Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017

Satoshi Kinoshita Tadaaki Oshio Tomoharu Mitsuhashi

Japan Patent Information Organization
{satoshi_kinoshita, t_oshio, t_mitsuhashi} @ japio.or.jp

Abstract

Japan Patent Information Organization (Japio) participates in patent subtasks (JPC-EJ/JE/CJ/KJ) with phrase-based statistical machine translation (SMT) and neural machine translation (NMT) systems which are trained with its own patent corpora in addition to the subtask corpora provided by organizers of WAT2017. In EJ and CJ subtasks, SMT and NMT systems whose sizes of training corpora are about 50 million and 10 million sentence pairs respectively achieved comparable scores for automatic evaluations, but NMT systems were superior to SMT systems for both official and in-house human evaluations.

1 Introduction

Japan Patent Information Organization (Japio) provides a patent information service named GPG/FX[1], 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 phrase-based statistical machine translation (SMT) system for Chinese-to-Japanese translation, and are preparing to change an English-to-Japanese translation system from a rule-based machine translation (RBMT) system to an SMT system. To improve translation quality, we have been building technical term dictionaries and parallel corpora, and the current corpora sizes are 300 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 13 million sentence pairs for adding Korean-to-Japanese translation to enable searching Korean patents as well.

Our current concern is neural machine translation (NMT), which has been used practically in the field of patent translation since last year (WIPO, 2016). The new approach has been reported to produce better translations than SMT by training with a smaller corpus than SMT. Our translation results in the 4th Workshop on Asian Translation (WAT2017) (Nakazawa et al., 2017) show the same conclusion.

2 Systems

2.1 Base Systems

We used three MT tools to produce translations for the workshop; two are SMTs and the rest is an NMT. The SMT tools are a phrase-based SMT toolkit licensed by NICT (Utiyama and Sumita, 2014), and Moses (Koehn et al., 2007). The former is used for EJ and CJ translation because 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. The latter is used for KJ translation where pre-ordering is not necessary because of linguistic similarities between Korean and Japanese. We used morphological analyzers mecab-ko[2] and juman version 7.0 (Kurohashi et al., 1994) for tokenizing Korean and Japanese respectively.

A toolkit we used for NMT is OpenNMT[3], whose default setting provides an attention-based NMT model which consists of a 2-layer LSTM with 500 hidden units.

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

[2] https://bitbucket.org/eunjeon/mecab-ko/

[3] http://opennmt.net/

Proceedings of the 4th Workshop on Asian Translation, pages 140–145,
Taipei, Taiwan, November 27, 2017. ©2017 AFNLP

Two major difference between its default and our experimental settings are: 1) a deep bidirectional recurrent neural network (DBRNN) is used instead of a standard recurrent neural network (RNN). 2) The value 100,000 is used as a vocabulary size if a size of training corpus is equal or more than 3 million sentence pairs whereas 50,000, a default value, is used for a smaller training corpus than that. For tokenizing corpus texts, Moses tokenizer, juman and kytea[4] are used to tokenize English, Japanese and Chinese, respectively.

2.2 Treatment of Out of Vocabulary

One of major problems to use an NMT system for translating patent documents, which include a large number of technical terms, is a limited number of vocabulary size. To solve the problem, various approaches have been proposed, such as using a model based on not words but characters or subwords, and a method to replace technical terms in a training corpus and source sentences with technical term tokens (Sennrich et al., 2015; Long et al., 2016).

We propose a method to extract out of vocabulary (OOV) words by the attention mechanism of OpenNMT and translate them with another NMT which has a character-based model. For EJ/JE/CJ NMT systems, such character-based models are trained by using a size of 1 million technical terms extracted from our technical term dictionaries. Japanese and Chinese words of the extracted dictionary entries are tokenized on a character basis while English words are divided by byte pair encoding. In translation, OpenNMT can output source tokens for unknown words instead of <unk> symbols by using attention weights[5]. They are translated by the above-mentioned character-based NMT systems and replaced with their translations.

2.3 Pre- and Post-processing

We include the following pre- and post-editing functions depending on translation systems and directions:

- Recovering lowercased out-of-vocabularies (OOVs) to their original spellings (EJ-SMT)
- Balancing unbalanced parentheses (KJ)
- Splitting long sentences into shorter ones (CJ-NMT)

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 consist of an original document in one language and its translations in other languages. Sentence alignment is performed by 2 alignment tools: one is a tool licensed by NICT (Utiyama and Isahara, 2007), and the other is E_align[6].

In patent subtask of WAT2016, we achieved the highest BLEU score 58.66 in JPC-CJ with an SMT system trained with about 49 million sentence pairs. However, we found that about 55% of sentences in the test sets were involved in the training corpus[7]. Although we built our corpora independently from those of Japan Patent Office corpora (JPC), methodological 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 its automatic evaluation result, but submitted another translation which was produced by an SMT which was trained by using a corpus of 4 million sentence pairs with no sentence in the test set. This year, we trained an SMT with a corpus of the 49 million sentence pairs where test set sentences are removed from the original corpus by using publication numbers embedded as data IDs in the JPC corpora. To train NMTs, we used the JPC-CJ corpus as a baseline, and added up to 9 million sentence pairs extracted from the above corpus.

Corpus for EJ translation was prepared as in the case for CJ. A corpus that we used for training an SMT for our service contained 24% of test set sentences. Therefore, we published the result, but did not request human evaluation. What we asked for human evaluation was a result which was translated by an SMT that was trained with a corpus without sentences in the test set. Similarly, to train NMTs, we used the JPC-EJ corpus as a baseline, and added up to 11 million sentence pairs from the corpus prepared for the above SMT.

In the case of KJ patent subtask, we used 8 million sentences pairs from our corpus in addition to

[4] http://www.phontron.com/kytea/

[5] To distinguish unknown words in target tokens, we modified source codes of OpenNMT to add a tag to them. The latest version of OpenNMT has a similar function.

[6] http://www.gsk.or.jp/catalog/gsk2017-a/

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

Subtask	#	DataID	System	Corpus Size (million)	Use official corpus	Automatic			Human	
						BLEU	RIBES	AMFM	pairwise	JPO adq.
JPC-EJ	1-1	1330*	SMT (PBSMT with preordering)	1	Yes	38.59	0.839141	0.733020	—	—
	1-2	1445	SMT (PBSMT with preordering)	100**	No	55.55	0.875667	0.802260	—	—
	1-3	1462	SMT (PBSMT with preordering)	50	No	**51.79**	0.864038	**0.781150**	41.000	—
	1-4	1451	NMT	1	Yes	44.69	0.864568	0.746720	—	—
	1-5	1453	NMT	5	Yes	48.39	0.880215	0.767720	—	—
	1-6	1454	NMT (Combination of 4 NMTs)	12	Yes	50.27	**0.886403**	0.776790	**56.250**	4.75
JPC-JE	2-1	1455	NMT	1	Yes	44.07	0.863385	0.699930	—	—
	2-2	1578	NMT	5	Yes	48.08	0.873093	0.715560	67.000	—
	2-3	1574	NMT (Combination of 3 NMTs)	11	Yes	**49.00**	**0.878298**	**0.724710**	**68.500**	4.79
JPC-CJ	3-1	1329*	SMT (PBSMT with preordering)	1	Yes	39.29	0.820339	0.733300	—	—
	3-2	1161*	SMT (PBSMT with preordering)	49**	No	58.66	0.868027	0.808090	—	—
	3-3	1447	SMT (PBSMT with preordering)	49	No	**50.52**	0.847793	0.774660	60.500	—
	3-4	1458	NMT	1	Yes	45.07	0.859883	0.754970	—	—
	3-5	1482	NMT	5	Yes	49.51	0.872625	0.777460	—	—
	3-6	1484	NMT (Combination of 3 NMTs)	10	Yes	50.06	**0.875398**	**0.779420**	**80.250**	4.46
JPC-KJ	4-1	1331*	SMT (Character-based PBSMT)	1	Yes	69.10	0.940367	0.859790	—	—
	4-2	1448	SMT (Word-based PBSMT)	9	Yes	**73.00**	0.946880	0.872510	**48.750**	4.84
	4-3	1449	SMT (Character-based PBSMT)	9	Yes	71.97	0.944435	0.868170	—	—
	4-4	1450	SMT (Combination of 2 SMTs)	9	Yes	**73.00**	**0.946985**	**0.873200**	48.500	—

* Submissions with '*' of their DataID are those submitted for WAT2016

** Traing data whose size are given '**' include some sentences of test set.

Table 1: Official Evaluation Results

DataID	Team	Method	Other Resources	Automatic			Human	
				BLEU	RIBES	AMFM	pairwise	JPO adq.
1407	Team-A	NMT	No	44.63	0.866722	0.747770	**60.000**	4.63
1406	Team-A	NMT	No	44.44	0.860998	0.747050	58.250	—
1454	Japio	NMT	Yes	50.27	**0.886403**	0.776790	56.250	**4.75**
1470	Team-B	NMT	No	38.91	0.845815	0.734010	49.500	4.40
1339	Team-C	NMT	Yes	50.60	0.879382	0.770480	48.500	—
1462	Japio	SMT	Yes	**51.79**	0.864038	**0.781150**	41.000	—

Table 2: Official Human Evaluation Results for JPC-EJ subtask

JPC-KJ corpus. By using 9 million sentence corpus, we trained two types of SMTs: one trained with a corpus that is tokenized on a character-basis, while the other with a corpus that is tokenized by mecab-ko.

4 System Combination

It was reported that an NMT system achieved better translation by ensembling multiple models (Sennrich et al., 2016). Because OpenNMT, which we use for our NMT systems, does not provide the function, we combined translations from multiple NMT systems as follows, in addition to using character-based NMTs to resolve OOVs.

(1) Combinations of NMTs that are trained for technical domains

For JPC-EJ, we trained 4 NMTs by using corpora whose data are selected based on its domain label, namely C, E, M and P, which are also given to test set sentences. They are used in addition to JPC-EJ corpus. In translating test set sentences, an appropriate NMT is used according to the domain.

(2) Usage of scores by OpenNMT

For JPC-JE and CJ, we could not complete training which was needed to make 4 domain models as we did for JPC-EJ by the submission deadline. Instead, we used scores which are given to each translation by OpenNMT, and selected a translation with the highest score.

In JPC-KJ, we chose a translation by a character-based SMT when a translation by a word-based SMT contains an OOV with at least one Hangul character.

Subtask	DataID	System	Corpus size (million)	BLEU
JPC-EJ	1462	SMT (PBSMT with preordering)	50	51.79
	1454	NMT (Combination of 4 NMTs)	12	50.27
JPC-CJ	1447	SMT (PBSMT with preordering)	49	50.52
	1484	NMT (Combination of 3 NMTs)	10	50.06

Table 3: Translations for in-house evaluations

	EJ	CJ
SMT is better	24	32
NMT is better	32	68
comparable	144	100

Table 4: Result of pairwise evaluations

Error Type	EJ		CJ	
	SMT	NMT	SMT	NMT
Insertion	10	2	8	4
Deletion	21	6	14	21
Mistranslation	26	31	29	54
Others	19	6	75	13
Total	76	45	126	92

Table 5: Errors of SMT and NMT for JPC-EJ/CJ

5 Results

Table 1 shows official evaluation results for our submissions[8]. In JPC-EJ and CJ, translations by SMTs trained with about 50 million sentence pairs are given comparable scores for automatic evaluation with those by NMTs trained with about 10 million sentence pairs. Human pairwise evaluation, however, gives much higher scores to translations by NMTs than those by SMTs.

Table 2 shows evaluation results for high-ranked submissions of JPC-EJ this year[9]. What is the most interesting for us is that a translation by an SMT which is given the highest scores for BLEU and AMFM is given a lower human evaluation score than those by NMTs trained with only 1 million sentence pairs. Furthermore, comparing results between NMT systems, a result whose DataID is 1339 and is given the highest BLEU score and a result whose DataID is 1454 and is given the highest RIBES and AMFM scores are given lower pairwise evaluation scores than those of Team-A, which are apparently given lower BLEU and AMFM scores than the formers. These results support previous findings that there is no correlation between automatic and human evaluations.

6 Discussion

To recognize a difference of translation quality between SMT and NMT systems, we conducted two kinds of human evaluations independently from the official evaluation: one is pairwise evaluation, and the other is an error analysis. We used the same sentences used for JPO adequacy evaluation in WAT2017, and one evaluator conducted both evaluations. Table 3 shows translations used for the in-house evaluation.

6.1 Pairwise Evaluation

We conducted pairwise evaluation based on adequacy. When evaluating a translation, which translation is better is determined based on how much of the meaning of a source sentence is expressed in its translation. Taking JPO adequacy into account, insertion and deletion of conjunctions which are considered not to convey important information are ignored if translations are grammatical. Fluency is also ignored.

Table 4 shows the result. In both EJ and CJ, NMTs are evaluated to produce more better translations than SMTs. The tendency is remarkable in

[8] Scores of BLEU, RIBES and AMFM for JPC-EJ/CJ/KJ are those calculated with tokens segmented by juman.

[9] A translation result whose DataID is 1339 was not evaluated last year because it was submitted after the deadline for human evaluation.

CJ, which is consistent with the official evaluation result shown in Table 1.

6.2 Error Analysis

In the error analysis, translation errors are categorized into the following 4 categories:

- Insertion
- Deletion
- Mistranslation
- Others (such as grammatical errors)

Note that insertions and deletions which are ignored in the pairwise evaluation are counted in this analysis.

Table 5 shows the result. On the whole, number of errors of the SMT translations is larger than that of NMT in both EJ and CJ. This is consistent with the results of official and in-house pairwise evaluations.

Number of mistranslations of NMT translations is however larger than that of SMT in both EJ and CJ. The reason we think is that technical terms of low frequencies are not properly translated by the following two reasons:

- A corpus that was used for training NMTs is much smaller than that for SMTs.
- In training NMTs, a vocabulary is limited by a pre-defined vocabulary size or vocabulary set, and words out of the involved vocabulary cannot be translated.

A character-based NMT which is used to resolve the OOV problem does not work as we expected. In addition, deletion errors of NMT are smaller than SMT in EJ, but are larger in CJ.

What is the most characteristic in the error analysis is that about 60% of errors of CJ SMT are categorized as "Others." This might be caused by low precision of preordering due to the difficulty of Chinese syntactic analysis.

7 Conclusion

In this paper, we described systems and corpora of Team Japio for submitting translations to WAT2017. To show potential of SMT and NMT in patent translation, we participated in patent subtasks (JPC-EJ/JE/CJ/KJ) with systems which are trained with its own patent corpora in addition to the corpora provided by organizers of WAT2017. The result shows that SMT and NMT systems whose sizes of training corpora are about 50 million and 10 million sentence pairs respectively achieved comparable scores for automatic evaluations in EJ and CJ subtasks. NMT systems were, however, superior to SMT systems for both official and in-house human evaluations.

References

Satoshi Kinoshita, Tadaaki Oshio, Tomoharu Mitsuhashi, Terumasa Ehara. 2016. Translation Using JAPIO Patent Corpora: JAPIO at WAT2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 133-138.

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.

Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto. 2016. Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 47-57.

Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig, Sadao Kurohashi and Eiichiro Sumita. 2016a. Overview of the 3rd Workshop on Asian Translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 1-46.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Graham Neubig, Hideto Kazawa, Yusuke Oda, Jun Harashima and Sadao Kurohashi. 2017. Overview of the 4th Workshop on Asian Translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017).*

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

Rico Sennrich, Barry Haddow and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *Proceedings of 54th ACL*, pages 1715-1725.

Rico Sennrich, Barry Haddow and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First*

Conference on Machine Translation, Volume 2: Shared Task Papers, pages 371–376.

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

WIPO. 2016. WIPO Develops Cutting-Edge Translation Tool For Patent Documents. http://www.wipo.int/pressroom/en/articles/2016/article_0014.html