Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation

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

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

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

Abstract

Neural machine translation (NMT), a new approach to machine translation, has achieved promising results comparable to those of traditional approaches such as statistical machine translation (SMT). Despite its recent success, NMT cannot handle a larger vocabulary because training complexity and decoding complexity proportionally increase with the number of target words. This problem becomes even more serious when translating patent documents, which contain many technical terms that are observed infrequently. In NMTs, words that are out of vocabulary are represented by a single unknown token. In this paper, we propose a method that enables NMT to translate patent sentences comprising a large vocabulary of technical terms. We train an NMT system on bilingual data wherein technical terms are replaced with technical term tokens; this allows it to translate most of the source sentences except technical terms. Further, we use it as a decoder to translate source sentences with technical term tokens and replace the tokens with technical term translations using SMT. We also use it to rerank the 1,000-best SMT translations on the basis of the average of the SMT score and that of the NMT rescoring of the translated sentences with technical term tokens. Our experiments on Japanese-Chinese patent sentences show that the proposed NMT system achieves a substantial improvement of up to 3.1 BLEU points and 2.3 RIBES points over traditional SMT systems and an improvement of approximately 0.6 BLEU points and 0.8 RIBES points over an equivalent NMT system without our proposed technique.

1 Introduction

Neural machine translation (NMT), a new approach to solving machine translation, has achieved promising results (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015; Jean et al., 2014; Luong et al., 2015a; Luong et al., 2015b). An NMT system builds a simple large neural network that reads the entire input source sentence and generates an output translation. The entire neural network is jointly trained to maximize the conditional probability of a correct translation of a source sentence with a bilingual corpus. Although NMT offers many advantages over traditional phrase-based approaches, such as a small memory footprint and simple decoder implementation, conventional NMT is limited when it comes to larger vocabularies. This is because the training complexity and decoding complexity proportionally increase with the number of target words. Words that are out of vocabulary are represented by a single unknown token in translations, as illustrated in Figure 1. The problem becomes more serious when translating patent documents, which contain several newly introduced technical terms.

There have been a number of related studies that address the vocabulary limitation of NMT systems. Jean el al. (2014) provided an efficient approximation to the softmax to accommodate a very large vocabulary in an NMT system. Luong et al. (2015b) proposed annotating the occurrences of a target unknown word token with positional information to track its alignments, after which they replace the tokens with their translations using simple word dictionary lookup or identity copy. Li et al. (2016) proposed to replace out-of-vocabulary words with similar in-vocabulary words based on a similarity model learnt from monolingual data. Sennrich et al. (2016) introduced an effective approach based on encoding rare and unknown words as sequences of subword units. Luong and Manning (2016) provided a character-level

Proceedings of the 3rd Workshop on Asian Translation,
pages 47–57, Osaka, Japan, December 11-17 2016.

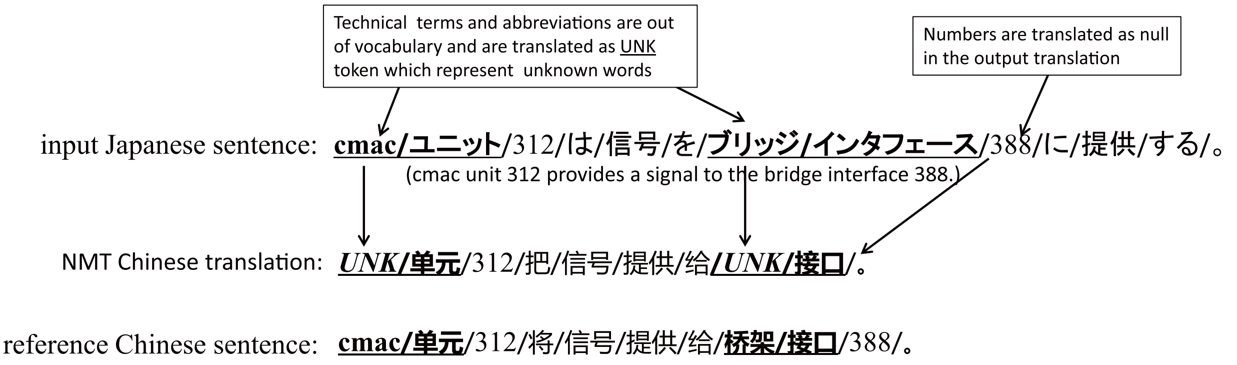

Technical terms and abbreviations are out of vocabulary and are translated as UNK token which represent unknown words
Numbers are translated as null in the output translation
input Japanese sentence: cmac/ユニット/312/は/信号/を/ブリッジ/インタフェース/388/に/提供/する/。
(cmac unit 312 provides a signal to the bridge interface 388.)
NMT Chinese translation: UNK/单元/312/把/信号/提供/给/UNK/接口/。
reference Chinese sentence: cmac/单元/312/将/信号/提供/给/桥架/接口/388/。

Figure 1: Example of translation errors when translating patent sentences with technical terms using NMT

and word-level hybrid NMT model to achieve an open vocabulary, and Costa-jussà and Fonollosa (2016) proposed a NMT system based on character-based embeddings.

However, these previous approaches have limitations when translating patent sentences. This is because their methods only focus on addressing the problem of unknown words even though the words are parts of technical terms. It is obvious that a technical term should be considered as one word that comprises components that always have different meanings and translations when they are used alone. An example is shown in Figure1, wherein Japanese word "ブリッジ"(bridge) should be translated to Chinese word "桥架" when included in technical term "bridge interface"; however, it is always translated as "桥".

In this paper, we propose a method that enables NMT to translate patent sentences with a large vocabulary of technical terms. We use an NMT model similar to that used by Sutskever et al. (2014), which uses a deep long short-term memories (LSTM) (Hochreiter and Schmidhuber, 1997) to encode the input sentence and a separate deep LSTM to output the translation. We train the NMT model on a bilingual corpus in which the technical terms are replaced with technical term tokens; this allows it to translate most of the source sentences except technical terms. Similar to Sutskever et al. (2014), we use it as a decoder to translate source sentences with technical term tokens and replace the tokens with technical term translations using statistical machine translation (SMT). We also use it to rerank the 1,000-best SMT translations on the basis of the average of the SMT and NMT scores of the translated sentences that have been rescored with the technical term tokens. Our experiments on Japanese-Chinese patent sentences show that our proposed NMT system achieves a substantial improvement of up to 3.1 BLEU points and 2.3 RIBES points over a traditional SMT system and an improvement of approximately 0.6 BLEU points and 0.8 RIBES points over an equivalent NMT system without our proposed technique.

2 Japanese-Chinese Patent Documents

Japanese-Chinese parallel patent documents were collected from the Japanese patent documents published by the Japanese Patent Office (JPO) during 2004-2012 and the Chinese patent documents published by the State Intellectual Property Office of the People's Republic of China (SIPO) during 2005-2010. From the collected documents, we extracted 312,492 patent families, and the method of Utiyama and Isahara (2007) was applied[1] to the text of the extracted patent families to align the Japanese and Chinese sentences. The Japanese sentences were segmented into a sequence of morphemes using the Japanese morphological analyzer MeCab[2] with the morpheme lexicon IPAdic,[3] and the Chinese sentences were segmented into a sequence of words using the Chinese morphological analyzer Stanford Word Segment (Tseng et al., 2005) trained using the Chinese Penn Treebank. In this study, Japanese-Chinese parallel patent sentence pairs were ordered in descending order of sentence-alignment score and we used the topmost 2.8M pairs, whose Japanese sentences contain fewer than 40 morphemes and

[1] Herein, we used a Japanese-Chinese translation lexicon comprising around 170,000 Chinese entries.

[2] http://mecab.sourceforge.net/

[3] http://sourceforge.jp/projects/ipadic/

Chinese sentences contain fewer than 40 words.[4]

3 Neural Machine Translation (NMT)

NMT uses a single neural network trained jointly to maximize the translation performance (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015; Luong et al., 2015a). Given a source sentence $\boldsymbol{x} = (x_1, \ldots, x_N)$ and target sentence $\boldsymbol{y} = (y_1, \ldots, y_M)$, an NMT system uses a neural network to parameterize the conditional distributions

$$p(y_l \mid y_{<l}, \boldsymbol{x})$$

for $1 \leq l \leq M$. Consequently, it becomes possible to compute and maximize the log probability of the target sentence given the source sentence

$$\log p(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{l=1}^{M} \log p(y_l | y_{<l}, \boldsymbol{x}) \tag{1}$$

In this paper, we use an NMT model similar to that used by Sutskever et al. (2014). It uses two separate deep LSTMs to encode the input sequence and output the translation. The encoder, which is implemented as a recurrent neural network, reads the source sentence one word at a time and then encodes it into a large vector that represents the entire source sentence. The decoder, another recurrent neural network, generates a translation on the basis of the encoded vector one word at a time.

One important difference between our NMT model and the one used by Sutskever et al. (2014) is that we added an attention mechanism. Recently, Bahdanau et al. (2015) proposed an attention mechanism, a form of random access memory, to help NMT cope with long input sequences. Luong et al. (2015a) proposed an attention mechanism for different scoring functions in order to compare the source and target hidden states as well as different strategies for placing the attention. In this paper, we utilize the attention mechanism proposed by Bahdanau et al. (2015), wherein each output target word is predicted on the basis of not only a recurrent hidden state and the previously predicted word but also a context vector computed as the weighted sum of the hidden states.

4 NMT with a Large Technical Term Vocabulary

4.1 NMT Training after Replacing Technical Term Pairs with Tokens

Figure 2 illustrates the procedure of the training model with parallel patent sentence pairs, wherein technical terms are replaced with technical term tokens "TT_1", "TT_2", ...

In the step 1 of Figure 2, we align the Japanese technical terms, which are automatically extracted from the Japanese sentences, with their Chinese translations in the Chinese sentences.[5] Here, we introduce the following two steps to identify technical term pairs in the bilingual Japanese-Chinese corpus:

1. According to the approach proposed by Dong et al. (2015), we identify Japanese-Chinese technical term pairs using an SMT phrase translation table. Given a parallel sentence pair $\langle S_J, S_C \rangle$ containing a Japanese technical term t_J, the Chinese translation candidates collected from the phrase translation table are matched against the Chinese sentence S_C of the parallel sentence pair. Of those found in S_C, t_C with the largest translation probability $P(t_C \mid t_J)$ is selected, and the bilingual technical term pair $\langle t_J, t_C \rangle$ is identified.

[4] In this paper, we focus on the task of translating patent sentences with a large vocabulary of technical terms using the NMT system, where we ignore the translation task of patent sentences that are longer than 40 morphemes in Japanese side or longer than 40 words in Chinese side.

[5] In this work, we approximately regard all the Japanese compound nouns as Japanese technical terms. These Japanese compound nouns are automatically extracted by simply concatenating a sequence of morphemes whose parts of speech are either nouns, prefixes, suffixes, unknown words, numbers, or alphabetical characters. Here, morpheme sequences starting or ending with certain prefixes are inappropriate as Japanese technical terms and are excluded. The sequences that include symbols or numbers are also excluded. In Chinese side, on the other hand, we regard Chinese translations of extracted Japanese compound nouns as Chinese technical terms, where we do not regard other Chinese phrases as technical terms.

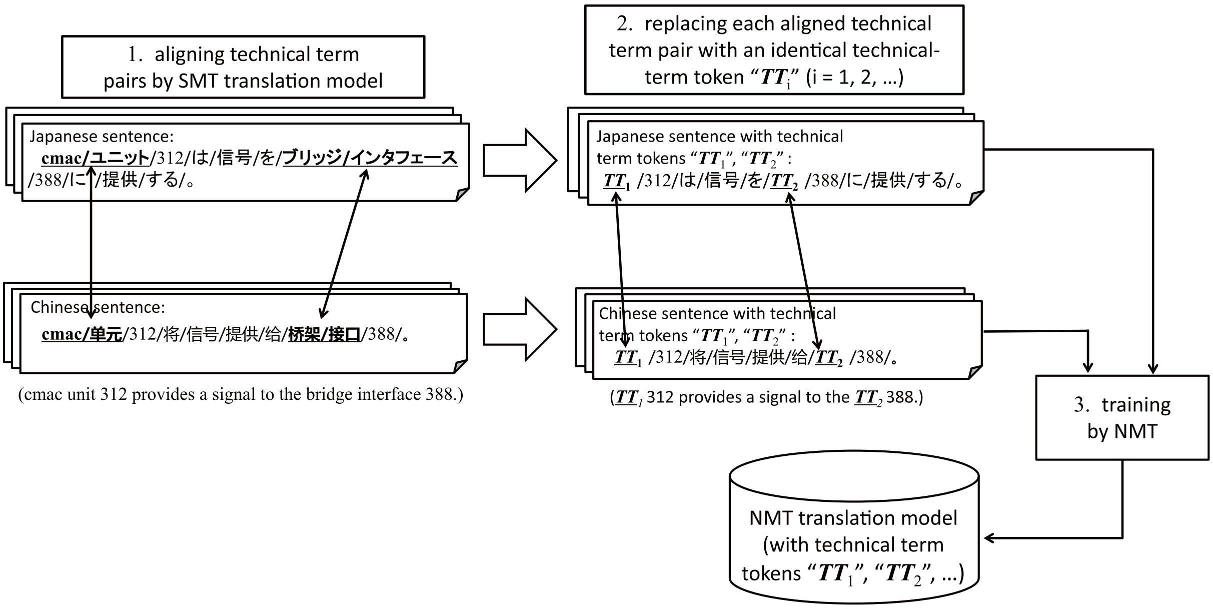

1. aligning technical term pairs by SMT translation model
2. replacing each aligned technical term pair with an identical technical-term token "TT_i" (i = 1, 2, …)
Japanese sentence:
cmac/ユニット/312/は/信号/を/ブリッジ/インタフェース/388/に/提供/する/。
Chinese sentence:
cmac/单元/312/将/信号/提供/给/桥架/接口/388/。
(cmac unit 312 provides a signal to the bridge interface 388.)
Japanese sentence with technical term tokens "TT_1", "TT_2" :
TT_1 /312/は/信号/を/TT_2 /388/に/提供/する/。
Chinese sentence with technical term tokens "TT_1", "TT_2" :
TT_1 /312/将/信号/提供/给/TT_2 /388/。
(TT_1 312 provides a signal to the TT_2 388.)
3. training by NMT
NMT translation model (with technical term tokens "TT_1", "TT_2", …)

Figure 2: NMT training after replacing technical term pairs with technical term tokens "TT_i" ($i = 1, 2, \ldots$)

2. For the Japanese technical terms whose Chinese translations are not included in the results of Step 1, we then use an approach based on SMT word alignment. Given a parallel sentence pair $\langle S_J, S_C \rangle$ containing a Japanese technical term t_J, a sequence of Chinese words is selected using SMT word alignment, and we use the Chinese translation t_C for the Japanese technical term t_J.[6]

As shown in the step 2 of Figure 2, in each of Japanese-Chinese parallel patent sentence pairs, occurrences of technical term pairs $\langle t_J^1, t_C^1 \rangle$, $\langle t_J^2, t_C^2 \rangle$, …, $\langle t_J^k, t_C^k \rangle$ are then replaced with technical term tokens $\langle TT_1, TT_1 \rangle$, $\langle TT_2, TT_2 \rangle$, …, $\langle TT_k, TT_k \rangle$. Technical term pairs $\langle t_J^1, t_C^1 \rangle$, $\langle t_J^2, t_C^2 \rangle$, …, $\langle t_J^k, t_C^k \rangle$ are numbered in the order of occurrence of Japanese technical terms t_J^i ($i = 1, 2, \ldots, k$) in each Japanese sentence S_J. Here, note that in all the parallel sentence pairs $\langle S_J, S_C \rangle$, technical term tokens "TT_1", "TT_2", … that are identical throughout all the parallel sentence pairs are used in this procedure. Therefore, for example, in all the Japanese patent sentences S_J, the Japanese technical term t_J^1 which appears earlier than other Japanese technical terms in S_J is replaced with TT_1. We then train the NMT system on a bilingual corpus, in which the technical term pairs is replaced by "TT_i" ($i = 1, 2, \ldots$) tokens, and obtain an NMT model in which the technical terms are represented as technical term tokens.[7]

4.2 NMT Decoding and SMT Technical Term Translation

Figure 3 illustrates the procedure for producing Chinese translations via decoding the Japanese sentence using the method proposed in this paper. In the step 1 of Figure 3, when given an input Japanese sentence, we first automatically extract the technical terms and replace them with the technical term tokens "TT_i" ($i = 1, 2, \ldots$). Consequently, we have an input sentence in which the technical term tokens "TT_i" ($i = 1, 2, \ldots$) represent the positions of the technical terms and a list of extracted Japanese technical terms. Next, as shown in the step 2-N of Figure 3, the source Japanese sentence with technical term tokens is translated using the NMT model trained according to the procedure described in Section 4.1, whereas the extracted Japanese technical terms are translated using an SMT phrase translation table in the step 2-S of Figure 3.[8] Finally, in the step 3, we replace the technical term tokens "TT_i" ($i = 1, 2, \ldots$)

[6] We discard discontinuous sequences and only use continuous ones.

[7] We treat the NMT system as a black box, and the strategy we present in this paper could be applied to any NMT system (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015; Luong et al., 2015a).

[8] We use the translation with the highest probability in the phrase translation table. When an input Japanese technical term has multiple translations with the same highest probability or has no translation in the phrase translation table, we apply a compositional translation generation approach, wherein Chinese translation is generated compositionally from the constituents

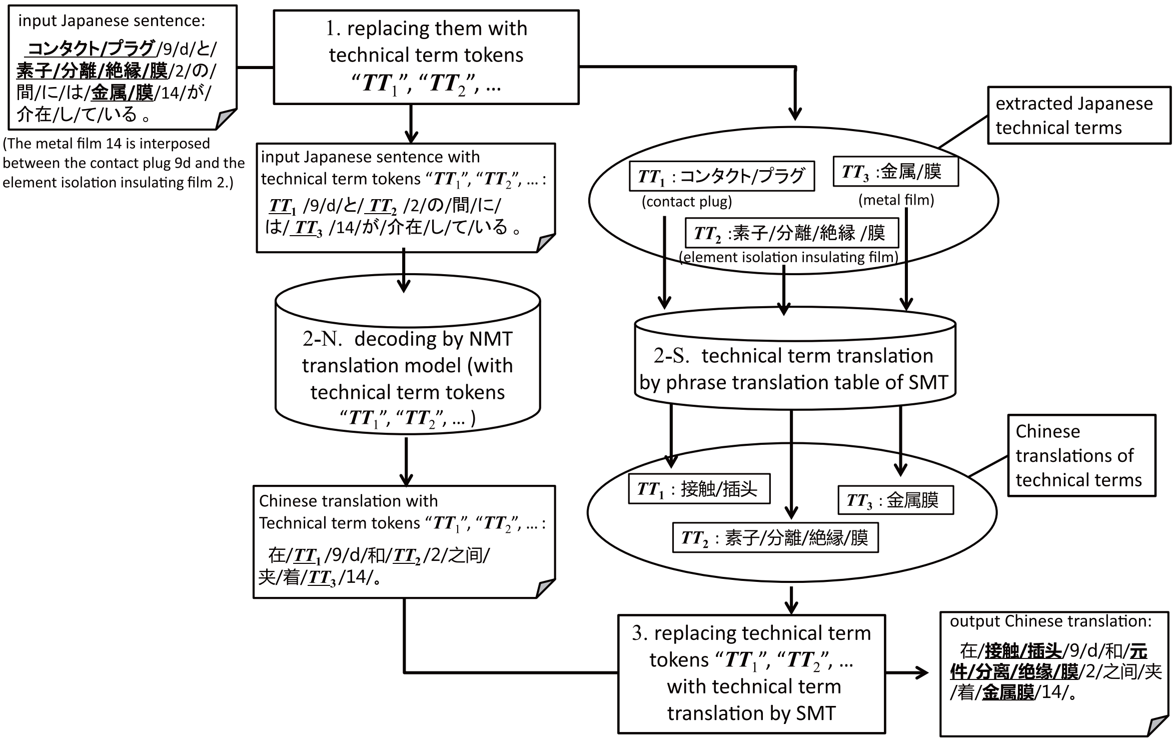

input Japanese sentence:
コンタクト/プラグ/9/d/と/素子/分離/絶縁/膜/2/の/間/に/は/金属/膜/14/が/介在/し/て/いる 。
(The metal film 14 is interposed between the contact plug 9d and the element isolation insulating film 2.)
1. replacing them with technical term tokens "TT1", "TT2", ...
input Japanese sentence with technical term tokens "TT1", "TT2", ... :
TT1 /9/d/と/ TT2 /2/の/間/に/は/ TT3 /14/が/介在/し/て/いる 。
extracted Japanese technical terms
TT1 : コンタクト/プラグ
(contact plug)
TT3 :金属/膜
(metal film)
TT2 :素子/分離/絶縁 /膜
(element isolation insulating film)
2-N. decoding by NMT translation model (with technical term tokens "TT1", "TT2", ...)
2-S. technical term translation by phrase translation table of SMT
Chinese translation with Technical term tokens "TT1", "TT2", ... :
在/TT1/9/d/和/TT2/2/之间/夹/着/TT3/14/。
Chinese translations of technical terms
TT1 : 接触/插头
TT3 : 金属膜
TT2 : 素子/分離/絶縁/膜
3. replacing technical term tokens "TT1", "TT2", ... with technical term translation by SMT
output Chinese translation:
在/接触/插头/9/d/和/元件/分离/绝缘/膜/2/之间/夹/着/金属膜/14/。

Figure 3: NMT decoding with technical term tokens "TT_i" ($i = 1, 2, \ldots$) and SMT technical term translation

of the sentence translation with SMT the technical term translations.

4.3 NMT Rescoring of 1,000-best SMT Translations

As shown in the step 1 of Figure 4, similar to the approach of NMT rescoring provided in Sutskever et al.(2014), we first obtain 1,000-best translation list of the given Japanese sentence using the SMT system. Next, in the step 2, we then replace the technical terms in the translation sentences with technical term tokens "TT_i" ($i = 1, 2, 3, \ldots$), which must be the same with the tokens of their source Japanese technical terms in the input Japanese sentence. The technique used for aligning Japanese technical terms with their Chinese translations is the same as that described in Section 4.1. In the step 3 of Figure 4, the 1,000-best translations, in which technical terms are represented as tokens, are rescored using the NMT model trained according to the procedure described in Section 4.1. Given a Japanese sentence S_J and its 1,000-best Chinese translations S_C^n ($n = 1, 2, \ldots,\ 1,000$) translated by the SMT system, NMT score of each translation sentence pair $\langle S_J, S_C^n \rangle$ is computed as the log probability $\log p(S_C^n \mid S_J)$ of Equation (1). Finally, we rerank the 1,000-best translation list on the basis of the average SMT and NMT scores and output the translation with the highest final score.

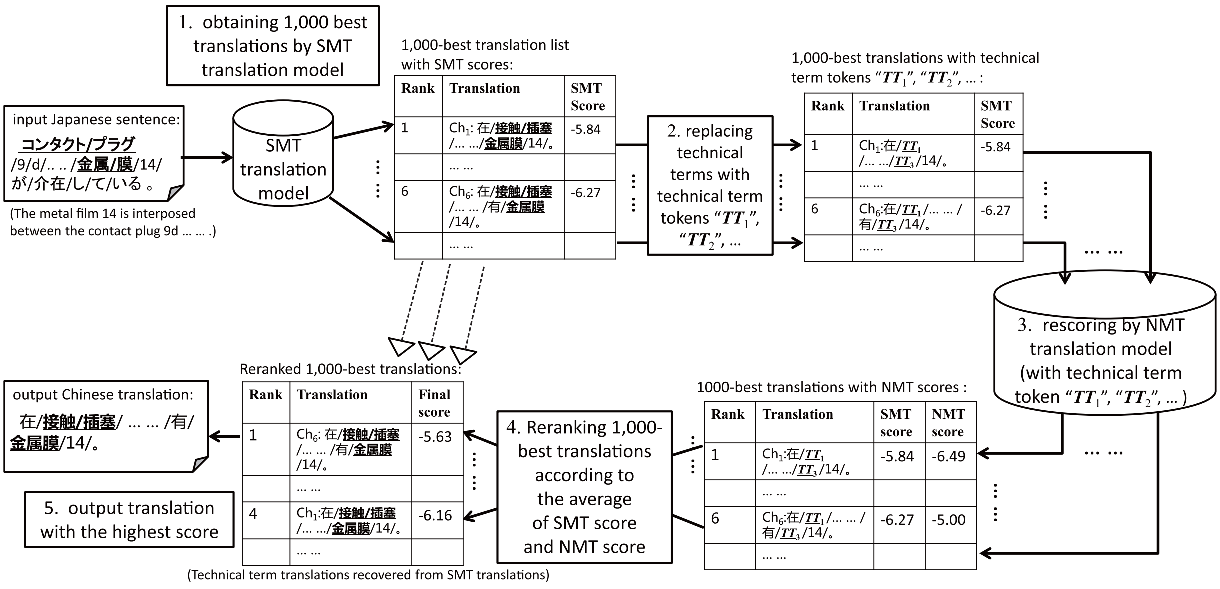
5 Evaluation

5.1 Training and Test Sets

We evaluated the effectiveness of the proposed NMT system in translating the Japanese-Chinese parallel patent sentences described in Section 2. Among the 2.8M parallel sentence pairs, we randomly extracted 1,000 sentence pairs for the test set and 1,000 sentence pairs for the development set; the remaining sentence pairs were used for the training set.

According to the procedure of Section 4.1, from the Japanese-Chinese sentence pairs of the training set, we collected 6.5M occurrences of technical term pairs, which are 1.3M types of technical term pairs with 800K unique types of Japanese technical terms and 1.0M unique types of Chinese technical terms.

of Japanese technical terms.


1. obtaining 1,000 best translations by SMT translation model
input Japanese sentence:
コンタクト/プラグ/9/d/.. .. /金属/膜/14/が/介在/し/て/いる 。
(The metal film 14 is interposed between the contact plug 9d … … .)
SMT translation model
1,000-best translation list with SMT scores:
2. replacing technical terms with technical term tokens “TT_1”, “TT_2”, …
1,000-best translations with technical term tokens “TT_1”, “TT_2”, … :
3. rescoring by NMT translation model (with technical term token “TT_1”, “TT_2”, …)
… …
1000-best translations with NMT scores :
4. Reranking 1,000-best translations according to the average of SMT score and NMT score
Reranked 1,000-best translations:
(Technical term translations recovered from SMT translations)
output Chinese translation:
在/接触/插塞/ … … /有/金属膜/14/。
5. output translation with the highest score

1,000-best translation list with SMT scores:

Rank	Translation	SMT Score
1	Ch$_1$: 在/接触/插塞/… …/金属膜/14/。	-5.84
	… …	
6	Ch$_6$: 在/接触/插塞/… … /有/金属膜/14/。	-6.27
	… …	

1,000-best translations with technical term tokens “TT_1”, “TT_2”, … :

Rank	Translation	SMT Score
1	Ch$_1$:在/TT_1/… …/TT_3/14/。	-5.84
	… …	
6	Ch$_6$:在/TT_1/… … /有/TT_3/14/。	-6.27
	… …	

1000-best translations with NMT scores :

Rank	Translation	SMT score	NMT score
1	Ch$_1$:在/TT_1/… …/TT_3/14/。	-5.84	-6.49
	… …		
6	Ch$_6$:在/TT_1/… … /有/TT_3/14/。	-6.27	-5.00
	… …		

Reranked 1,000-best translations:

Rank	Translation	Final score
1	Ch$_6$: 在/接触/插塞/… … /有/金属膜/14/。	-5.63
	… …	
4	Ch$_1$:在/接触/插塞/… …/金属膜/14/。	-6.16
	… …	

Figure 4: NMT rescoring of 1,000-best SMT translations with technical term tokens “TT_i” ($i = 1, 2, \ldots$)

Out of the total 6.5M occurrences of technical term pairs, 6.2M were replaced with technical term tokens using the phrase translation table, while the remaining 300K were replaced with technical term tokens using the word alignment.[9] We limited both the Japanese vocabulary (the source language) and the Chinese vocabulary (the target language) to 40K most frequently used words.

Within the total 1,000 Japanese patent sentences in the test set, 2,244 occurrences of Japanese technical terms were identified, which correspond to 1,857 types.

5.2 Training Details

For the training of the SMT model, including the word alignment and the phrase translation table, we used Moses (Koehn et al., 2007), a toolkit for a phrase-based SMT models.

For the training of the NMT model, our training procedure and hyperparameter choices were similar to those of Sutskever et al. (2014). We used a deep LSTM neural network comprising three layers, with 512 cells in each layer, and a 512-dimensional word embedding. Similar to Sutskever et al. (2014), we reversed the words in the source sentences and ensure that all sentences in a minibatch are roughly the same length. Further training details are given below:

- All of the LSTM's parameter were initialized with a uniform distribution ranging between -0.06 and 0.06.
- We set the size of a minibatch to 128.
- We used the stochastic gradient descent, beginning at a learning rate of 0.5. We computed the perplexity of the development set using the currently produced NMT model after every 1,500 minibatches were trained and multiplied the learning rate by 0.99 when the perplexity did not decrease with respect to the last three perplexities. We trained our model for a total of 10 epoches.
- Similar to Sutskever et al. (2014), we rescaled the normalized gradient to ensure that its norm does not exceed 5.

[9] There are also Japanese technical terms (3% of all the extracted terms) for which Chinese translations can be identified using neither the SMT phrase translation table nor the SMT word alignment.

Table 1: Automatic evaluation results

System	NMT decoding and SMT technical term translation		NMT rescoring of 1,000-best SMT translations	
	BLEU	RIBES	BLEU	RIBES
Baseline SMT (Koehn et al., 2007)	52.5	88.5	-	-
Baseline NMT	53.5	90.0	55.0	89.1
NMT with technical term translation by SMT	55.3	**90.8**	**55.6**	89.3

Table 2: Human evaluation results (the score of pairwise evaluation ranges from -100 to 100 and the score of JPO adequacy evaluation ranges from 1 to 5)

System	NMT decoding and SMT technical term translation		NMT rescoring of 1,000-best SMT translations	
	pairwise evaluation	JPO adequacy evaluation	pairwise evaluation	JPO adequacy evaluation
Baseline SMT (Koehn et al., 2007)	-	3.5	-	-
Baseline NMT	5.0	3.8	28.5	4.1
NMT with technical term translation by SMT	**36.5**	**4.3**	31.0	4.1

We implement the NMT system using TensorFlow,[10] an open source library for numerical computation. The training time was around two days when using the described parameters on an 1-GPU machine.

5.3 Evaluation Results

We calculated automatic evaluation scores for the translation results using two popular metrics: BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010). As shown in Table 1, we report the evaluation scores, on the basis of the translations by Moses (Koehn et al., 2007), as the baseline SMT[11] and the scores based on translations produced by the equivalent NMT system without our proposed approach as the baseline NMT. As shown in Table 1, the two versions of the proposed NMT systems clearly improve the translation quality when compared with the baselines. When compared with the baseline SMT, the performance gain of the proposed system is approximately 3.1 BLEU points if translations are produced by the proposed NMT system of Section 4.3 or 2.3 RIBES points if translations are produced by the proposed NMT system of Section 4.2. When compared with the result of decoding with the baseline NMT, the proposed NMT system of Section 4.2 achieved performance gains of 0.8 RIBES points. When compared with the result of reranking with the baseline NMT, the proposed NMT system of Section 4.3 can still achieve performance gains of 0.6 BLEU points. Moreover, when the output translations produced by NMT decoding and SMT technical term translation described in Section 4.2 with the output translations produced by decoding with the baseline NMT, the number of unknown tokens included in output translations reduced from 191 to 92. About 90% of remaining unknown tokens correspond to numbers, English words, abbreviations, and symbols.[12]

In this study, we also conducted two types of human evaluation according to the work of Nakazawa et al. (2015): pairwise evaluation and JPO adequacy evaluation. During the procedure of pairwise eval-

[10] `https://www.tensorflow.org/`

[11] We train the SMT system on the same training set and tune it with development set.

[12] In addition to the two versions of the proposed NMT systems presented in Section 4, we evaluated a modified version of the propsed NMT system, where we introduce another type of token corresponding to unknown compound nouns and integrate this type of token with the technical term token in the procedure of training the NMT model. We achieved a slightly improved translation performance, BLEU/RIBES scores of 55.6/90.9 for the proposed NMT system of Section 4.2 and those of 55.7/89.5 for the proposed NMT system of Section 4.3.

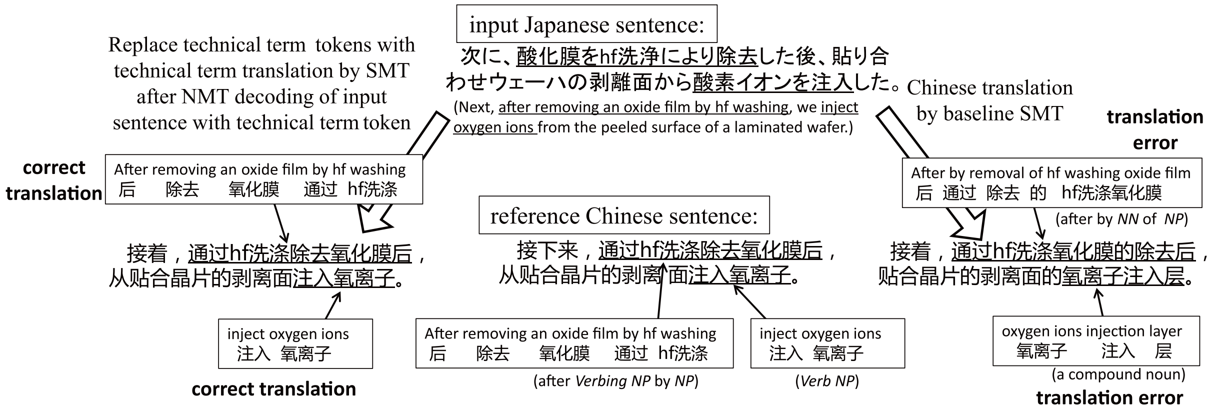

input Japanese sentence:
次に、酸化膜をhf洗浄により除去した後、貼り合わせウェーハの剥離面から酸素イオンを注入した。
(Next, after removing an oxide film by hf washing, we inject oxygen ions from the peeled surface of a laminated wafer.)
Replace technical term tokens with technical term translation by SMT after NMT decoding of input sentence with technical term token
correct translation
After removing an oxide film by hf washing
后 除去 氧化膜 通过 hf洗涤
接着，通过hf洗涤除去氧化膜后，从贴合晶片的剥离面注入氧离子。
inject oxygen ions
注入 氧离子
correct translation
reference Chinese sentence:
接下来，通过hf洗涤除去氧化膜后，从贴合晶片的剥离面注入氧离子。
After removing an oxide film by hf washing
后 除去 氧化膜 通过 hf洗涤
(after Verbing NP by NP)
inject oxygen ions
注入 氧离子
(Verb NP)
Chinese translation by baseline SMT
translation error
After by removal of hf washing oxide film
后 通过 除去 的 hf洗涤氧化膜
(after by NN of NP)
接着，通过hf洗涤氧化膜的除去后，贴合晶片的剥离面的氧离子注入层。
oxygen ions injection layer
氧离子 注入 层
(a compound noun)
translation error

Figure 5: Example of correct translations produced by the proposed NMT system with SMT technical term translation (compared with baseline SMT)

uation, we compare each of translations produced by the baseline SMT with that produced by the two versions of the proposed NMT systems, and judge which translation is better, or whether they are with comparable quality. The score of pairwise evaluation is defined by the following formula, where W is the number of better translations compared to the baseline SMT, L the number of worse translations compared to the baseline SMT, and T the number of translations having their quality comparable to those produced by the baseline SMT:

$$score = 100 \times \frac{W - L}{W + L + T}$$

The score of pairwise evaluation ranges from -100 to 100. In the JPO adequacy evaluation, Chinese translations are evaluated according to the quality evaluation criterion for translated patent documents proposed by the Japanese Patent Office (JPO).[13] The JPO adequacy criterion judges whether or not the technical factors and their relationships included in Japanese patent sentences are correctly translated into Chinese, and score Chinese translations on the basis of the percentage of correctly translated information, where the score of 5 means all of those information are translated correctly, while that of 1 means most of those information are not translated correctly. The score of the JPO adequacy evaluation is defined as the average over the whole test sentences. Unlike the study conducted Nakazawa et al. (Nakazawa et al., 2015), we randomly selected 200 sentence pairs from the test set for human evaluation, and both human evaluations were conducted using only one judgement. Table 2 shows the results of the human evaluation for the baseline SMT, the baseline NMT, and the proposed NMT system. We observed that the proposed system achieved the best performance for both pairwise evaluation and JPO adequacy evaluation when we replaced technical term tokens with SMT technical term translations after decoding the source sentence with technical term tokens.

Throughout Figure 5∼Figure 7, we show an identical source Japanese sentence and each of its translations produced by the two versions of the proposed NMT systems, compared with translations produced by the three baselines, respectively. Figure 5 shows an example of correct translation produced by the proposed system in comparison to that produced by the baseline SMT. In this example, our model correctly translates the Japanese sentence into Chinese, whereas the translation by the baseline SMT is a translation error with several erroneous syntactic structures. As shown in Figure 6, the second example highlights that the proposed NMT system of Section 4.2 can correctly translate the Japanese technical term "貼り合わせウェーハ"(laminated wafer) to the Chinese technical term "贴合晶片". The translation by the baseline NMT is a translation error because of not only the erroneously translated unknown token but also the Chinese word "贴合的", which is not appropriate as a component of a Chinese technical term. Another example is shown in Figure 7, where we compare the translation of a reranking SMT 1,000-best

[13] `https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/tokkyohonyaku_hyouka/01.pdf` (in Japanese)

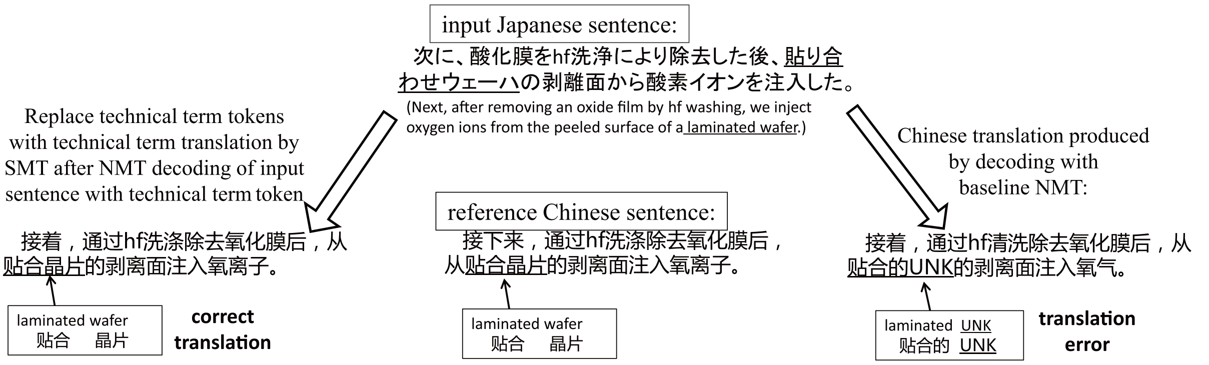

input Japanese sentence:
次に、酸化膜をhf洗浄により除去した後、貼り合わせウェーハの剥離面から酸素イオンを注入した。
(Next, after removing an oxide film by hf washing, we inject oxygen ions from the peeled surface of a laminated wafer.)
Replace technical term tokens with technical term translation by SMT after NMT decoding of input sentence with technical term token
接着，通过hf洗涤除去氧化膜后，从贴合晶片的剥离面注入氧离子。
laminated wafer
贴合 晶片
correct translation
reference Chinese sentence:
接下来，通过hf洗涤除去氧化膜后，从贴合晶片的剥离面注入氧离子。
laminated wafer
贴合 晶片
Chinese translation produced by decoding with baseline NMT:
接着，通过hf清洗除去氧化膜后，从贴合的UNK的剥离面注入氧气。
laminated UNK
贴合的 UNK
translation error

Figure 6: Example of correct translations produced by the proposed NMT system with SMT technical term translation (compared to decoding with the baseline NMT)

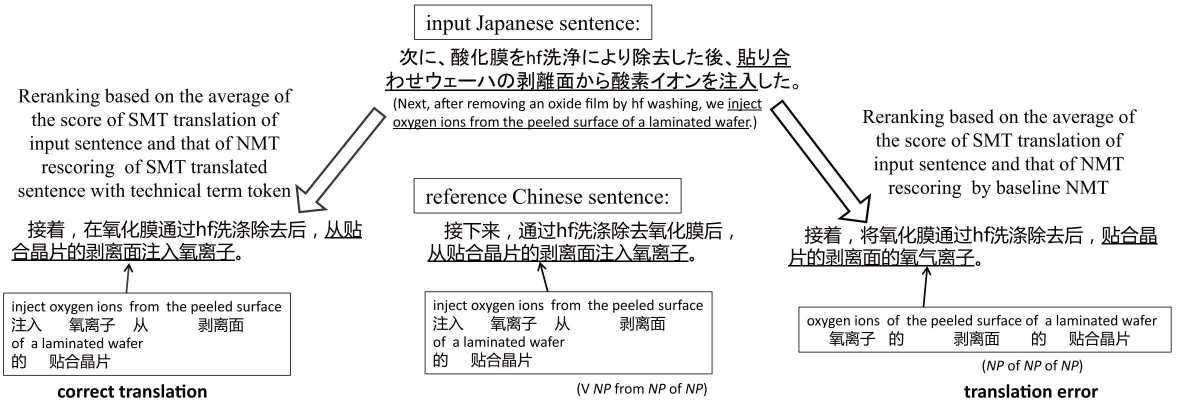

input Japanese sentence:
次に、酸化膜をhf洗浄により除去した後、貼り合わせウェーハの剥離面から酸素イオンを注入した。
(Next, after removing an oxide film by hf washing, we inject oxygen ions from the peeled surface of a laminated wafer.)
Reranking based on the average of the score of SMT translation of input sentence and that of NMT rescoring of SMT translated sentence with technical term token
接着，在氧化膜通过hf洗涤除去后，从贴合晶片的剥离面注入氧离子。
inject oxygen ions from the peeled surface
注入 氧离子 从 剥离面
of a laminated wafer
的 贴合晶片
correct translation
reference Chinese sentence:
接下来，通过hf洗涤除去氧化膜后，从贴合晶片的剥离面注入氧离子。
inject oxygen ions from the peeled surface
注入 氧离子 从 剥离面
of a laminated wafer
的 贴合晶片
(V NP from NP of NP)
Reranking based on the average of the score of SMT translation of input sentence and that of NMT rescoring by baseline NMT
接着，将氧化膜通过hf洗涤除去后，贴合晶片的剥离面的氧气离子。
oxygen ions of the peeled surface of a laminated wafer
氧离子 的 剥离面 的 贴合晶片
(NP of NP of NP)
translation error

Figure 7: Example of correct translations produced by reranking the 1,000-best SMT translations with the proposed NMT system (compared to reranking with the baseline NMT)

translation produced by the proposed NMT system with that produced by reranking with the baseline NMT. It is interesting to observe that compared with the baseline NMT, we obtain a better translation when we rerank the 1,000-best SMT translations using the proposed NMT system, in which technical term tokens represent technical terms. It is mainly because the correct Chinese translation "晶片"(wafter) of Japanese word "ウェーハ" is out of the 40K NMT vocabulary (Chinese), causing reranking with the baseline NMT to produce the translation with an erroneous construction of "noun phrase of noun phrase of noun phrase". As shown in Figure 7, the proposed NMT system of Section 4.3 produced the translation with a correct construction, mainly because Chinese word "晶片"(wafter) is a part of Chinese technical term "贴合晶片"(laminated wafter) and is replaced with a technical term token and then rescored by the NMT model (with technical term tokens "TT_1", "TT_2", ...).

6 Conclusion

In this paper, we proposed an NMT method capable of translating patent sentences with a large vocabulary of technical terms. We trained an NMT system on a bilingual corpus, wherein technical terms are replaced with technical term tokens; this allows it to translate most of the source sentences except the technical terms. Similar to Sutskever et al. (2014), we used it as a decoder to translate the source sentences with technical term tokens and replace the tokens with technical terms translated using SMT. We also used it to rerank the 1,000-best SMT translations on the basis of the average of the SMT score and that of NMT rescoring of translated sentences with technical term tokens. For the translation of Japanese patent sentences, we observed that our proposed NMT system performs better than the phrase-based SMT system as well as the equivalent NMT system without our proposed approach.

One of our important future works is to evaluate our proposed method in the NMT system proposed by Bahdanau et al. (2015), which introduced a bidirectional recurrent neural network as encoder and is the state-of-the-art of pure NMT system recently. However, the NMT system proposed by Bahdanau et al. (2015) also has a limitation in addressing out-of-vocabulary words. Our proposed NMT system is expected to improve the translation performance of patent sentences by applying approach of Bahdanau et al. (2015). Another important future work is to quantitatively compare our study with the work of Luong et al. (2015b). In the work of Luong et al. (2015b), they replace low frequency single words and translate them in a post-processing Step using a dictionary, while we propose to replace the whole technical terms and post-translate them with phrase translation table of SMT system. Therefore, our proposed NMT system is expected to be appropriate to translate patent documents which contain many technical terms comprised of multiple words and should be translated together. We will also evaluate the present study by reranking the n-best translations produced by the proposed NMT system on the basis of their SMT rescoring. Next, we will rerank translations from both the n-best SMT translations and n-best NMT translations. As shown in Section 5.3, the decoding approach of our proposed NMT system achieved the best RIBES performance and human evaluation scores in our experiments, whereas the reranking approach achieved the best performance with respect to BLEU. A translation with the highest average SMT and NMT scores of the n-best translations produced by NMT and SMT, respectively, is expected to be an effective translation.

References

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. 3rd ICLR*.

K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proc. EMNLP*.

M. R. Costa-jussà and J. A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proc. 54th ACL*, pages 357–361.

L. Dong, Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. 2015. Collecting bilingual technical terms from Japanese-Chinese patent families by SVM. In *Proc. PACLING*, pages 71–79.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proc. EMNLP*, pages 944–952.

S. Jean, K. Cho, Y. Bengio, and R. Memisevic. 2014. On using very large target vocabulary for neural machine translation. In *Proc. 28th NIPS*, pages 1–10.

N. Kalchbrenner and P. Blunsom. 2013. Recurrent continous translation models. In *Proc. EMNLP*, pages 1700–1709.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pages 177–180.

X. Li, J. Zhang, and C. Zong. 2016. Towards zero unknown word in neural machine translation. In *Proc. 25th IJCAI*, pages 2852–2858.

M. Luong and C. D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proc. 54th ACL*, pages 1054–1063.

M. Luong, H. Pham, and C. D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*.

M. Luong, I. Sutskever, O. Vinyals, Q. V. Le, and W. Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *Proc. 53rd ACL*, pages 11–19.

T. Nakazawa, H. Mino, I. Goto, G. Neubig, S. Kurohashi, and E. Sumita. 2015. Overview of the 2nd workshop on asian translation. In *Proc. 2nd WAT*, pages 1–28.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pages 311–318.

R. Sennrich, B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. 54th ACL*, pages 1715–1725.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural machine translation. In *Proc. 28th NIPS*.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pages 168–171.

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