Translation Using JAPIO Patent Corpora: JAPIO at WAT2016

Satoshi Kinoshita Tadaaki Oshio Tomoharu Mitsuhashi Terumasa Ehara[1]
Japan Patent Information Organization
{satoshi_kinoshita, t_oshio, t_mitsuhashi}@ japio.or.jp
eharate@ gmail.com

Abstract

Japan Patent Information Organization (JAPIO) participates in scientific paper subtask (ASPEC-EJ/CJ) and patent subtask (JPC-EJ/CJ/KJ) with phrase-based SMT systems which are trained with its own patent corpora. Using larger corpora than those prepared by the workshop organizer, we achieved higher BLEU scores than most participants in EJ and CJ translations of patent subtask, but in crowdsourcing evaluation, our EJ translation, which is best in all automatic evaluations, received a very poor score. In scientific paper subtask, our translations are given lower scores than most translations that are produced by translation engines trained with the in-domain corpora. But our scores are higher than those of general-purpose RBMTs and online services. Considering the result of crowdsourcing evaluation, it shows a possibility that CJ SMT system trained with a large patent corpus translates non-patent technical documents at a practical level.

1 Introduction

Japan Patent Information Organization (JAPIO) provides a patent information service named GPG-FX[2], which enables users to do cross-lingual information retrieval (CLIR) on patent documents by translating English and Chinese patents into Japanese and storing the translations in a full-text search engine.

For this purpose, we use a rule-based machine translation (RBMT) system and a phrase-based statistical machine translation (SMT) system for English-to-Japanese and Chinese-to-Japanese translation respectively. To improve translation quality, we have been collecting technical terms and building parallel corpora, and the current corpora sizes are 250 million sentence pairs for English-Japanese (EJ) and 100 million for Chinese-Japanese (CJ). We have also built a Korean-Japanese (KJ) corpus which contains about 5 million sentence pairs for adding Korean-to-Japanese translation to enable searching Korean patents as well.

The Japan Patent Office (JPO) and National Institute of Information and Communications Technology (NICT) have also built very large parallel corpora in patent domain. Their EJ, CJ and KJ corpora whose sizes are 350, 130 and 80 million sentence pairs are available at ALAGIN[3] for research purposes. Considering this trend, we think it important to make a research on a methodology to use very large parallel corpora for building a practical SMT system, as well as a research for creating a framework that can provide high automatic evaluation scores using a corpus of small size. This consideration led us to attend the 3rd Workshop on Asian Translation (WAT2016) (Nakazawa et al, 2016) in order to confirm the effectiveness of our own large patent parallel corpora.

[1] Guest researcher

[2] http://www.japio.or.jp/service/service05.html

[3] https://alaginrc.nict.go.jp/

Proceedings of the 3rd Workshop on Asian Translation,
pages 133–138, Osaka, Japan, December 11-17 2016.

2 Systems

We used two SMT systems to produce translations for the workshop.

The first one is a phrase-based SMT toolkit licensed by NICT (Utiyama and Sumita, 2014). It includes a pre-ordering module, which changes word order of English and Chinese source sentences into a head-final manner to improve translation into Japanese. We used it for EJ and CJ translation.

The second is Moses (Koehn et al., 2007), which is used for KJ translation. We used no morphological analyser for tokenizing Korean sentences. Instead, we simply decompose them into tokens which consist of only one Hangul character, and add a special token which represents a blank. To tokenize Japanese sentences, we used juman version 7.0 (Kurohashi et al., 1994). Distortion limit is set to 0 when the decoder runs whatever MERT estimates because of linguistic similarity between Korean and Japanese.

In addition, we include the following post-editing functions depending on translation directions and subtasks:

- Changing Japanese punctuation marks "、" to commas, and some patent-specific expressions to what are common in scientific papers (ASPEC-EJ/CJ)
- Recovering lowercased out-of-vocabularies (OOVs) to their original spellings (EJ)
- Balancing unbalanced parentheses (KJ) (Ehara, 2015)

3 Corpora and Training of SMT

Our patent parallel corpora, hereafter JAPIO corpora, are built automatically from pairs of patent specifications called "patent families," which typically consists of an original document in one language and its translations in other languages. Sentence alignment is performed by an alignment tool licensed by NICT (Utiyama and Isahara, 2007).

When we decided to attend WAT2016, we had EJ and CJ SMT systems which were built for research purposes, whose maximum training corpus sizes were 20 and 49 million sentence pairs respectively, and we thought what we had to do was to translate test sets except for KJ patent subtask. However, we found that about 24% and 55% of sentences in the patent subtask test sets were involved in JAPIO corpora for EJ and CJ respectively[4]. Although we built our corpora independently from those of Japan Patent Office corpora (JPC), a similarity to use patent-family documents may have led the situation. In order to make our submission to WAT more meaningful, we determined that we would publish automatic evaluation results of translations by the above SMT systems, but would not ask for human evaluation, and started retraining of SMT systems with corpora which exclude sentences in JPC test sets.

By the deadline of submission, we finished training CJ SMT with 4 million sentence pairs. As for EJ SMT, we finished training with 5 million sentence pairs, and added 1 million sentences of JPC corpus for an extra result.

In the case of KJ patent subtask, JAPIO corpus contains only 0.6% of JPC test set sentences, which are smaller than that of JPC training set[4]. So we used our KJ corpus without removing sentences contained in JPC test set. One thing we'd better to mention here is that 2.6 million sentence pairs out of 5 million, and 2.3 million out of 6 million, were filtered by corpus-cleaning of Moses because of limitation for maximum number of tokens per sentence. This is because we tokenized Korean sentences not by morphological analysis but based on Hangul characters.

As for scientific paper subtask, we did not use ASPEC corpus (Nakazawa et al, 2016), which is provided for this task, but used only our patent corpus. Since ASPEC corpus and our corpus were built from different data sources, our EJ corpus contains no sentence of ASPEC-EJ test set, and CJ corpus contains only 2 sentences of CJ test set. Therefore, we used SMT systems which are trained with our original corpora. For a submission of EJ translations, we chose a result translated by an SMT which was trained with 10 million sentence pairs because its BLEU score was higher than that with 20 million sentence pairs.

Finally, all development sets used in MERT process are from our corpora, whose sizes are about 3,000, 5,000 and 1,900 for EJ, CJ and KJ respectively.

[4] JPC training sets contain 1.1%, 2.3% and 1.0% of sentences of EJ, CJ and KJ test sets respectively.

4 Results

Table 1 shows official evaluation results for our submissions[5].

On patent subtask, the result shows that using a larger corpus does not necessarily lead to a higher BLEU score. Translation with our 5 million corpus achieved a lower score than that with 1 million JPC corpus in JPC-KJ subtask although training with our corpora achieved higher BLEU scores than most of the participants in EJ and CJ translations. In addition, those for KJ translations are lower than many of the task participants although our corpus is much larger than JPC corpus. In crowdsourcing evaluation, our EJ result, which received best scores in all automatic evaluations among the results submitted for human evaluation, received a poorer score than we expected.

On scientific paper subtask, we cannot achieve scores which are comparable with scores of translations that are produced by translation engines trained with ASPEC corpora. However, our scores are higher than those of general-purpose RBMTs and online services. Considering the result of crowdsourcing evaluation, this suggests a possibility that a CJ SMT system trained with a large patent corpus translates non-patent technical documents at a practical level even though the used resource is out of domain.

#	Subtask	System	Corpus	Size (million)	BLEU	RIEBS	AMFM	HUMAN
1	JPC-EJ	JAPIO-a	JAPIO-test	5	45.57	0.851376	0.747910	17.750
2		JAPIO-b	JAPIO-test+JPC	6	47.79	0.859139	0.762850	26.750
3		JAPIO-c	JAPIO	5	50.28	0.859957	0.768690	—
4		JAPIO-d	JPC	1	38.59	0.839141	0.733020	—
5	JPC-CJ	JAPIO-a	JAPIO-test	3	43.87	0.833586	0.748330	43.500
6		JAPIO-b	JAPIO-test	4	44.32	0.834959	0.751200	46.250
7		JAPIO-c	JAPIO	49	58.66	0.868027	0.808090	—
8		JAPIO-d	JPC	1	39.29	0.820339	0.733300	—
9	JPC-KJ	JAPIO-a	JAPIO	5	68.62	0.938474	0.858190	-9.000
10		JAPIO-b	JAPIO+JPC	6	70.32	0.942137	0.863660	17.500
11		JAPIO-c	JPC	1	69.10	0.940367	0.859790	—
12	ASPEC-EJ	JAPIO-a	JAPIO	10	20.52	0.723467	0.660790	4.250
13		Online x	—	—	18.28	0.706639	0.677020	49.750
14		RBMT x	—	—	13.18	0.671958	—	—
15	ASPEC-CJ	JAPIO-a	JAPIO	49	26.24	0.790553	0.696770	16.500
16		Online x	—	—	11.56	0.589802	0.659540	-51.250
17		RBMT x	—	—	19.24	0.741665	—	—

Table 1: Official Evaluation Results

5 Discussion

5.1 Error Analysis of Patent Subtask

We analysed errors which are involved in translations of EJ, CJ and KJ patent subtask by comparing our translations with the given references. Analysed translations are the first 200 sentences of each test set, and are from translation #1(EJ), #6(CJ) and #9(KJ) in Table 1.

Table 2 shows the result. Numbers of mistranslation for content words are comparable although that of KJ is less than those of EJ and CJ. This type of error can only be resolved by adding translation examples to a training corpus. Other errors which are critical in EJ and CJ translation are mistranslation

[5] Scores of BLEU, RIEBS and AMFM in the table are those calculated with tokens segmented by juman. Evaluation results of an online service and RBMT systems are also listed for the sake of comparison in ASPEC-EJ and CJ subtasks.

of functional words and errors of part of speech (POS) and word order which seem due to errors in pre-ordering. This suggests that improvement of pre-ordering might be more effective to better translation quality than increasing parallel corpora for EJ and CJ translation, which seems compatible with a future work derived from an analysis of crowdsourcing evaluation, which shows a poor correlation between automatic and human evaluations in JPC-EJ, and JPO adequacy evaluation.

Error Type	EJ	CJ	KJ
Insertion	0	0	6
Deletion	4	9	1
OOV	6	9	2
Mistranslation(content word)	44	41	30
Mistranslation(functional word)	21	51	0
Pre-ordering	33	45	0
Other	6	7	2
Total	114	162	41

Table 2: Errors of patent subtask

5.2 Error Analysis of Scientific Paper Subtask

We analysed errors of translations in EJ and CJ scientific paper subtask from a viewpoint of domain adaptation. As described in section 3, what we used to train SMTs for this subtask are not ASPEC corpora but our patent corpora. Therefore, some of the mistranslations must be recognized as domain-specific errors. That is, words and expressions which appear frequently in scientific papers but seldom in patent documents must have tendencies to be mistranslated. Similarly, what appear frequently in patents but seldom in papers and what appear frequently in both domains but are often translated differently might also be mistranslated. We call these types of error as “type A” and “type B” error respectively. Table 3 shows their examples. In example 1, word “paper(academic article)” is mistranslated as “紙(physical paper),” which can be categorized as a type A error. Word “discusses” is mistranslated as “開示されている(disclose),” which can be categorized as a type B error. Example 2 shows another type B error, where word “我们(we)” is mistranslated as “本発明者ら(the inventors).”

Example 1	
Source	This paper discusses the mechanism of the heat return reaction.
Reference	熱戻り反応の機構を議論した
MT	この紙は，熱戻り反応の機構が開示されている。

Example 2	
Source	由此，伴随中国乡镇向城市化发展而增加的环境负荷，我们从大气污染角度着手并利用环境库兹涅茨曲线进行环境分析，再将与他发达国家的城市环境相比较，探讨了降低环境负荷的可能性。
Reference	このような状況から,中国の都市化に伴う環境負荷の増大について大気汚染に着目して環境クズネック曲線を用いて分析し,先進諸国の都市の動向と比較して,その環境負荷低減策の可能性について考察した。
MT	これにより，中国タウン都市化発展に増加した環境負荷を伴って，本発明者らは，大気汚染の観点から着手し，利用環境库兹涅茨曲線環境分析を行い，さらに，彼と先進国の都市環境と比較して，環境負荷を低減する可能性を検討した。

Table 3: Examples of translation error

Table 4 shows the result. We analysed mistranslated content words from 200 translations of ASPEC-EJ/CJ test sets.

Error Type	EJ	CJ
Type A Error	9 (3.7%)	0
Type B Error	5 (2.1%)	6 (2.7%)
OOV	68 (28.2%)	48 (21.7%)
Others	159 (66.0%)	167 (75.6%)
Total	241	221

Table 4: Errors in translations of scientific paper subtask

As the table shows, domain-specific errors, that is type A and B errors, are only 5.8% in EJ translation and 2.7% in CJ. Rest of errors are related to OOVs or errors which come from the statistical characteristics of training corpora. As in the analysis of 5.1, OOVs can only be resolved by adding translation examples to a training corpus. Some of the other type of errors might, however, be resolved by modifying data in patent corpora. One idea is to remove numbering expressions such as 1 or 1a in "XX system 1" or "YY device 1a." Because usage of numbering in scientific papers is limited compared to that in patent documents, removing uncommon numbering expressions in scientific papers from patent corpora may generate better translation and language models for the domain.

6 Conclusion

In this paper, we described systems and corpora of Team JAPIO for submitting translations to WAT2016. The biggest feature of our experimental settings is that we use larger patent corpora than those prepared by the workshop organizer. We used 3 to 6 million sentence pairs for training SMT systems for patent subtask (JPC-EJ/CJ/KJ) and 10 and 49 million sentence pairs for scientific paper subtask (ASPEC-EJ/CJ). Using the corpora, we achieved higher BLEU scores than most participants in EJ and CJ translations of patent subtask. In crowdsourcing evaluation, however, our EJ translation, which is best in all automatic evaluations, received a very poor score.

In scientific paper subtask, our translations are given lower scores than most translations that are produced by translation engines trained with the in-domain corpora. But our scores are higher than those of general-purpose RBMTs and online services. Considering the result of crowdsourcing evaluation, it shows a possibility that a CJ SMT system trained with a large patent corpus translates non-patent technical documents at a practical level.

References

Terumasa Ehara. 2015. System Combination of RBMT plus SPE and Preordering plus SMT. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015).*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28.

Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig, Sadao Kurohashi and Eiichiro Sumita. 2016. Overview of the 3rd Workshop on Asian Translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016).*

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi and Hitoshi Isahara. 2016. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proceedings of the 10th Conference on International Language Resources and Evaluation (LREC2016).*

Masao Utiyama and Hiroshi Isahara. 2007. A Japanese-English Patent Parallel Corpus. In *MT summit XI*, pages 475-482.

Masao Utiyama and Eiichiro Sumita. 2014. AAMT Nagao Award Memorial lecture. http://www2.nict.go.jp/astrec-att/member/mutiyama/pdf/AAMT2014.pdf