Patent NMT integrated with Large Vocabulary Phrase Translation by SMT at WAT 2017

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

Abstract

Neural machine translation (NMT) cannot handle a larger vocabulary because the 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. Long et al. (2017) proposed to select phrases that contain out-of-vocabulary words using the statistical approach of branching entropy. The selected phrases are then replaced with tokens during training and post-translated by the phrase translation table of SMT. In this paper, we apply the method proposed by Long et al. (2017) to the WAT 2017 Japanese-Chinese and Japanese-English patent datasets. Evaluation on Japanese-to-Chinese, Chinese-to-Japanese, Japanese-to-English and English-to-Japanese patent sentence translation proved the effectiveness of phrases selected with branching entropy, where the NMT model of Long et al. (2017) achieves a substantial improvement over a baseline NMT model without the technique proposed by Long et al. (2017).

1 Introduction

Neural machine translation (NMT), a new approach to solving machine translation, has achieved promising results (Bahdanau et al., 2015; Cho et al., 2014; Jean et al., 2014; Kalchbrenner and Blunsom, 2013; Luong et al., 2015a,b; Sutskever et al., 2014). 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 the 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 "$\langle unk \rangle$" 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 et al. (2014) provided an efficient approximation to the softmax function to accommodate a very large vocabulary in an NMT system. Luong et al. (2015b) proposed annotating the occurrences of the out-of-vocabulary token in the target sentence 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 replacing 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 out-of-vocabulary words as sequences of subword units. Luong and Manning (2016) provided a character-level and word-level hybrid NMT model to achieve an open vocabulary, and Costa-jussà and Fonollosa (2016) proposed an NMT system that uses character-based embeddings.

Proceedings of the 4th Workshop on Asian Translation, pages 110–118,
Taipei, Taiwan, November 27, 2017. ©2017 AFNLP

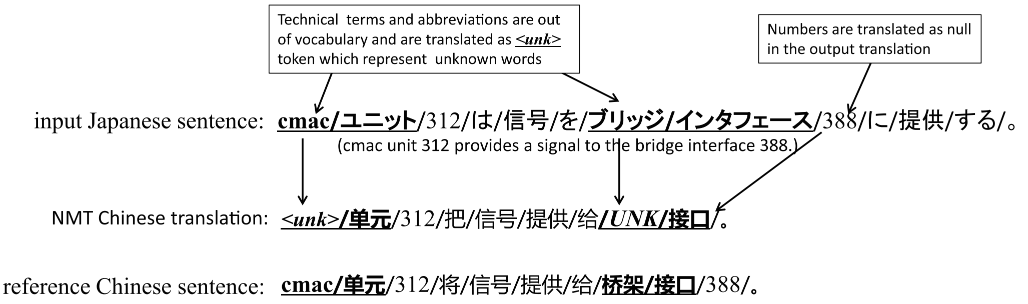

Technical terms and abbreviations are out of vocabulary and are translated as <unk> token which represent unknown words
Numbers are translated as null in the output translation
input Japanese sentence: cmac/ユニット/312/は/信号/を/ブリッジ/インタフェース/388/に/提供/する/。
(cmac unit 312 provides a signal to the bridge interface 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

However, these previous approaches have limitations when translating patent sentences. This is because their methods only focus on addressing the problem of out-of-vocabulary 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 Figure 1, where the 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".

To address this problem, Long et al. (2016) proposed extracting compound nouns as technical terms and replacing them with tokens. Long et al. (2017) proposed to select phrase pairs using the statistical approach of branching entropy; this allows the proposed technique to be applied to the translation task on any language pair without needing specific language knowledge to formulate the rules for technical term identification. In this paper, we apply the method proposed by Long et al. (2017) to the WAT 2017 Japanese-Chinese and Japanese-English patent datasets. On the WAT 2017 Japanese-Chinese JPO patent dataset, the NMT model of Long et al. (2017) achieves an improvement of 1.4 BLEU points over a baseline NMT model when translating Japanese sentences into Chinese, and an improvement of 0.8 BLEU points when translating Chinese sentences into Japanese. On the WAT 2017 Japanese-English JPO patent dataset, the NMT model of Long et al. (2017) achieves an improvement of 0.8 BLEU points over a baseline NMT model when translating Japanese sentences into English, and an improvement of 0.7 BLEU points when translating English sentences into Japanese. Moreover, the number of translation error of under-translations[1] by PosUnk model proposed by Luong et al. (2015b) reduces to around 30% by the NMT model of Long et al. (2017).

2 Neural Machine Translation

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

$$p(y_z \mid y_{<z}, \boldsymbol{x})$$

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

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

In this paper, we use an NMT model similar to that used by Bahdanau et al. (2015), which consists of an encoder of a bidirectional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and another LSTM as decoder. In the model of Bahdanau et al. (2015), the encoder consists of forward and backward LSTMs. The forward LSTM reads the source sentence as it is ordered (from x_1 to x_N) and calculates a sequence of forward hidden states, while the backward LSTM reads the source sentence in the reverse order

[1] It is known that NMT models tend to have the problem of the under-translation. Tu el al. (2016) proposed coverage-based NMT which considers the problem of the under-translation.

(from x_N to x_1) , resulting in a sequence of backward hidden states. The decoder then predicts target words using not only a recurrent hidden state and the previously predicted word but also a context vector as followings:

$$p(y_z \mid y_{<z}, \boldsymbol{x}) = g(y_{z-1}, s_{z-1}, c_z)$$

where s_{z-1} is an LSTM hidden state of decoder, and c_z is a context vector computed from both of the forward hidden states and backward hidden states, for $1 \leq z \leq M$.

3 Phrase Pair Selection using Branching Entropy

Branching entropy has been applied to the procedure of text segmentation (e.g., (Jin and Tanaka-Ishii, 2006)) and key phrases extraction (e.g., (Chen et al., 2010)). In this work, we use the left/right branching entropy to detect the boundaries of phrases, and thus select phrase pairs automatically.

3.1 Branching Entropy

The left branching entropy and right branching entropy of a phrase $\boldsymbol{w}$ are respectively defined as

$$H_l(\boldsymbol{w}) = -\sum_{v \in V_l^{\boldsymbol{w}}} p_l(v) \log_2 p_l(v)$$

$$H_r(\boldsymbol{w}) = -\sum_{v \in V_r^{\boldsymbol{w}}} p_r(v) \log_2 p_r(v)$$

where $\boldsymbol{w}$ is the phrase of interest (e.g., "ブリッジ/インターフェース" in the Japanese sentence shown in Figure 1, which means "bridge interface"), $V_l^{\boldsymbol{w}}$ is a set of words that are adjacent to the left of $\boldsymbol{w}$ (e.g., "を" in Figure 1, which is a Japanese particle) and $V_r^{\boldsymbol{w}}$ is a set of words that are adjacent to the right of $\boldsymbol{w}$ (e.g., "388" in Figure 1). The probabilities $p_l(v)$ and $p_r(v)$ are respectively computed as

$$p_l(v) = \frac{f_{v,\boldsymbol{w}}}{f_{\boldsymbol{w}}} \qquad p_r(v) = \frac{f_{\boldsymbol{w},v}}{f_{\boldsymbol{w}}} \qquad (1)$$

where $f_{\boldsymbol{w}}$ is the frequency count of phrase $\boldsymbol{w}$, and $f_{v,\boldsymbol{w}}$ and $f_{\boldsymbol{w},v}$ are the frequency counts of sequence "v,$\boldsymbol{w}$" and sequence "$\boldsymbol{w}$,v" respectively. According to the definition of branching entropy, when a phrase $\boldsymbol{w}$ is a technical term that is always used as a compound word, both its left branching entropy $H_l(\boldsymbol{w})$ and right branching entropy $H_r(\boldsymbol{w})$ have high values because many different words, such as particles and numbers, can be adjacent to the phrase. However, the left/right branching entropy of substrings of $\boldsymbol{w}$ have low values because words contained in $\boldsymbol{w}$ are always adjacent to each other.

3.2 Selecting Phrase Pairs

Given a parallel sentence pair $\langle S_s, S_t \rangle$, all n-grams phrases of source sentence S_s and target sentence S_t are extracted and aligned using phrase translation table and word alignment of SMT according to the approaches described in Long et al. (2016). Next, phrase translation pair $\langle t_s, t_t \rangle$ obtained from $\langle S_s, S_t \rangle$ that satisfies all the following conditions is selected as a phrase pair and is extracted:

(1) Either t_s or t_t contains at least one out-of-vocabulary word.

(2) Neither t_s nor t_t contains predetermined stop words.

(3) Entropies $H_l(t_s)$, $H_l(t_t)$, $H_r(t_s)$ and $H_r(t_t)$ are larger than a lower bound, while the left/right branching entropy of the substrings of t_s and t_t are lower than or equal to the lower bound.

Here, the maximum length of a phrase as well as the lower bound of the branching entropy are tuned with the validation set.[2] All the selected source-target phrase pairs are then used in the next section as phrase pairs.

4 NMT with a Large Phrase Vocabulary

In this work, the NMT model is trained on a bilingual corpus in which phrase pairs are replaced with tokens. The NMT system is then used as a decoder to translate the source sentences and replace the tokens with phrases translated using SMT.

[2] Throughout the evaluations on patent translation of both language pairs of Japanese-Chinese and Japanese-English, the maximum length of the extracted phrases is tuned as 7. The lower bounds of the branching entropy are tuned as 5 for patent translation of the language pair of Japanese-Chinese, and 8 for patent translation of the language pair of Japanese-English. We also tune the number of stop words using the validation set, and use the 200 most-frequent Japanese morphemes and Chinese words as stop words for the language pair of Japanese-Chinese, use the 100 most-frequent Japanese morphemes and English words as stop words for the language pair of Japanese-English.

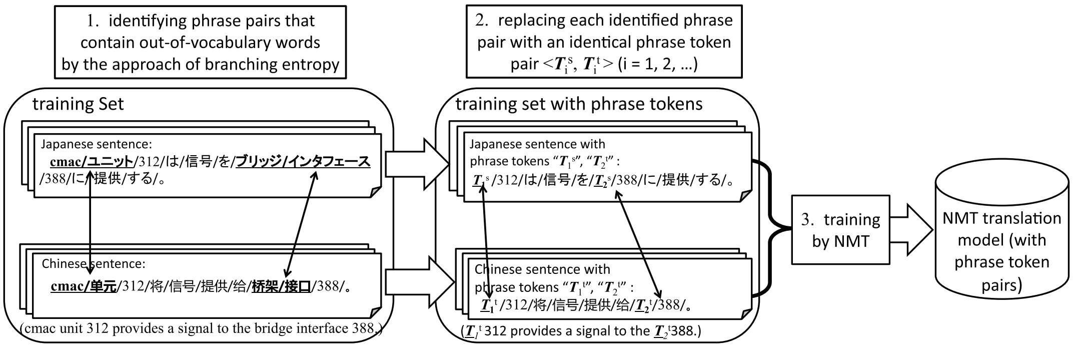

1. identifying phrase pairs that contain out-of-vocabulary words by the approach of branching entropy
2. replacing each identified phrase pair with an identical phrase token pair <T_i^s, T_i^t> (i = 1, 2, …)
training Set
Japanese sentence:
cmac/ユニット/312/は/信号/を/ブリッジ/インタフェース/388/に/提供/する/。
Chinese sentence:
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.)
training set with phrase tokens
Japanese sentence with phrase tokens "T_1^s", "T_2^s" :
T_1^s/312/は/信号/を/T_2^s/388/に/提供/する/。
Chinese sentence with phrase tokens "T_1^t", "T_2^t" :
T_1^t/312/将/信号/提供/给/T_2^t/388/。
(T_1^t 312 provides a signal to the T_2^t388.)
3. training by NMT
NMT translation model (with phrase token pairs)

Figure 2: NMT training after replacing phrase pairs with token pairs $\langle T_i^s, T_i^t \rangle$ $(i = 1, 2, \ldots)$

4.1 NMT Training after Replacing Phrase Pairs with Tokens

Figure 2 illustrates the procedure for training the model with parallel patent sentence pairs in which phrase pairs are replaced with phrase token pairs $\langle T_1^s, T_1^t \rangle$, $\langle T_2^s, T_2^t \rangle$, and so on.

In the step 1 of Figure 2, source-target phrase pairs that contain at least one out-of-vocabulary word are selected from the training set using the branching entropy approach described in Section 3.2. As shown in the step 2 of Figure 2, in each of the parallel patent sentence pairs, occurrences of phrase pairs $\langle t_1^s, t_1^t \rangle$, $\langle t_2^s, t_2^t \rangle$, …, $\langle t_k^s, t_k^t \rangle$ are then replaced with token pairs $\langle T_1^s, T_1^t \rangle$, $\langle T_2^s, T_2^t \rangle$, …, $\langle T_k^s, T_k^t \rangle$. Phrase pairs $\langle t_1^s, t_1^t \rangle$, $\langle t_2^s, t_2^t \rangle$, …, $\langle t_k^s, t_k^t \rangle$ are numbered in the order of occurrence of the source phrases t_1^s $(i = 1, 2, \ldots, k)$ in each source sentence S_s. Here note that in all the parallel sentence pairs $\langle S_s, S_t \rangle$, the tokens pairs $\langle T_1^s, T_1^t \rangle$, $\langle T_2^s, T_2^t \rangle$, … that are identical throughout all the parallel sentence pairs are used in this procedure. Therefore, for example, in all the source patent sentences S_s, the phrase t_1^s which appears earlier than other phrases in S_s is replaced with T_1^s. We then train the NMT model on a bilingual corpus, in which the phrase pairs are replaced by token pairs $\langle T_i^s, T_i^t \rangle$ $(i = 1, 2, \ldots)$, and obtain an NMT model in which the phrases are represented as tokens.

Table 1: Statistics of datasets

	training set	validation set	test set
ja ↔ ch	998,054	2,000	2,000
ja ↔ en	999,636	2,000	2,000

4.2 NMT Decoding and SMT Phrase Translation

Figure 3 illustrates the procedure for producing target translations by decoding the input source sentence using the NMT model of Long et al. (2017).

In the step 1 of Figure 3, when given an input source sentence, we first generate its translation by decoding of SMT translation model. Next, as shown in the step 2 of Figure 3, we automatically extract the phrase pairs by branching entropy according to the procedure of Section 3.2, where the input sentence and its SMT translation are considered as a pair of parallel sentence. Phrase pairs that contains at least one out-of-vocabulary word are extracted and are replaced with phrase token pairs $\langle T_i^s, T_i^t \rangle$ $(i = 1, 2, \ldots)$. Consequently, we have an input sentence in which the tokens "T_i^s" $(i = 1, 2, \ldots)$ represent the positions of the phrases and a list of SMT phrase translations of extracted Japanese phrases. Next, as shown in the step 3 of Figure 3, the source Japanese sentence with tokens is translated using the NMT model trained according to the procedure described in Section 4.1. Finally, in the step 4, we replace the tokens "T_i^t" $(i = 1, 2, \ldots)$ of the target sentence translation with the phrase translations of the SMT.

5 Evaluation

5.1 DataSets

We evaluated the effectiveness of the NMT model of Long et al. (2017) on the WAT 2017 Japanese-Chinese and Japanese-English JPO dataset.[3] Out of the training set of the WAT 2017 Japanese-Chinese JPO dataset, we used 998,954 patent sentence pairs, whose Japanese sentences contain

[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html

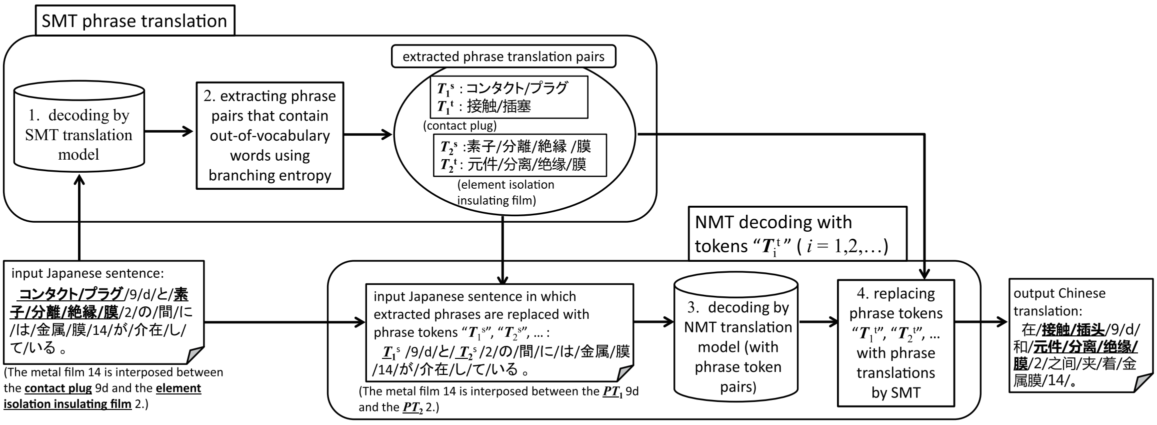

SMT phrase translation
1. decoding by SMT translation model
2. extracting phrase pairs that contain out-of-vocabulary words using branching entropy
extracted phrase translation pairs
T_1^s : コンタクト/プラグ
T_1^t : 接触/插塞
(contact plug)
T_2^s :素子/分離/絶縁 /膜
T_2^t : 元件/分离/绝缘/膜
(element isolation insulating film)
NMT decoding with tokens "T_i^t" (i = 1,2,…)
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.)
input Japanese sentence in which extracted phrases are replaced with phrase tokens "T_1^s", "T_2^s", … :
T_1^s /9/d/と/ T_2^s /2/の/間/に/は/金属/膜/14/が/介在/し/て/いる 。
(The metal film 14 is interposed between the PT_1 9d and the PT_2 2.)
3. decoding by NMT translation model (with phrase token pairs)
4. replacing phrase tokens "T_1^t", "T_2^t", … with phrase translations 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 tokens "T_i^s" ($i = 1, 2, \ldots$) and the SMT phrase translation

fewer than 100 morphemes, Chinese sentences contain fewer than 100 words. Out of the training set of the WAT 2017 Japanese-English JPO dataset, we used 999,636 sentence pairs whose Japanese sentences contain fewer than 100 morphemes and English sentences contain fewer than 100 words. In both cases, we used all of the sentence pairs contained in the development sets of the WAT 2017 JPO datasets as development sets, and we used all of the sentence pairs contained in the test sets of the WAT 2017 JPO datasets as test sets. Table 1 show the statistics of the dataset.

According to the procedure of Section 3.2, from the Japanese-Chinese sentence pairs of the training set, we collected 102,630 occurrences of Japanese-Chinese phrase pairs, which are 69,387 types of phrase pairs with 52,786 unique types of Japanese phrases and 67,456 unique types of Chinese phrases. Within the total 2,000 Japanese patent sentences in the Japanese-Chinese test set, 266 occurrences of Japanese phrases were extracted, which correspond to 247 types. With the total 2,000 Chinese patent sentences in the Japanese-Chinese test set, 417 occurrences of Chinese phrases were extracted, which correspond to 382 types.

From the Japanese-English sentence pairs of the training set, we collected 38,457 occurrences of Japanese-English phrase pairs, which are 35,544 types of phrase pairs with unique 34,569 types of Japanese phrases and 35,087 unique types of English phrases. Within the total 2,000 Japanese patent sentences in the Japanese-English test set, 249 occurrences of Japanese phrases were extracted, which correspond to 221 types. With the total 2,000 English patent sentences in the Japanese-English test set, 246 occurrences of English phrases were extracted, which correspond to 230 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 phrase-based SMT models. We trained the SMT model on the training set and tuned it with the validation set.

For the training of the NMT model, our training procedure and hyperparameter choices were similar to those of Bahdanau et al. (2015). The encoder consists of forward and backward deep LSTM neural networks each consisting of three layers, with 512 cells in each layer. The decoder is a three-layer deep LSTM with 512 cells in each layer. Both the source vocabulary and the target vocabulary are limited to the 40K most-frequently used morphemes / words in the training set. The size of the word embedding was set to 512. We ensured that all sentences in a minibatch were roughly the same length. Further training details are given below: (1) We set the size of a minibatch to 128. (2) All of the LSTM's parameter were initialized with a uniform distribution ranging between -0.06 and 0.06. (3) We used the stochastic gradient descent, beginning at a fixed learning rate of 1. We trained our model for a total of 10 epochs, and we began to halve the learning rate every epoch after the first seven epochs. (4) Similar to Sutskever et al.(2014), we rescaled the normalized gradient to ensure that its norm does not exceed 5. We trained the NMT model on the training set. The training time was around two days when using the described parameters on a 1-GPU

Table 2: Automatic evaluation results (BLEU)

System	ja → ch	ch → ja	ja → en	en → ja
Baseline SMT (Koehn et al., 2007)	30.0	36.2	28.0	29.4
Baseline NMT	34.2	40.8	43.1	41.8
NMT with PosUnk model (Luong et al., 2015b)	34.5	41.0	43.5	42.0
NMT with phrase translation by SMT (Long et al., 2017)	**35.6**	**41.6**	**43.9**	**42.5**

Table 3: Human evaluation results of pairwise evaluation

System	ja → ch	ch → ja	ja → en	en → ja
NMT with PosUnk model (Luong et al., 2015b)	13	12.5	9.5	14.5
NMT with phrase translation by SMT (Long et al., 2017)	**23.5**	**22.5**	**15.5**	**19**

machine.

We compute the branching entropy using the frequency statistics from the training set.

5.3 Evaluation Results

In this work, we calculated automatic evaluation scores for the translation results using a popular metrics called BLEU (Papineni et al., 2002). As shown in Table 2, we report the evaluation scores, using the translations by Moses (Koehn et al., 2007) as the baseline SMT and the scores using the translations produced by the baseline NMT system without the approach proposed by Long et al. (2017) as the baseline NMT. As shown in Table 2, the BLEU score obtained by the NMT model of Long et al. (2017) is clearly higher than those of the baselines. Here, as described in Section 3, the lower bounds of branching entropy for phrase pair selection are tuned as 5 throughout the evaluation of language pair of Japanese-Chinese, and tuned as 8 throughout the evaluation of language pair of Japanese-English, respectively. On the WAT 2017 Japanese-Chinese JPO patent dataset, when compared with the baseline SMT, the performance gains of the NMT model of Long et al. (2017) are approximately 5.6 BLEU points when translating Japanese into Chinese and 5.4 BLEU when translating Chinese into Japanese. On the WAT 2017 Japanese-English JPO patent dataset, when compared with the baseline SMT, the performance gains of the NMT model of Long et al. (2017) are approximately 15.9 BLEU points when translating Japanese into English and 13.1 BLEU when translating English into Japanese. When compared with the result of the baseline NMT, the NMT model of Long et al. (2017) achieved performance gains of 1.4 BLEU points on the task of translating Japanese into Chinese and 0.8 BLEU points on the task of translating Chinese into Japanese. When compared with the result of the baseline NMT, the NMT model of Long et al. (2017) achieved performance gains of 0.8 BLEU points on the task of translating Japanese into English and 1.4 BLEU points on the task of translating English into Japanese.

Furthermore, we quantitatively compared our study with the work of Luong et al. (2015b). Table 2 compares the NMT model with the PosUnk model, which is the best model proposed by Luong et al. (2015b) The NMT model of Long et al. (2017) achieves performance gains of 0.9 BLEU points when translating Japanese into Chinese, and performance gains of 0.6 BLEU points when translating Chinese into Japanese. The NMT model of Long et al. (2017) achieves performance gains of 0.4 BLEU points when translating Japanese into English, and performance gains of 0.5 BLEU points when translating English into Japanese

In this study, we also conducted two types of human evaluations according to the work of Nakazawa et al. (2015): pairwise evaluation and JPO adequacy evaluation. In the pairwise evaluation, we compared each translation produced by the baseline NMT with that produced by the NMT model of Long et al. (2017) as well as the NMT model with PosUnk model, and judged which translation is better or whether they have

Table 4: Human evaluation results of JPO adequacy evaluation

System	ja → ch	ch → ja	ja → en	en → ja
Baseline SMT (Koehn et al., 2007)	3.1	3.2	2.9	3.0
Baseline NMT	3.6	3.6	3.7	3.7
NMT with PosUnk model (Luong et al., 2015b)	3.8	3.9	3.9	3.9
NMT with phrase translation by SMT (Long et al., 2017)	**4.1**	**4.1**	**4.2**	**4.1**

Table 5: Evaluation results from WAT 2017

Evaluation	System	ja → ch	ch → ja	ja → en	en → ja
Automatic evaluation (BLEU)	Baseline (PBSMT)	32.1	38.5	30.8	34.3
	NMT with phrase translation by SMT (Long et al., 2017)	**33.2**	**40.5**	**37.3**	**41.1**
Pairwise evaluation	NMT with phrase translation by SMT (Long et al., 2017)	21.8	40.1	51.5	49.5
JPO adequacy evaluation	NMT with phrase translation by SMT (Long et al., 2017)	4.1	3.9	4.2	4.3

comparable quality. In contrast to the study conducted by 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 3 and Table 4 show the results of the human evaluation for the baseline SMT, baseline NMT, NMT model with PosUnk model, and the NMT model of Long et al. (2017). We observe that the NMT model of Long et al. (2017) achieves the best performance for both the pairwise and JPO adequacy evaluations when we replace the tokens with SMT phrase translations after decoding the source sentence with the tokens.

Moreover, Table 5 shows the results of automatic evaluation, pairwise evaluation and JPO adequacy evaluation from the WAT 2017 (Nakazawa et al., 2017).[4] We observe that the NMT model of Long et al. (2017) achieves a substantial improvement over the WAT 2017 baseline.

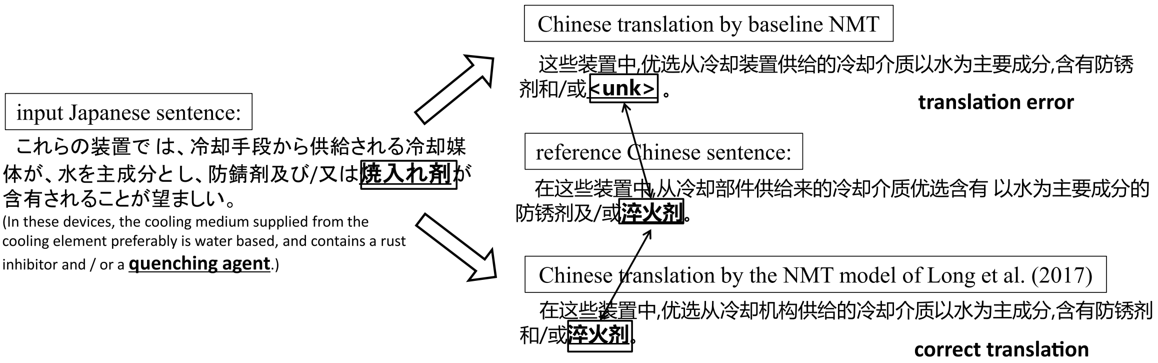
For the test sets, we also counted the numbers of the untranslated words of input sentences. As shown in Table 6, the number of untranslated words by the baseline NMT reduced to around 65% by the NMT model of Long et al. (2017). This is mainly because part of untranslated source words are out-of-vocabulary, and thus are untranslated by the baseline NMT. The NMT model of Long et al. (2017) extracts those out-of-vocabulary words as a part of phrases and replaces those phrases with tokens before the decoding of NMT. Those phrases are then translated by SMT and inserted in the output translation, which ensures that those out-of-vocabulary words are translated.

Figure 4 compares an example of correct translation produced by the NMT model of Long et al. (2017) with one produced by the baseline NMT. In this example, the translation is a translation error because the Japanese word "焼入れ (quenching)" is an out-of-vocabulary word and is erroneously translated into the "$\langle unk \rangle$" token. The NMT model of Long et al. (2017) correctly translated the Japanese sentence into Chinese, where the out-of-vocabulary word "焼入れ" is correctly selected by the approach of branching entropy as a part of the Japanese phrase "焼入れ剤 (quenching agent)". The selected Japanese phrase is then translated by the phrase translation table of SMT. Figure 5 shows another example of correct translation produced by the NMT model of Long et al. (2017) with one produced by the baseline NMT. As shown in Figure 5, the translation produced by baseline NMT is a translation error because the out-of-vocabulary English words "eukaryotic" and "promoters" are untranslated words and their translations are not contained in the output translation of the baseline NMT. The NMT model of

[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/

Table 6: Numbers of untranslated morphemes / words of input sentences

System	ja → ch	ch → ja	ja → en	en → ja
NMT with PosUnk model (Luong et al., 2015b)	1,112	846	1,031	794
NMT with phrase translation by SMT (Long et al., 2017)	736	581	655	571


input Japanese sentence:
Chinese translation by baseline NMT
translation error
reference Chinese sentence:
Chinese translation by the NMT model of Long et al. (2017)
correct translation

これらの装置では、冷却手段から供給される冷却媒体が、水を主成分とし、防錆剤及び/又は焼入れ剤が含有されることが望ましい。

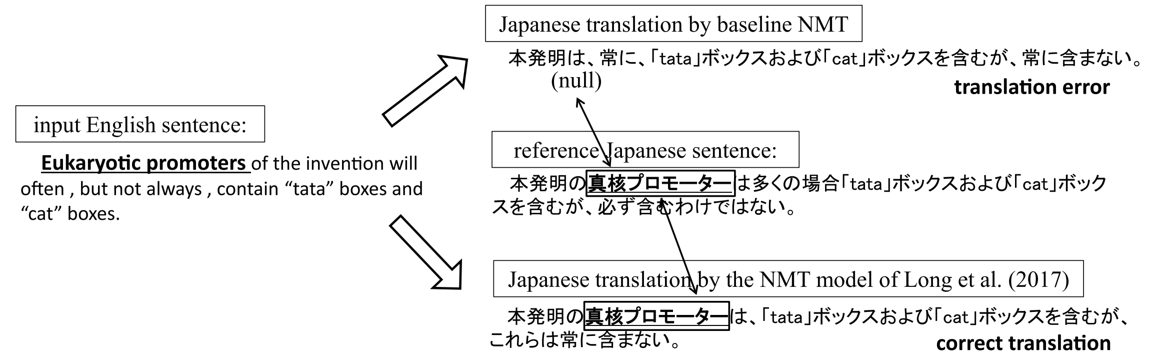
(In these devices, the cooling medium supplied from the cooling element preferably is water based, and contains a rust inhibitor and / or a **quenching agent**.)

这些装置中,优选从冷却装置供给的冷却介质以水为主要成分,含有防锈剂和/或<unk>。

在这些装置中,从冷却部件供给来的冷却介质优选含有 以水为主要成分的防锈剂及/或淬火剂。

在这些装置中,优选从冷却机构供给的冷却介质以水为主成分,含有防锈剂和/或淬火剂。

Figure 4: An example of correct translations produced by the NMT model of Long et al. (2017) when addressing the problem of out-of-vocabulary words (Japanese-to-Chinese)


input English sentence:
Japanese translation by baseline NMT
(null)
translation error
reference Japanese sentence:
Japanese translation by the NMT model of Long et al. (2017)
correct translation

Eukaryotic promoters of the invention will often , but not always , contain “tata” boxes and “cat” boxes.

本発明は、常に、「tata」ボックスおよび「cat」ボックスを含むが、常に含まない。

本発明の真核プロモーターは多くの場合「tata」ボックスおよび「cat」ボックスを含むが、必ず含むわけではない。

本発明の真核プロモーターは、「tata」ボックスおよび「cat」ボックスを含むが、これらは常に含まない。

Figure 5: An example of correct translations produced by the NMT model of Long et al. (2017) when addressing the problem of under-translation (English-to-Japanese)

Long et al. (2017) correctly translated those English words into Japanese because those English words “eukaryotic” and “promoters” are selected as an English phrase “Eukaryotic promoters” with branching entropy and then are translated by SMT.

6 Conclusion

Long et al. (2017) proposed selecting phrases that contain out-of-vocabulary words using the branching entropy. These selected phrases are then replaced with tokens and post-translated using an SMT phrase translation. In this paper, we apply the method proposed by Long et al. (2017) to the WAT 2017 Japanese-Chinese and Japanese-English patent datasets. We observed that the NMT model of Long et al. (2017) performed much better than the baseline NMT system in all of the language pairs: Japanese-to-Chinese/Chinese-to-Japanese and Japanese-to-English/English-to-Japanese. One of our important future tasks is to compare the translation performance of the NMT model of Long et al. (2017) with that based on subword units (e.g. (Sennrich et al., 2016)). Another future work is to integrate the reranking framework for minimizing untranslated content (Goto and Tanaka, 2017) into the NMT model of Long et al. (2017), which is expected to further reduce the number of untranslated words. This future work is roughly based on the observation reported in Kimura et al. (2017), where the NMT model of Long et al. (2017) is not only effective in reducing the untranslated content without any specific framework of minimizing the untranslated content, but also successfully reduced the estimated volumes of the untranslated content, which was proposed by Goto and Tanaka (2017).

References

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

Y. Chen, Y. Huang, S. Kong, and L. Lee. 2010. Automatic key term extraction from spoken course lectures using branching entropy and prosodic/semantic features. In *Proc. 2010 IEEE SLT Workshop*, pages 265–270.

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*, pages 1724–1734.

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

I. Goto and H. Tanaka. 2017. Detecting untranslated content for neural machine translation. In *Proc. 1st NMT*, pages 47–55.

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

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.

Z. Jin and K. Tanaka-Ishii. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. In *Proc. COLING/ACL 2006*, pages 428–435.

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

R. Kimura, Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. 2017. Effect on reducing untranslated content by neural machine translation with a large vocabulary of technical terms. In *Proc. 7th PSLT*, pages 9–20.

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.

Z. Long, R. Kimura, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. 2017. Neural machine translation model with a large vocabulary selected by branching entropy. In *Proc. MT Summit XVI*, pages 227–240.

Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. 2016. Translation of patent sentences with a large vocabulary of technical terms using neural machine translation. In *Proc. 3rd WAT*, pages 47–57.

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*, pages 1412–1421.

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, S. Higashiyama, C. Ding, H. Mino, I. Goto, G. Neubig, H. Kazawa, Y. Oda, J. Harashima, and S. Kurohashi. 2017. Overview of the 4th workshop on Asian translation. In *Proc. 4th WAT*.

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. 27th NIPS*, pages 3104–3112.

Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li. 2016. Modeling coverage for neural machine translation. In *Proc. ACL 2016*, pages 76–85.