation</th>
<th>Chinese translation candidate</th>
<th>feature $f_1$</th>
<th>feature $f_2$</th>
<th>feature $f_3$</th>
<th>feature $f_4$</th>
<th>feature $f_5$</th>
<th>feature $f_6$</th>
<th>feature $f_7$</th>
<th>feature $f_8$</th>
<th>reference judgement</th>
<th>judgement by SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>カバー/絕縁/膜 (cover insulating layer)</td>
<td>盤絶縁膜 (substring of word “cover” + insulating layer)</td>
<td>501 ≤</td>
<td>1,000</td>
<td>0.05</td>
<td>3</td>
<td>21 ≤jcf</td>
<td>30</td>
<td>10</td>
<td>0.57</td>
<td>translation error</td>
<td>correct translation</td>
</tr>
<tr>
<td>駆動/回路 (drive circuit)</td>
<td>駆動回路電路 (drive circuit)</td>
<td>10,001 ≤</td>
<td>jcf</td>
<td>0.006</td>
<td>3</td>
<td>101 ≤jcf</td>
<td>200</td>
<td>5</td>
<td>0</td>
<td>correct translation</td>
<td>translation error</td>
</tr>
</tbody>
</table>

Next, Table 7 shows examples of correct and erroneous SVMs’ judgments. As shown in Table 7(a), a Japanese-Chinese technical term pair (“置換/基”, “取代基”) are correctly judged by SVM, mainly because its translation probability in the phrase translation table ($f_3$) is high and the rank of the Chinese translation candidate ($f_4$) is 1. In addition, its translation probability of compositional translation generation ($f_5$) are relatively high, and its number of Chinese translation candidates ($f_7$) are relatively small. Compared with this result of correct translation by the framework of this paper based on Chinese sentences segmented by characters, on the other hand, the framework of Dong et al. (2015) based on Chinese sentences segmented by morphemes translates the Japanese technical term “置換/基” not only into the correct Chinese translation “取代基”, but also into an erroneous Chinese translation “取替/基如” (which means “substituent such as”). This is mainly because of the error in Chinese morphological analysis where the two Chinese characters “基” and “如” are concatenated into one morpheme, and then, SVM trained with the phrase translation table with Chinese sentences segmented by morphemes can not judge “取替/基如” as an erroneous translation. Also, another Japanese-Chinese technical term pair (“気/液/分離/器”, “気液/反应/器”) is correctly judged by SVM to be a translation error, mainly because its values of $f_3$ and $f_8$ are 0 or quite small, while those of $f_4$ and $f_7$ are fairly large.

Table 7(b) shows erroneous judgements by SVM. The first bilingual technical term pair (“カバー/絶縁/膜”, “盖/絕緣/層”) is a translation error because the Chinese character “盖” is a substring of the Chinese word “覆盖” (which means “cover”), while the correct translation should be (“カバー/絶縁/膜”, “盖/絕緣/層/層”). In the framework based on Chinese sentences segmented by characters, however, both of these two bilingual technical term pair are judged as correct translations. The erroneous bilingual technical term pair (“カバー/絶縁/膜”, “盖/絕緣/層/層”) is judged to be a correct translation mainly because its values of $f_3$ and $f_8$ are not quite small, while the rank of $f_4$ is relatively high and the value of $f_7$ is relatively small. Compared with this result of erroneous translation by the framework of this paper based on Chinese
sentences segmented by characters, on the other hand, in the framework of Dong et al. (2015) based on Chinese sentences segmented by morphemes, SVM judges only “覆蓋/被覆層” as the correct translation of the Japanese technical term “カバー/被覆層”, while it judges “電/電纜層” as the erroneous translation of “カバー/被覆層”. This is mainly because, for the erroneous translation pair (“カバー/被覆層”, “電/電纜層”), the value of $f_9$ is low.

The second bilingual technical term pair, (“騒動/騒音”, “騒動/騒音”), on the other hand, is a correct translation according to the reference judgement. However, this technical term pair is judged by SVM to be a translation error mainly because the value of the translation probability of compositional translation generation $f_9$ is 0. In this example, although the Japanese-Chinese constituent phrase translation pair (“騒動”, “騒音”) exists in the phrase translation table, its translation probability is below the pre-determined lower bound $^{14}$.

### 6 Related Work

Among related works on acquiring bilingual lexicon from text, Itagaki et al. (2007) focused on automatic validation of translation pairs available in the phrase translation table trained by an SMT model. Itagaki et al. (2007) especially studied to apply a Gaussian mixture model based classifier to the task of automatic validation of translation pairs available in the phrase translation table. Yasuda and Sumita (2013) also studied to extract bilingual terms from comparable patents, where, they first extract parallel sentences from patent families, and then extract bilingual terms from parallel sentences. Yasuda and Sumita (2013) especially studied to exploit kanji character similarity between Japanese and Chinese languages in the task of extracting Japanese-Chinese bilingual term pairs. It is also reported that two types of SMT phrase translation tables are integrated in this task. Haque et al. (2014a) also presented a bilingual terminology extraction method using the phrase translation table trained by a phrase-based SMT. One of the major differences of our approach and those proposed in Itagaki et al. (2007), Yasuda and Sumita (2013) and Haque et al. (2014b) is that we apply the SVM-based classifier learning framework to the task of identifying bilingual technical term pairs from parallel patent sentences, where we examine various features extracted from parallel patent sentences themselves as well as the phrase translation table of a statistical machine translation model trained with those parallel patent sentences.

Lu and Tsou (2009) also studied to extract English-Chinese bilingual terms from patent families, where they first extract parallel sentences from patent families, and then extract bilingual terms from parallel sentences based on an SVM classifier. One of the major differences of our approach and that proposed in Lu and Tsou (2009) is that our features studied in this paper are much finer-grained and cover wider range of information that are available from parallel patent sentences themselves as well as the phrase translation table of a statistical machine translation model trained with those parallel patent sentences.

Morishita et al. (2008) studied to acquire Japanese-English technical term translation lexicon from the phrase translation tables, which are trained by a phrase-based SMT model with parallel sentences automatically extracted from patent families. The approach taken in Morishita et al. (2008) is based on integrating the phrase translation table and compositional translation generation based on an existing bilingual lexicon for human use. This approach is quite effective in the case of language pairs such as Japanese and English, where an existing bilingual lexicon for human use is widely available. However, this is not always the case in the case of other language pairs such as Japanese and Chinese. Compared with the approach of Morishita et al. (2008), our approach is advantageous in that we concentrate on utilizing information that are available from patent families, but not rely on information source other than patent fami-

---

$^{14}$In this paper, we introduce a lower bound of the translation probability of constituent phrase translation pairs in the procedure of compositional translation generation of the feature $f_9$, and set the lower bound as 0.005.
lies. Also, compared with the features of SVM examined in Morishita et al. (2008), those we employed in this paper cover much wider range of information that are available from patent families. Especially, we concentrate more on utilizing features that are based on statistics of all the parallel sentences of the patent families rather than a single parallel sentence pair. In our proposed framework, we introduce the feature of the number of Chinese translation candidates ($f_7$), which was not examined in Morishita et al. (2008). We also use the rate of parallel sentences where phrase alignment is consistent with word alignments as a feature ($f_8$), while Morishita et al. (2008) used a binary feature which judges for each parallel sentence pair whether a phrase alignment is consistent with word alignments. Finally, we use the feature of the translation probability of compositional translation generation, which, through a preliminary evaluation, is proved to perform better than the binary feature of compositional translation generation employed in Morishita et al. (2008).

7 Conclusion

In this paper, we segment the Chinese sentences into characters instead of segmenting them into morphemes as in (Dong et al., 2015), and represent Japanese-Chinese patent families in terms of character-segmented Chinese sentences and morpheme-segmented Japanese sentences. Then, to those Japanese-Chinese patent families, we apply the framework (Dong et al., 2015) of identifying bilingual technical terms. As a result, we achieve the performance of over 90% precision with the condition of more than or equal to 60% recall.

As a future work, in order to avoid errors caused by character-based segmentation of Chinese sentences as discussed in section 5.5, we plan to integrate two types of a Japanese-Chinese phrase translation table, where Chinese sentences are segmented not only by characters, but also by morphemes. Another future work is to integrate phonetic level (Xu et al., 2006) as well as character level (Chu et al., 2013) correspondences between Japanese and Chinese within our feature framework such as the one of the translation probability of compositional translation generation ($f_9$). As the phonetic level correspondences, introducing the framework of transliteration based on Katakana-Pinyin correspondence (Xu et al., 2006) is expected to improve the performance. As the character level correspondences, introducing the correspondence based on shared Chinese characters between Japanese and Chinese languages is expected to improve the performance.

References


Resampling Approach for Instance-based Domain Adaptation from Patent Domain to Newspaper Domain in Statistical Machine Translation

Keisuke Noguchi
noguchi@ai.cs.ehime-u.ac.jp
Department of Computer Science, Faculty of Engineering, Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan

Takashi Ninomiya
ninomiya@cs.ehime-u.ac.jp
Graduate School of Science and Engineering, Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan

Abstract
In this paper, we investigate a resampling approach for domain adaptation from a resource-rich domain (patent domain) to a resource-scarce target domain (newspaper domain) in Statistical Machine Translation (SMT). We propose two resampling methods for domain adaptation in SMT: random resampling and resampling for instance weighting. The random resampling randomly adds sentence pairs from the resource-rich parallel corpus to the target-domain parallel corpus. Instance weighting is a method which provides a weight to each sample in the resource-rich domain. The problem of instance weighting in SMT is how to provide a weight to each sentence pair. We approximate the instance weights by resampling sentence pairs according to the ratio of sentence-pair probabilities between the two domains. We also explore a method of selecting samples that have instance weights larger than some threshold.

1 Introduction
In the last few decades, Statistical Machine Translation (SMT) has been widely studied in the field of machine translation because SMT is mathematically well-defined in terms of the probabilistic models it uses, and it can be learned automatically from large bilingual parallel corpora by using publicly available SMT tools. One of the key issues in SMT is how to develop a large parallel corpus. Developing a large-scale parallel corpus by hand is very expensive, and in the past, SMT systems were usually trained with only tens of thousands sentence pairs. In the last decade, large-scale parallel corpora consisting of around millions of sentence pairs were developed by using the methods for automatically acquiring parallel sentence pairs from comparable corpora (Utiyama and Isahara, 2003; Koehn, 2005; Callison-Burch et al., 2009). For example, a large-scale Japanese-English patent parallel corpus consisting of around 3 million sentence pairs was automatically developed from patent documents (Utiyama and Isahara, 2003). The Europarl parallel corpus consists of around 0.4 to 2 million sentence pairs for each language pair, which were extracted from the proceedings of the European Parliament (Koehn, 2005). The French-English Gigaword \(10^9\) parallel corpus consists of around 22 million sentence pairs crawled from Canadian and European Internet pages (Callison-Burch et al., 2009). These automatically acquired parallel corpora are large enough for SMT training if their domain and the target domain are the same. However, in general, the target domains of machine
translation are often different from the domains of the automatically acquired parallel corpora. It is empirically known that translation quality drastically deteriorates when an SMT system trained on one domain is applied to other domains (Foster et al., 2010). Therefore, domain adaptation is needed from the domains of these automatically acquired parallel corpora to the target domain.

This paper investigates a resampling approach for domain adaptation from the patent domain to the newspaper domain in SMT. We propose two resampling methods for domain adaptation in SMT: random resampling and resampling for instance weighting. The random resampling method randomly adds sentence pairs from the large-scale parallel corpus to the target-domain parallel corpus. Instance weighting (Jiang and Zhai, 2007) gives each sample a weight calculated by dividing the probability of the sample in the target domain by the probability of the sample in the large-scale data domain; i.e., the weight is an estimated frequency of the sample in the target domain. The problem of instance weighting in SMT is how to provide a weight to each sentence-pair. As SMT tools have many complex components, wrapper methods which do not modify the tools themselves are preferable. We approximate the instance weights by resampling sentence-pairs according to a ratio of sentence-pair probabilities between the two domains. We also explore a method of selecting samples that have instance weights larger than some threshold.

2 Domain Adaptation

Domain adaptation, which is also called transfer learning, is a method for adapting the model or data in a resource-rich domain to a resource-scarce target domain. The goal of domain adaptation is to increase the performance of systems in the resource-scarce domain by leveraging the model or data in the resource-rich domain. The resource-scarce domain is called the target domain or in-domain, and the resource-rich domain is called the source domain or out-domain. In what follows, we call the data/model in the resource-scarce domain the in-domain data/model and the data/model in the resource-rich domain the out-domain data/model.

2.1 Related Work

Domain adaptation can be classified into two types of adaptation: model adaptation and instance weighting (Foster and Kuhn, 2007; Jiang and Zhai, 2007). In the model adaptation, models are learned from the in-domain data and the out-domain data separately, and then a mixture model is developed by mixing the in-domain model and the out-domain model so that the mixture model performs better for the in-domain data. Foster and Kuhn (2007) proposed model adaptation for SMT. They used a log-linear model for the mixture model learned by using MERT (Och, 2003) with other log-linear parameters. They also proposed four distance metrics to measure the weight for each model; TF-IDF, LSA, perplexity, and EM. The mixture model learned by using MERT is also used as a baseline in (Foster et al., 2010). Moore and Lewis (2010) used perplexity for language model adaptation.

Instance weighting can be further classified into the metrics approach, weight optimization approach, and covariate shift approach. The metrics approach gives each sample a weight representing the distance between the sample as an in-domain sample and the sample as an out-domain sample. There have been several studies on the metrics approach in SMT, wherein the metrics used were the cross-entropy (Yasuda et al., 2008) or cross-entropy difference (Axelrod et al., 2011, 2012) for a sentence-pair. In the metrics approach, sentence-pairs in the out-domain corpus are selected if their metrics is closer than some threshold. Yasuda et al. (2008) also used model adaptation in addition to the metrics approach. In the weight optimization approach, a weight is assigned to each sentence/phrase in the training corpus, and these weights are optimized so as to maximize the objective function for the in-domain development.
corpus (Matsoukas et al., 2009; Foster et al., 2010). Matsoukas et al. (2009)’s model was originally proposed for data selection, not for domain adaptation, but it can also be applied to domain adaptation. The covariate shift approach (Shimodaira, 2000; Jiang and Zhai, 2007; Xia et al., 2013, 2014) gives each sample a weight calculated by dividing the in-domain probability of the sample by the out-domain probability of the sample; i.e., the weight is an estimated frequency of the sample in the target domain. Most of the previous studies on instance weighting directly incorporate the weights into the objective function or select the samples having weights higher than some threshold. Our method is also based on instance weighting, but our method is designed for SMT and uses resampling for estimating the probability ratio.

The most similar approach to our method is that of Gascó et al. (2012). They proposed the resampling method for domain adaptation for SMT. The difference between our method and theirs is that we use the covariate shift approach using both in-domain probabilities and out-domain probabilities, whereas they use only the in-domain probabilities.

2.2 Instance Weighting for SMT

Given parallel sentence pairs \((s_i, t_i)_{i=1}^N\) as a training data set, where \(s_i\) is a sentence in the source language and \(t_i\) is a sentence in the target language, the parameter estimation for SMT in the maximum likelihood estimation framework is defined as follows.

\[
\hat{\theta} = \arg \max_{\theta} \sum_{s \in S} \sum_{t \in T} p(s, t) \log p(t|s; \theta)
\]

\[
\approx \arg \max_{\theta} \sum_{s \in S} \sum_{t \in T} \tilde{p}(s, t) \log p(t|s; \theta) = \arg \max_{\theta} \sum_{i=1}^N \log p(t_i|s_i; \theta)
\]  

where \(S\) is the source language, \(T\) is the target language, and \(\tilde{p}\) is an empirical distribution. The instance weighting is derived as follows (Jiang and Zhai, 2007).

\[
\hat{\theta} = \arg \max_{\theta} \sum_{s \in S} \sum_{t \in T} p_{in}(s, t) \log p(t|s; \theta)
\]

\[
= \arg \max_{\theta} \sum_{s \in S} \sum_{t \in T} \frac{p_{in}(s, t)}{p_{out}(s, t)} p_{out}(s, t) \log p(t|s; \theta)
\]

\[
\approx \arg \max_{\theta} \sum_{i=1}^N \frac{p_{in}(s_i, t_i)}{p_{out}(s_i, t_i)} \log p(t_i|s_i; \theta)
\]  

where \(p_{in}\) is the in-domain probabilities and \(p_{out}\) is the out-domain probabilities. Instance weighting gives each sentence pair a weight calculated by dividing the in-domain probability of the sentence pair by the out-domain probability of the sentence pair; i.e., the weight is an estimated in-domain frequency of the sentence pair.

3 Random Resampling for SMT

In domain adaptation, there are two simple and strong baseline methods: one which trains a model only on the in-domain data set and one which trains a model on the union of the in-domain and out-domain data sets (Daume III, 2007). The baseline method using both data sets is strong, but in our experiments it was worse than the baseline method using only the in-domain data set. We think that this deterioration is caused by adding too many data samples from the out-domain data set to the in-domain data set.
To solve the problem of the baseline methods mentioned above, we introduced a random resampling method (He and Garcia, 2009) for domain adaptation in SMT. Resampling was originally studied to solve the imbalanced data problem in binary classification, but it can also be applied to domain adaptation. There are two methods for random resampling: oversampling and undersampling (He and Garcia, 2009). Both methods use both an in-domain corpus and out-domain corpus as a training data set by default. Oversampling increases the size of the in-domain corpus by copying randomly selected sentences in the in-domain corpus. It may add the same sentences in the in-domain corpus many times but it increases the weights for the in-domain sentences in the objective function compared to the out-domain sentences. Undersampling decreases the weights for the out-domain corpus by removing randomly selected sentences in the out-domain corpus. In the experiments, we used undersampling as the random resampling method.

4 Resampling for Instance Weighting in SMT

We propose a resampling method for instance weighting in SMT. We approximate the probability ratio in Equation 2 with in-domain and out-domain language models as follows.

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^{N} \frac{p_{in}(s_i, t_i)}{p_{out}(s_i, t_i)} \log p(t_i | s_i; \theta) \approx \arg \max_{\theta} \sum_{i=1}^{N} \frac{p_{in}(t_i)}{p_{out}(t_i)} \log p(t_i | s_i; \theta)$$

(3)

where \(p_{in}(t_i)\) is an in-domain language model and \(p_{out}(t_i)\) is an out-domain language model. Both language models are defined as \(n\)-gram language models.

In the experiments, we used 5-gram models with Kneser-Ney smoothing for the \(n\)-gram language models. 5-gram models with Kneser-Ney smoothing was learned by using the SRILM tool kit (Stolcke et al., 2011). We calculate the probability of a sentence \(w_1 w_2 \ldots w_n\) as follows.

$$p(w_1 w_2 \ldots w_n) \approx \prod_{j=1}^{n} p(w_j | w_{j-4} w_{j-3} w_{j-2} w_{j-1})$$

(4)

where \(p(w_j | w_{j-4} w_{j-3} w_{j-2} w_{j-1})\) represents 5-gram probabilities. Given a sentence \(t\), \(p_{in}(t)\) and \(p_{out}(t)\) are calculated by using 5-gram models learned from the in-domain parallel corpus and the out-domain parallel corpus, respectively.

Given a sentence \(t\), let \(w(t)\) be the weight \(p_{in}(t) / p_{out}(t)\) in Equation 3. The weight \(w(t)\) for sentence \(t\) represents the in-domain frequency of \(t\). In our resampling method, a sentence pair \((s, t)\) in the out-domain corpus is selected with the probability of \(w(t)\) if \(w(t)\) is less than 1. A sentence pair \((s, t)\) is always selected if \(w(t) \geq 1\). Theoretically, a sentence pair should be resampled multiple times if \(w(t)\) is greater than 1, but we resample the sentence only once because \(w(t)\) can be an extremely large number. Formally, we have the modified resampling number \(w'(t)\) for sentence \(t\) as follows.

$$w(t) = \frac{p_{in}(t)}{p_{out}(t)}, \quad w'(t) = \begin{cases} w(t) & \text{if } w(t) < 1 \\ 1 & \text{otherwise} \end{cases}$$

(5)

In the same way as the previous methods that select sentences using thresholds, we also experimented with the sample selection approach by using thresholds and the weight \(w(t)\) in Equation 5. In this approach, a sentence pair \((s, t)\) is selected if the weight \(w(t)\) is greater than a threshold. All sentence pairs that satisfy this condition in the out-domain parallel corpus are selected.
Table 1: Specification of English-Japanese parallel corpora

<table>
<thead>
<tr>
<th>domain</th>
<th>training set (# of sentence-pairs)</th>
<th>development set (# of sentence-pairs)</th>
<th>test set (# of sentence-pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>patent</td>
<td>3,166,284</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>newspaper</td>
<td>130,000</td>
<td>500</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Table 2: Results

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline (in-domain)</td>
<td>13.93</td>
</tr>
<tr>
<td>baseline (in-domain + out-domain)</td>
<td>12.67</td>
</tr>
<tr>
<td>random resampling</td>
<td>14.24</td>
</tr>
<tr>
<td>instance weighting (resampling)</td>
<td>14.15</td>
</tr>
<tr>
<td>instance weighting (w. thresholds)</td>
<td>14.47</td>
</tr>
</tbody>
</table>

5 Experiments

We evaluated the performance of the random resampling method and the resampling method for instance weighting in English-Japanese SMT.

5.1 Settings

In the experiments, we regarded the newspaper domain as the target domain (in-domain) and the patent domain as the resource-rich domain (out-domain). We used a English-Japanese patent parallel corpus consisting of 3,166,284 sentence pairs; the same one was used for the shared task in NTCIR10 (PatentMT) (Goto et al., 2013). We also used the English-Japanese newspaper parallel corpus, JENAAD (Utiyama and Isahara, 2003). JENAAD consists of 150,000 sentence pairs extracted from the comparable corpora: the Yomiuri Shimbun and the Daily Yomiuri. Table 1 shows the details of the data sets. The language models were learned as 5-gram models with Kneser-Ney smoothing by using the SRILM tool kit (Stolcke et al., 2011). The in-domain language model was learned from the training corpus in the newspaper parallel corpus, and the out-domain language model was learned from the training corpus in the patent parallel corpus.

We used GIZA++ 1.0.7 for word alignment, SRILM 1.5.12 for learning n-gram language models, Moses for SMT (Koehn et al., 2007) and MERT for tuning. The value of distortion limit was infinite. We used Mecab 0.98 with ipadic 2.7.0 for tokenizing Japanese sentences. We used BLEU for measuring translation accuracy. We evaluated SMT only in a direction from English to Japanese.

In the experiments, we also evaluated the SMT trained with only the in-domain parallel corpus and the SMT trained with the union of the in-domain parallel corpus and the out-domain parallel corpus as baseline methods.

5.2 Results

Table 2 shows the experimental results. “baseline (in-domain)” indicates the baseline method using only the in-domain corpus. “baseline (in-domain + out-domain)” indicates the baseline method trained on the union of the in-domain corpus and the out-domain corpus. In this experiment, we used undersampling in the random resampling method. “instance weighting (resampling)” means the resampling method for instance weighting. “instance weighting (w. threshold)” means the instance weighting method that selects the sentence pairs having larger...
Table 3: Change in BLEU with the size of resampled data

<table>
<thead>
<tr>
<th># of added sentence pairs</th>
<th>random resampling (BLEU (%))</th>
<th>instance weighting (w. threshold) (BLEU (%))</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline (in-domain)</td>
<td>0</td>
<td>13.93</td>
</tr>
<tr>
<td></td>
<td>5,525</td>
<td>13.74</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>13.73</td>
</tr>
<tr>
<td></td>
<td>20,000</td>
<td>14.10</td>
</tr>
<tr>
<td></td>
<td>40,000</td>
<td>13.51</td>
</tr>
<tr>
<td>resampling</td>
<td>80,000</td>
<td>14.24</td>
</tr>
<tr>
<td></td>
<td>130,000</td>
<td>13.58</td>
</tr>
<tr>
<td></td>
<td>500,000</td>
<td>13.22</td>
</tr>
<tr>
<td></td>
<td>1,000,000</td>
<td>12.83</td>
</tr>
<tr>
<td></td>
<td>2,000,000</td>
<td>12.92</td>
</tr>
<tr>
<td>baseline (in-domain + out-domain)</td>
<td>3,166,284</td>
<td>12.67</td>
</tr>
</tbody>
</table>

weights than a threshold\(^1\)^. The resampling method for instance weighting selected 5,525 sentence pairs from the out-domain parallel corpus. As seen in the table, the random resampling and instance weighting achieved better BLEU scores than the baseline methods. The instance weighting with thresholds achieved the best BLEU, but the resampling method for instance weighting also achieved comparative results, using fewer out-domain sentence pairs. The random resampling method was also comparable to instance weighting with thresholds.

Table 3 shows the change in BLEU with the number of added sentence pairs in the random resampling method and the instance weighting method with thresholds. In the results of the random resampling method, the best result was achieved with 80,000 sentence pairs, and BLEU increased by 0.31 from the baseline. In the results of the instance weighting method with thresholds, the best result was achieved with 10,000 sentence pairs, and BLEU increased by 0.54 from the baseline. The number of sentence pairs selected by the resampling method for instance weighting, 5,525 sentence pairs, was close to the number of sentence pairs selected by the instance weighting method with thresholds. This means that the resampling method for instance weighting provided a good estimation for the weights without brute-force searching for the threshold.

6 Conclusion

We investigated a resampling approach for domain adaptation from the patent domain to the newspaper domain in SMT. The random resampling method randomly selects sentence pairs from the out-domain parallel corpus. The resampling method for instance weighting selects sentence pairs according to the ratio of sentence-pair probabilities between the two domains. Instance weighting selects the sentence pairs that are likely to be the in-domain sentence pairs from the out-domain parallel corpus. We also explored instance weighting with thresholds, which selects all the sentence pairs having higher weights than a threshold. In this study, \(n\)-gram language models were used for calculating the in-domain and out-domain probabilities.

\(^1\)As an implementation issue, we first sorted the sentence pairs in the out-domain corpus in descending order of their weights. Then we selected the sentence pairs which had high weights.

\(^2\)However, in the experiments, the number of added sentences in the random resampling method and the thresholds in instance weighting with thresholds were wrongly determined by using the test set. So, the true BLEU scores of the random resampling method and instance weighting with thresholds might be lower than the BLEU scores in Table 2.
for sentence pairs. In the experiments, instance weighting with thresholds achieved the best results, but both random resampling and resampling for instance weighting also achieved comparable results. Though the BLEU score of the resampling method for instance weighting was worse than other two methods, the number of resampled sentence pairs was very close to that of the instance weighting method with thresholds. The advantage of the resampling method for instance weighting is that it does not require tuning for finding the thresholds while other methods need a tuning process.

Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number 25280084.

References


Towards Cross-lingual Patent Wikification

Takashi Tsunakawa  
tuna@inf.shizuoka.ac.jp
Hiroyuki Kaji  
kaji@inf.shizuoka.ac.jp
College of Informatics, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan

Abstract

This paper demonstrates the effectiveness of cross-lingual patent wikification, which links technical terms in a patent application document to their corresponding Wikipedia articles in different languages. The number of links increases definitely because different language versions of Wikipedia cover different sets of technical terms. We present an experiment of Japanese-to-English cross-lingual anchor text extraction using a dedicated technical term extraction system and a patent parallel corpus. Cross-lingual anchor text extraction retrieves about 10% more technical terms linked to Wikipedia articles than monolingual extraction. We also show that restricting anchor texts to technical terms in a specified Wikipedia category has effect of reducing the number of destination article candidates.

1. Introduction

Wikification refers to a task of linking phrases in a text to their corresponding Wikipedia articles. It greatly enhances the understandability of the text; namely, readers of a wikified text can easily figure out the meaning of an unfamiliar phrase by clicking it. Applying wikification to patent application documents is helpful for understanding technical terms in them. With the rapidly increasing quantity and improved quality of Wikipedia articles in technical domains, the effectiveness of patent wikification will be enhanced.

A promising extension of wikification is cross-lingual wikification, which links phrases in a text to articles in languages other than that of the text. Wikipedia has more than 35 million articles in more than 290 languages. Since each language version is being edited independently, the quantity and quality of articles are very diverse among languages. While English Wikipedia has nearly five million articles, most other language versions have less than one million articles as of 2015. This indicates that a huge number of entities in the world are explained only in English. Thus, we can enrich wikification results by considering links from other languages to English. Consider a Japanese patent application document that includes the technical term “アーク抑制 (aku yokusei),” which is not explained by any Japanese Wikipedia article but by the English Wikipedia article “Arc suppression.” Cross-lingual wikification links the technical term “アーク抑制 (aku yokusei)” to the English Wikipedia article “Arc suppression.” Such cross-lingual links would be very useful for revealing important entities that could not be found by monolingual wikification.

In this paper, we demonstrate the potential effectiveness of cross-lingual patent wikification. The main target is patent application documents in languages other than English (such as Chinese and Japanese), which have been increasing by leaps and bounds. We showed in an experiment that cross-lingual wikification to Japanese patent application documents could significantly increase the number of links by adding English Wikipedia as the linking destination.
2. Related work

Wikification consists of two steps: anchor text extraction and disambiguation to Wikipedia articles. We describe studies on the former step that is the main concern of this paper. Anchor texts are phrases that should be linked to Wikipedia articles. One of the most important types of information for anchor text extraction is keyphraseness, namely link probabilities that phrases are used as anchor texts for linking to Wikipedia articles (Mihalcea and Csomai, 2007). Another is relatedness with co-occurring phrases (Milne and Witten, 2008) because a text tends to be linked to articles related to it. Keyword extraction techniques are also applicable to anchor text extraction, because keywords in a text should be anchor texts (Mihalcea and Csomai, 2007). There have been many studies for keyword extraction using syntactic and statistical information (Jacquemin and Bourigault, 2003).

With ever more attention being focused on wikification, cross-lingual wikification has been recognized as a new challenging task. Cross-lingual wikification consists of anchor text extraction, translation, and disambiguation. Translation quality severely affects the wikification results with this approach. To our knowledge, translation of technical terms for cross-lingual wikification has not been studied. In the previous studies, a large part of anchor texts are named entities, translation of which requires special techniques such as transliteration, name translation mining from comparable corpora, and information extraction-based techniques (McNamee et al., 2011; Cassidy et al., 2012; Miao et al., 2013).

3. Monolingual vs. cross-lingual anchor text extraction

In this section, we give a detailed description of our approach to cross-lingual patent wikification. Since our present goal is to estimate increases in the number of links by introducing cross-lingual wikification, we focus mainly on the anchor text extraction problem.

What kinds of phrases should be extracted as anchor texts depends on the domain of the text and the application of wikification results. Most anchor texts to be extracted in patent wikification are technical terms, while original wikification (Mihalcea and Csomai, 2007) also extracts named entities. Named entities, such as personal names and place names, seldom occur in patent application documents. In most cases, these are of no interest to readers for understanding what is invented in the patent. We therefore extract technical terms from patent application documents as anchor texts. In order to discriminate technical terms from other kinds of anchor texts, we employ a technical term extraction system.

Our anchor text extraction system built for an experiment consists of three parts: technical term extraction and monolingual/cross-lingual anchor text extraction (Fig. 1).

1) Technical term extraction

We first extract technical terms from patent application documents as anchor candidates by a technical term extraction system. To extract terms from Japanese patent application documents, we employ termex,¹ the automatic domain terminology extraction system developed at Nakagawa Laboratory, University of Tokyo and Mori Laboratory, Yokohama National University. This system is based on occurrence and concatenation frequencies of simple and compound nouns (Nakagawa and Mori, 2003). Termex assigns a score that approximates termhood for each extracted technical term and outputs the ranked list of technical terms for an input document. We assume the whole output of termex as candidates of anchor texts without filtering by the scores.

¹ http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html
2) **Monolingual anchor text extraction**

From anchor candidates, we next extract anchor texts by matching with prior anchor texts in the same language. Prior anchor texts are n-grams that have been anchor texts used to link to Wikipedia articles at least once. Each prior anchor text is associated with one or more destination articles; for example, a Japanese prior anchor text “圧縮 (asshuku)” has possible destinations “圧縮 (asshuku; compression),” “データ圧縮 (deta asshuku; data compression),” etc. It is likely that occurrence of the anchor text in a text will be linked to one of the possible destinations. Accordingly, we regard anchor candidates that match one prior anchor text as anchor texts. Extracting more anchor texts indicates the possibility of enriching patent application documents with more links to Wikipedia.

Optionally, we use Wikipedia category information to extract only anchor texts related to the domain of the patent application document. Every Wikipedia article belongs to at least one category, which is systematically organized in a taxonomy style. By specifying a category of users’ interest, we can only extract anchor texts if one of their destination articles belongs to a category subsumed by the specified category. Another advantage is removing non-technical terms from anchor candidates extracted by the technical term extraction system.

3) **Cross-lingual anchor text extraction**

In the cross-lingual part, extracted anchor candidates are translated to English by looking up a phrase table constructed from a Japanese-English patent parallel corpus. We generate n-best translations for each anchor candidate by using phrase pairs whose phrase translation probabilities are not below a threshold value $\theta$. If one of the translations matches an English Wikipedia article title, we regard the original Japanese anchor candidate as an anchor text.

---

2 Some prior anchor texts may have different meanings from all possible destination articles in current Wikipedia. We ignore such rare cases because they seem not to significantly affect our evaluation results.

3 We do not apply full phrase-based statistical machine translation to patent application documents because it might cause more translation errors due to the high complexity of sentence-level translation. We consider that phrase-level translation is enough to anchor text extraction for cross-lingual wikification.
The reason we do not use English prior anchor texts is to avoid unexpected matches occurring from translation. For example, consider the Japanese anchor text “細胞 (saibo)”, which means a cell in biology. The phrase table gives its translation as “cell,” and one of the destination articles is “Electrochemical cell,” which the Japanese anchor text does not mean. Cross-lingual links may be overestimated when the destination article does not describe the anchor text in a patent application document.

4. Experiment

In this section, we describe an experiment for proving the effectiveness of cross-lingual patent wikification. We compare the numbers of technical terms extracted by monolingual and cross-lingual anchor text extraction. We also confirm that use of Wikipedia categories has effect on disambiguation to Wikipedia articles for anchor texts.

4.1. Data used

In our experiment we used the NTCIR-7 PATMT test collection, which consists of unexamined Japanese patent application documents published in 1993-2002 and USPTO English patent grant data published in 1993-2000. From the test collection, we extracted 10,000 Japanese patent application documents published in 2000 with the international patent classification (IPC) class G06 (COMPUTING; CALCULATING; COUNTING) as the input.

We employed a phrase table provided from a research group in University of Tsukuba, which was used for identification of bilingual synonymous technical terms (Liang et al., 2011). This phrase table was constructed from a Japanese-English patent parallel corpus (Utiyama and Isahara, 2007) in the NTCIR-7 PATMT test collection by using the Moses toolkit (Koehn et al., 2007). Setting the phrase translation probability threshold $\theta$ to 0.01, we generated the 10-best translations for each anchor candidate for cross-lingual anchor text extraction.

We used Japanese and English Wikipedia data dumped in March 2013 for collecting Japanese prior anchor texts, English article titles, and category information. For an experiment using categories, we manually specified the category “計算機科学 (Computer science)”, which is strongly related to the IPC class cited above. We extract anchor texts if at least one of their destination articles belongs to the categories subsumed by the specified category. To reduce computational complexity, the subsumed categories are limited to categories that can be reached from the specified category by iteratively tracking subcategories within 10 times.

4.2. Anchor text extraction results

Table 1 shows an example of monolingual and cross-lingual anchor text extraction results for anchor candidates extracted by termex from an input patent application document. Score indicates a termhood score output by termex. Check marks in the Mono and CL columns show that the anchor candidate was selected as anchor texts that could be linked to an article by monolingual and cross-lingual anchor text extraction, respectively. For example, the term “復号化 (decoding)” in the second row is checked only in the CL column because it was linked to an English article “Decoding” but Japanese Wikipedia did not have an independent article to describe it. The InCat column shows whether one of the Japanese destination articles belongs to a category subsumed by the specified category “計算機科学 (Computer science).”

After extracting all anchor texts in the 10,000 patent application documents, we obtained proportions of anchor candidates selected as anchor texts in each interval of the termhood scores: 1-10, 10-100, 100-1000, 1000-3000, 3000-5000, 5000-10000, and 10000-. The total numbers of anchor candidates in each score interval are described in Table 2.
Table 1. Examples of extracted anchor texts.

<table>
<thead>
<tr>
<th>Extracted anchor candidate</th>
<th>Score</th>
<th>Anchor text decision Mono</th>
<th>CL</th>
<th>InCat</th>
</tr>
</thead>
<tbody>
<tr>
<td>情報 (information)</td>
<td>2544.40</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>複号化 (decoding)</td>
<td>2318.54</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>遊技プログラム情報 (game program information)</td>
<td>2090.44</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>書き込み情報 (write information)</td>
<td>1815.47</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>複号手段 (decoder)</td>
<td>1648.58</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>複号書き込み情報 P (decoding write information P)</td>
<td>1621.06</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>複号キー (decoding key)</td>
<td>1464.71</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>複号キー K (decoding key K)</td>
<td>1162.16</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>複号書き込み情報 (decoding write information)</td>
<td>1128.33</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>メモリ (memory)</td>
<td>155.88</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>自己検証回路 (self verification circuit)</td>
<td>135.04</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>遊技 (game)</td>
<td>122.38</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>暗号化処理 (encrypting/encoding)</td>
<td>119.73</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>キー (key)</td>
<td>119.65</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>アルゴリズム (algorithm)</td>
<td>39.50</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>制御 (control)</td>
<td>34.50</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>遊技内容 (game content)</td>
<td>33.08</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ROMライタ (ROM writer)</td>
<td>30.32</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Numbers of anchor candidates.

<table>
<thead>
<tr>
<th>Score interval</th>
<th>1-10</th>
<th>10-100</th>
<th>100-1000</th>
<th>1000-5000</th>
<th>5000-10000</th>
<th>10000-</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>1732k</td>
<td>825k</td>
<td>257k</td>
<td>37k</td>
<td>7973</td>
<td>5944</td>
</tr>
</tbody>
</table>

With the category specified, similar tendency between proportions and scores was observed. Though the percentage of anchor texts in the subsumed categories strongly depends on the specified category, it ranged from one third to a half. This shows that cross-lingual anchor text extraction is also effective for technical terms with categories limited.
The proportions slightly improve in higher score sections. This indicates that the employed technical term extraction method would be suitable for anchor text extraction of patent wikification, because the system preferentially selects technical terms that are described in Wikipedia, namely, that are of interest to Wikipedians.

4.3. Disambiguation of anchor texts by category information

The wikification process includes a disambiguation step to find a correct destination article for each anchor text. Cross-lingual disambiguation is a more challenging task than monolingual disambiguation because the translation process reduces the impact of most of the effective disambiguation features, such as keyphraseness and context information. Among these features, category information (Cucerzan, 2007) is cross-lingually available because Wikipedia categories have inter-language links across languages: in our case, Japanese category “計算機科学” and English corresponding category “Computer science” are connected by an inter-language link. In our experiment, we simply estimated how category information contributes to disambiguation of anchor texts by eliminating destination articles that do not belong to categories subsumed by the specified category. For example, anchor candidate “キー (key)” in Table 1 has 19 possible destination articles, including “鍵 (key (lock))”, “キーボード (コンピュータ) (computer keyboard),” “調 (key (music))”, and so on. Among them, only four destination articles such as “キーボード (コンピュータ) (computer keyboard)” belong to the subsumed categories. There were 5.56 possible destination articles on average. After removing destination articles that did not belong to the subsumed categories, only 3.31 destination articles on average remained.

Disambiguation of anchor texts by category information might be strengthened by automatically specifying categories relevant to each patent application document. Using such categories would be more specific than using all categories subsumed by a manually specified category. Hence, it would be able to eliminate more articles irrelevant to an anchor text from the destination candidates than manually specifying a category. One promising method to make this possible is using classification tags for patents, such as tags showing IPC classes. To develop such a method it will be important to find a way to identify Wikipedia categories corresponding to each IPC class.

5. Conclusion and future work

In this paper, we demonstrated that cross-lingual patent wikification is promising for enriching patent application documents by increasing links to Wikipedia articles. The experiment we conducted indicated that more than 60% of important technical terms in patent application documents could be linked to Wikipedia articles with cross-lingual anchor text extraction, while 50% of them with monolingual one.

As a topic for future research, we plan to automatically determine the categories related to each patent application document for improving cross-lingual patent wikification. We also plan to tackle the problem of how to disambiguate anchor texts after applying a category-based method to them.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K16096. We express our appreciation to Professor Takehito Utsuro and Professor Mikio Yamamoto at the University of Tsukuba for providing a phrase table built from a patent parallel corpus in the NTCIR data set.
References


