Effect on Reducing Untranslated Content by Neural Machine Translation with a Large Vocabulary of Technical Terms

Ryuichiro Kimura
Zi Long
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), 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 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. (2016) proposed a method that enables NMT to translate patent sentences comprising a large vocabulary of technical terms. The proposed NMT system is trained 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. The selected phrases are then replaced with tokens during training and post-translated by the phrase translation table of SMT. Based on the discussion as well as experimental evaluation results reported in Long et al. (2016), this paper further studies the effect of the proposed NMT model with phrase translation by the SMT model with respect to reducing untranslated content. The issue of untranslated content is among those most important problems of NMT. This paper employs the back translation probability which is proposed by Goto and Tanaka (2017) to apply to the task of detecting untranslated content in NMT. Then, we show the evaluation results of both predicting untranslated contents and of manually counting the numbers of words in the input Japanese sentences untranslated into English in the task of Japanese to English NMT, where the proposed NMT model with phrase translation by the SMT model outperforms the baseline NMT model.

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. 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.

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. To address this problem, Long et al. (2016) proposed extracting compound nouns as technical terms and replacing them with tokens. These compound nouns then are post-translated with the phrase translation table of the statistical machine translation (SMT) system. Based on the discussion as well as experimental evaluation results reported in Long et al. (2016), this paper further studies the effect of the proposed NMT model with phrase translation by the SMT model with respect to reducing untranslated content. The issue of untranslated content is among those most important problems of NMT. This paper employs the back translation probability which is proposed by Goto and Tanaka (2017) to apply to the task of detecting untranslated content in NMT. Then, we show the evaluation results of both predicting untranslated contents and of manually counting the numbers of words words in the input Japanese sentences untranslated into English in the task of Japanese to English NMT, where the proposed NMT model with phrase translation by the SMT model outperforms the baseline NMT model.

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 (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 Aligning Phrase Pairs by SMT Translation Model

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

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

2. For the Japanese technical terms whose English 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 $\langle S_s, S_t \rangle$ containing a Japanese technical term t_s, a sequence of English words is selected using SMT word alignment, and we use the English translation t_t for the Japanese technical term t_s. [2]

[1] 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 English side, on the other hand, we regard English translations of extracted Japanese compound nouns as English technical terms, where we do not regard other English phrases as technical terms.

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

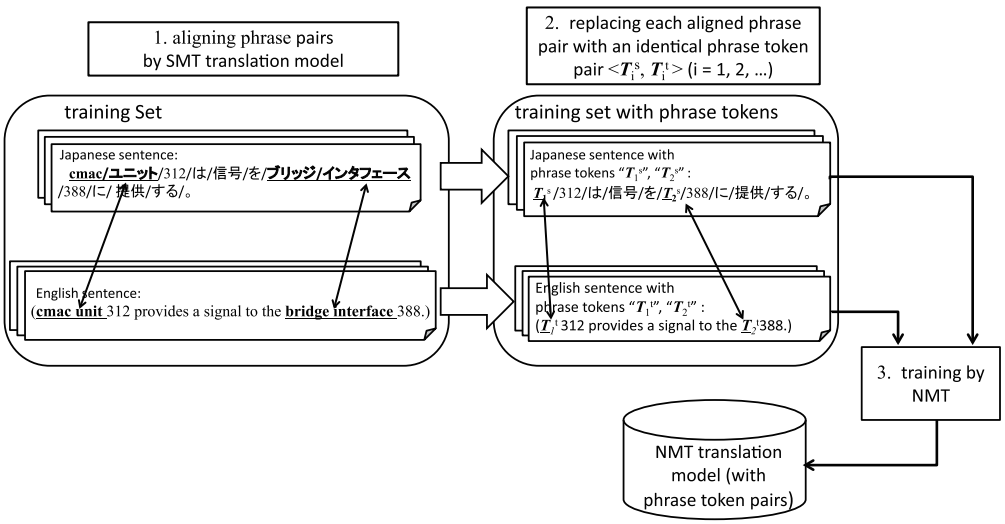

Figure 1: NMT training after replacing phrase pairs with token pairs $\langle T_i^s, T_i^t \rangle$ $(i = 1, 2, \ldots)$

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.

4.1 NMT Training after Replacing Phrase Pairs with Tokens

Figure 1 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 1, each source-target phrase pair, whose Japanese side is regarded as a compound noun, is aligned as described in Section 3. As shown in the step 2 of Figure 1, in each of the parallel patent sentence pairs, occurrences of phrase pairs $\langle t_1^s, t_1^t \rangle$, $\langle t_2^s, t_2^t \rangle$, $\ldots$, $\langle t_k^s, t_k^t \rangle$ are then replaced with token pairs $\langle T_1^s, T_1^t \rangle$, $\langle T_2^s, T_2^t \rangle$, $\ldots$, $\langle T_k^s, T_k^t \rangle$. Phrase pairs $\langle t_1^s, t_1^t \rangle$, $\langle t_2^s, t_2^t \rangle$, $\ldots$, $\langle t_k^s, t_k^t \rangle$ 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 $\langle T_1^s, T_1^t \rangle$, $\langle T_2^s, T_2^t \rangle$, $\ldots$ 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.[3]

4.2 NMT Decoding and SMT Phrase Translation

Figure 2 illustrates the procedure for producing target translations by decoding the input source sentence using the method proposed in this paper.

In the step 1 of Figure 2, 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 2, we

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

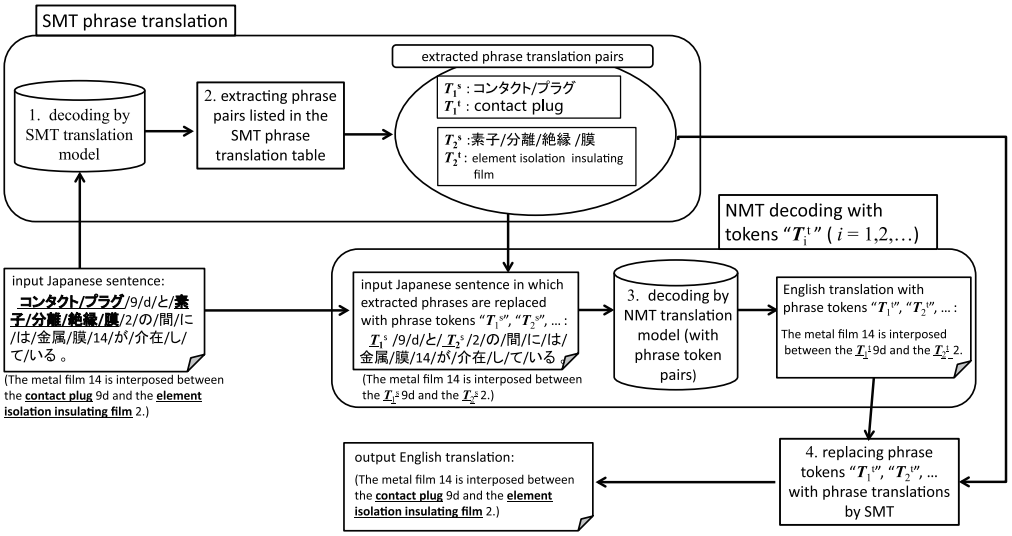

Figure 2: NMT decoding with tokens "T_i^s" ($i = 1, 2, \ldots$) and the SMT phrase translation

automatically extract the phrase pairs whose Japanese sides are regarded as compound nouns according to the procedure of Section 3. Extracted phrase pairs 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 2, 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 Resource and Evaluation Procedures

5.1 Patent Documents

Japanese-English patent documents are provided in the NTCIR-7 workshop (Fujii et al., 2008), which are collected from the 10 years of unexamined Japanese patent applications published by the Japanese Patent Office (JPO) and the 10 years patent grant data published by the U.S. Patent & Trademark Office (USPTO) in 1993-2000. The numbers of documents are approximately 3,500,000 for Japanese and 1,300,000 for English. From these document sets, patent families are automatically extracted and the fields of "Background of the Invention" and "Detailed Description of the Preferred Embodiments" are selected. Then, the method of Utiyama and Isahara (2007) is applied to the text of those fields, and Japanese and English sentences are aligned. The Japanese sentences were segmented into a sequence of morphemes using the Japanese morphological analyzer MeCab[4] with the morpheme lexicon IPAdic. In this study, out of the provided 1.8M Japanese-English parallel sentences, 1.1M parallel sentences whose Japanese sentences contain fewer than 40 morphemes and English sentences contain fewer than 40 words are used.

Table 1: Statistics of datasets

	training set	validation set	test set
Japanese-English	1,167,198	1,000	1,000

Table 2: Automatic evaluation results (BLEU)

System	ja → en
Baseline SMT (Koehn et al., 2007)	32.3
Baseline NMT	38.2
NMT with phrase translation by SMT	39.8

5.2 Training and Test Sets

We evaluated the effectiveness of the proposed NMT model at translating parallel patent sentences described in Section 5.1. Among the selected parallel sentence pairs, we randomly extracted 1,000 sentence pairs for the test set and 1,000 sentence pairs for the validation set; the remaining sentence pairs were used for the training set. Table 1 shows statistics of the dataset.

From the Japanese-English sentence pairs of the training set, we collected 2,785,108 occurrences of Japanese-English phrase pairs, which are 704,346 types of phrase pairs with unique 422,269 types of Japanese phrases and 511,633 unique types of English phrases. Within the total 1,000 Japanese patent sentences in the Japanese-English test set, 2,539 occurrences of Japanese phrases were extracted, which correspond to 2,171 types.

5.3 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 256 cells in each layer. The decoder is a three-layer deep LSTM with 256 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 256. 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 machine.

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

6 Evaluation Results with BLEU

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

[4] http://taku910.github.io/mecab/

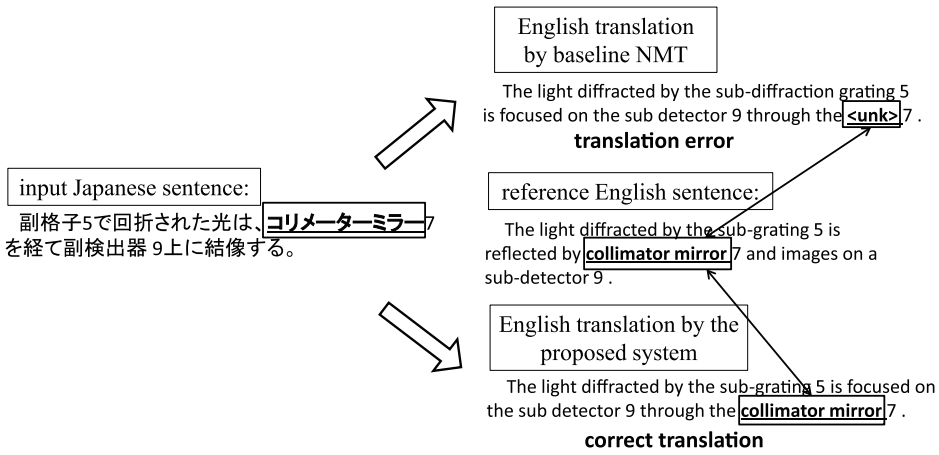

Figure 3: An example of correct translations produced by the proposed NMT model when addressing the problem of out-of-vocabulary words (Japanese-to-English)

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 our proposed approach as the baseline NMT. As shown in Table 2, the BLEU score obtained by the proposed NMT model is clearly higher than those of the baselines. When compared with the baseline SMT, the performance gains of the proposed system are approximately 7.5 BLEU points. When compared with the result of the baseline NMT, the proposed NMT model achieved performance gains of 1.6 BLEU points. Those evaluation details including other language pairs Japanese to Chinese, Chinese to Japanese, and English to Japanese are reported in Long et al. (2016, 2017). It is also important to note that we evaluate whether the phrase tokens are translated from the source sentences to the target sentences in the step 3 of Figure 2, without being missed during phrase token translation by the proposed NMT model. In this evaluation, the proposed NMT model achieved to miss no phrase token during the proposed NMT procedure.

Figure 3 compares an example of correct translations produced by the proposed system with those produced by the baseline NMT.

7 Effect on Reducing Untranslated Content

7.1 Predicting Untranslated Content

Goto and Tanaka (2017) proposed methods of detecting untranslated content within a framework of neural machine translation as well as that of improving translation evaluation results in terms of BLEU by reranking based on translation scores which are designed to minimize untranslated content. Two types of probabilities are studied in the task of detecting untranslated content, out of which we employed the *back translation probability*. More specifically, we evaluate the back translation probabilities of translation by both the baseline NMT model as well as the proposed NMT model with phrase translation by the SMT model. Then, we show that we achieve improvement in terms of the results of predicting untranslated content based on the back translation probabilities.

7.1.1 Back Translation Probability

The back translation BT is defined as the forced decoding from an MT output to its input sentence. When the content of a source word is missing in the MT output, the BT probability of the source word is expected to be small. This expectation is used as a clue for predicting

Table 3: Evaluation Results of Predicting Untranslated Contents (for the test set)

(a) Back translation ratio score BT-R averaged over the test set

System	ja → en
Baseline NMT	16.3
NMT with phrase translation by SMT	14.0

(b) Distribution of the difference of back translation ratio score between the proposed NMT with phrase translation by SMT and the baseline NMT (BT-R($\boldsymbol{x}$, $\boldsymbol{y}^d$) − BT-R($\boldsymbol{x}$, $\boldsymbol{y'}^d$)) (%) (over the test set)

< 0					> 0				
< -20	$-20 \sim -10$	$-10 \sim -5$	$-5 \sim -1$	$-1 \sim 0$	$0 \sim 1$	$1 \sim 5$	$5 \sim 10$	$10 \sim 20$	> 20
4.9	8.4	12.3	19.1	12.9	12.9	14.8	8.0	4.4	2.3
57.6					42.4				

untranslated content. A BT probability score (BT-P) b_j^d based on the BT probability of an input word x_j $(1 \leq j \leq N)$ from an n-best MT output $\boldsymbol{y}^d$ $(1 \leq d \leq n)$ is given as below:

$$b_j^d \quad = \quad -\log p(x_j | x_{<j}, \boldsymbol{y}^d)$$

For both of the MT outputs produced by the baseline NMT model as well as the proposed NMT model with phrase translation by the SMT model, this probability is calculated based on the baseline NMT model presented in section 2, while the NMT model is trained in the direction of English to Japanese translation with the training set without phrase tokens.

In this BT probability formalization, the following assumption of the "existence of translations" is employed:

Assumption: Existence of translations The translation of an arbitrary input word x_j $(1 \leq j \leq N)$ exists somewhere in the n-best outputs $\boldsymbol{y}^d$ $(1 \leq d \leq n)$, except when x_j does not inherently correspond to any target words.

and accordingly, $\min_{1 \leq d' \leq n} b_j^{d'}$ is assumed to be the score of an output that contains the content of x_j simply because it is assumed that the n-best output with the minimum score is most likely to contain the content of x_j. Then, as a score of missing the content of x_j from $\boldsymbol{y}^d$, a score based on a probability ratio is introduced and the BT ratio score (BT-R) q_j^d is defined below:

$$q_j^d \quad = \quad b_j^d - \min_{1 \leq d' \leq n} b_j^{d'}$$

which is the difference of the BT probability score of the output and that of the n-best output with the minimum score, being assumed to contain the content of x_j Finally, by summing this score across all the input words within the whole source sentence $\boldsymbol{x} = (x_1, \ldots, x_N)$, the BT ratio score (BT-R) of the MT output $\boldsymbol{y}^d$ for the source sentence $\boldsymbol{x}$ is obtained below:

$$\text{BT-R}(\boldsymbol{x}, \boldsymbol{y}^d) \quad = \quad \sum_j q_j^d$$

7.1.2 Prediction Results

We measured the back translation ratio score BT-R for the test set, averaged over the test set for both the proposed NMT model with phrase translation by the SMT model and the baseline

Table 4: Manual Evaluation Results on the Numbers of Words in the Input Japanese Sentences Untranslated into English (for the 100 test sentences)

(a) Numbers of words in the input Japanese sentences untranslated into English

System	ja → en
Baseline NMT	73
NMT with phrase translation by SMT	51

(b) Distribution of the numbers of untranslated words (%)

System	numbers of untranslated words										
	0	1	2	3	4	5	6	7	8	9	≥ 10
Baseline NMT	64	24	6	2	0	1	1	0	1	0	1
NMT with phrase translation by SMT	74	14	4	4	3	1	0	0	0	0	0

model as shown in Table 3. As can be seen from this result, the proposed NMT model with phrase translation by the SMT model achieves the BT-R score lower than that of the baseline NMT model. This result shows that the BT-R score predicts less untranslated content within those MT outputs by the proposed NMT model with phrase translation by the SMT model than those by the baseline NMT model.

Next, for each test sentence $\boldsymbol{x}$ within the test set, we measure the difference of the BT-R score between the proposed NMT with phrase translation by the SMT model and the baseline NMT model as below where BT-R($\boldsymbol{x}$, $\boldsymbol{y}^d$) and BT-R($\boldsymbol{x}$, $\boldsymbol{y'}^d$) are the BT-R scores of the proposed NMT with phrase translation by the SMT model and the baseline NMT model, respectively:

$$\text{BT-R}(\boldsymbol{x}, \boldsymbol{y}^d) - \text{BT-R}(\boldsymbol{x}, \boldsymbol{y'}^d)$$

The distribution of those differences over the test set is shown in Table 3, where, for 57.6% of the test set, the BT-R score of the MT outputs by the proposed NMT model with phrase translation by the SMT model is smaller than that of the MT outputs by the baseline NMT model. Out of those 57.6% test sentences, 25.6% have the absolute value of the difference of the BT-R score greater than 5. For the remaining 42.4% test sentences for which the BT-R score of the MT outputs of the baseline NMT model is smaller than that of the MT outputs by the proposed NMT model, only 14.7% have the absolute value of the difference of the BT-R score greater than 5.

7.2 Manual Evaluation on the Numbers of Words in the Input Japanese Sentences Untranslated into English

For the 100 test sentences selected at random, we counted the numbers of words in the input Japanese sentences untranslated into English in the task of Japanese to English NMT. As shown in Table 4, the number of words untranslated by the baseline NMT reduced to around 70% . Table 4 also shows the distribution of the numbers of untranslated words. The proposed NMT model with phrase translation by the SMT model contributes to reducing untranslated content within the MT outputs, mainly because part of untranslated source words are out-of-vocabulary, and thus are untranslated by the baseline NMT. The proposed system 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.

Table 5: An Example of Reducing Untranslated Content by the Proposed NMT Model with Phrase Translation by the SMT Model (Untranslated parts are underlined)

(a) Translation by the proposed NMT model with phrase translation by the SMT model

Input Japanese Sentence	プロジェクター ユニット 12 は 、 光源 16 と 、 空間 変調 素子 として の 液晶 表示 素子 18 a 、 18 b 、 18 c と 、 投写 レンズ 20 と を 含む 。
Reference English Translation	the projection unit 12 includes a light source 16 , liquid crystal display elements 18a , 18b , and 18c as space modulation elements , and a projection lens 20 .
MT Output with Phrase Tokens	the T_1^t 12 includes a light source 16 , T_3^t 18a , 18b and 18c as a T_2^t and a T_4^t 20 .
MT Output	the projection unit 12 includes a light source 16 , liquid crystal display element 18a , 18b and 18c as a spatial modulating element and a projection lens 20 .
BT-R score	13.2

(b) Translation by the baseline NMT model

Input Japanese Sentence	プロジェクター ユニット 12 は 、 光源 16 と 、 空間 変調 素子 として の 液晶 表示 素子 18 a 、 18 b 、 18 c と 、 投写 レンズ 20 と を 含む 。
Reference English Translation	the projection unit 12 includes a light source 16 , liquid crystal display elements 18a , 18b , and 18c as space modulation elements , and a projection lens 20 .
MT Output	the projector unit 12 includes a light source 16 , liquid crystal display elements 18a , 18b and 18c , and a projection lens 20 .
BT-R score	47.5

Table 5 compares examples of the MT output by the proposed NMT model with phrase translation by the SMT model with that by the baseline NMT model. It is clearly shown that the proposed NMT model with phrase translation by the SMT model successfully reduced untranslated contents compared with the baseline NMT model.

8 Conclusion

Long et al. (2016) proposed a method that enables NMT to translate patent sentences comprising a large vocabulary of technical terms. The proposed NMT system is trained 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. The selected phrases are then replaced with tokens during training and post-translated by the phrase translation table of SMT. Based on the discussion as well as experimental evaluation results reported in Long et al. (2016), this paper further studied the effect of the proposed NMT model with phrase translation by the SMT model with respect to reducing untranslated content. We showed the evaluation results of both predicting untranslated contents and of manually counting the numbers of words in the input Japanese sentences untranslated into English in the task of Japanese to English NMT, where the proposed NMT model with phrase translation by the SMT model outperformed the baseline NMT model. Our future tasks include integrating the reranking framework of Goto and Tanaka (2017) which is based on translation scores designed to minimize untranslated content into the proposed NMT model with phrase translation by the SMT model. One of another important future tasks is to compare the proposed NMT model with phrase translation by the SMT model with that based on subword units (e.g. Sennrich et al. (2016)).

References

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

Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. EMNLP*, pages 1724–1734.

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

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

Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T. (2008). Toward the evaluation of machine translation using patent information. In *Proc. 8th AMTA*, pages 97–106.

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

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

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

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

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

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

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

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

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

Luong, M., Pham, H., and Manning, C. D. (2015a). Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*, pages 1412–1421.

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

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

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

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural machine translation. In *Proc. 27th NIPS*, pages 3104–3112.

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