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

Zi Long† Takehito Utsuro† Tomoharu Mitsuhashi‡ Mikio Yamamoto†
†Graduate School of Systems and Information Engineering, University of Tsukuba
‡Japan Patent Information Organization

1 Introduction

Neural machine translation (NMT), a new approach to solving machine translation, has achieved promising results [8, 1]. However, a 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. 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. Among them, Luong et al. [5] 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. However, this previous approach has limitations when translating patent sentences. This is because their method only focuses 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.

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. [8], and 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. [8], 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.

2 Neural Machine Translation

Neural Machine Translation (NMT) uses a single neural network trained jointly to maximize the translation performance [8, 1]. 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 $p(y_l \mid y_{<l}, \boldsymbol{x})$ $(1 \leq l \leq M)$. Consequently, it becomes possible to compute and maximize the log probability of the target sentence given the source sentence

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

In this paper, we use an NMT model similar to that used by Bahdanau et al. [1], which consists of a bidirectional long short-term memory (LSTM) as an encoder and a decoder that predicts target words 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 of encoder.

3 NMT with a Large Technical Term Vocabulary

3.1 NMT Training after Replacing Technical Term Pairs with Tokens

Figure 1 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 1, we align the

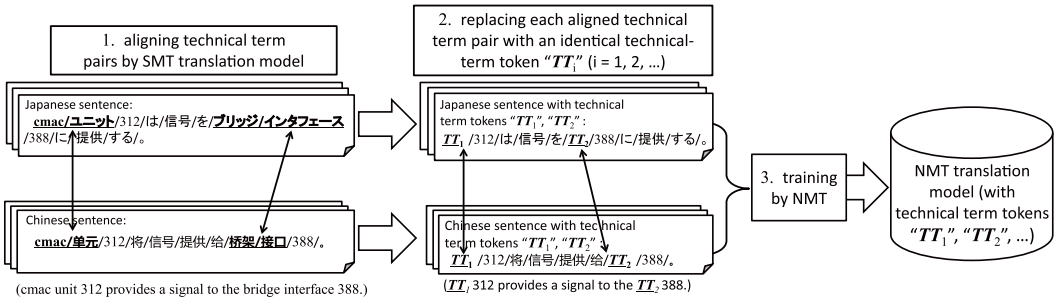

Figure 1: NMT training after replacing technical term pairs with technical term tokens "TT_i" ($i = 1, 2, \ldots$)

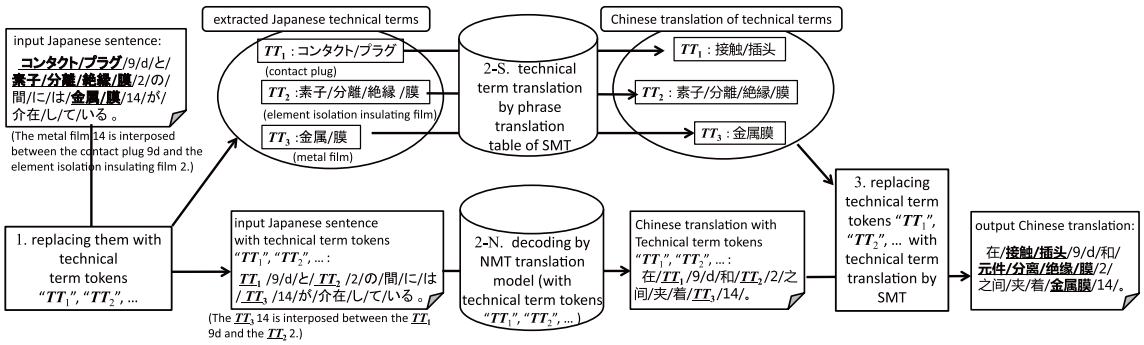

Figure 2: NMT decoding with technical term tokens "TT_i" ($i = 1, 2, \ldots$) and SMT technical term translation

Japanese technical terms, which are automatically extracted from the Japanese sentences, with their Chinese translations in the Chinese sentences.[1] As shown in the step 2 of Figure 1, 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.[2]

3.2 NMT Decoding and SMT Technical Term Translation

Figure 2 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 2, 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 2, the source Japanese sentence with technical term tokens is translated using the NMT model trained according to the procedure described in Section 3.1, whereas the extracted Japanese technical terms are translated using an SMT phrase translation table in the step 2-S of Figure 2.[3] Finally, in the step 3, we replace the technical term to-

[1] Details of the procedure of identifying technical term pairs in the bilingual Japanese-Chinese corpus can be found in the work of Long et al. [4].

[2] We treat the NMT system as a black box, and the strategy we present in this paper could be applied to any NMT system [8, 1].

[3] 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 of Japanese technical terms.

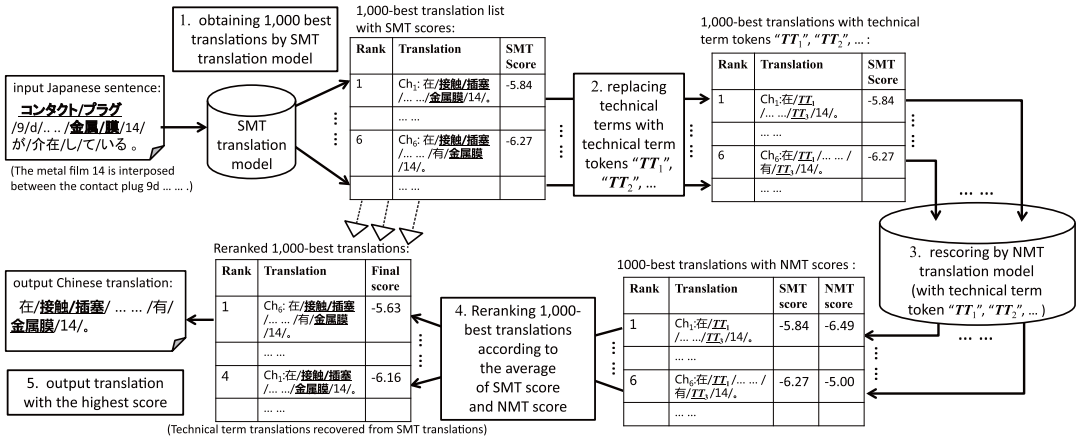

Figure 3: NMT rescoring of 1,000-best SMT translations with technical term tokens "TT_i" ($i = 1, 2, \ldots$)

kens "TT_i" ($i = 1, 2, \ldots$) of the sentence translation with SMT the technical term translations.

3.3 NMT Rescoring of 1,000-best SMT Translations

As shown in the step 1 of Figure 3, similar to the approach of NMT rescoring provided in Sutskever et al. [8], 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 3.1. In the step 3 of Figure 3, 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 3.1. 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 score.

4 Evaluation

4.1 Training and Test Sets

We evaluated the effectiveness of the proposed NMT system with the corpus of 2.8M Japanese-Chinese parallel patent sentences described in Long et al. [4]. 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 3.1, 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. 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.

4.2 Training Details

For the training of the SMT model, including the word alignment and the phrase translation table, we used Moses [3], 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 Bahdanau et al. [1]. The encoder consists of forward and backward deep LSTM neural networks each having three layers, with 512 cells in each layer. The decoder is a three layer deep LSTM with 512 cells in each layer, as well. Further training details are found in the work of Long et al. [4].

4.3 Evaluation Results

We calculated automatic evaluation scores for the translation results using two popular metrics: BLEU [7] and

Table 1: Automatic evaluation results

System	NMT decoding		NMT rescoring	
	BLEU	RIBES	BLEU	RIBES
Baseline SMT [3]	52.5	88.5	-	-
Baseline NMT	55.5	90.1	55.9	89.1
NMT with technical term translation by SMT	**56.8**	**91.1**	56.2	89.6
NMT with PosUnk model [5]	55.7	90.9	56.0	89.4

Table 2: Human evaluation results [PE: Pairwise Evaluation (scores range from -100 to 100) and JAE: JPO Adequacy Evaluation (scores range from 1 to 5)]

System	NMT decoding		NMT rescoring	
	PE	JAE	PE	JAE
Baseline SMT [3]	-	3.5	-	-
Baseline NMT	23.0	4.2	30.0	4.2
NMT with technical term translation by SMT	**39.5**	**4.6**	31.5	4.2

RIBES [2]. As shown in Table 1, we report the evaluation scores, on the basis of the translations by Moses [3], as the baseline SMT[4] 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 4.3 BLEU points and 2.6 RIBES if translations are produced by the proposed NMT system of Section 3.2. When compared with the result of decoding with the baseline NMT, the proposed NMT system of Section 3.2 achieved performance gains of 1.3 BLEU points and 1.0 RIBES points. When compared with the result of reranking with the baseline NMT, the proposed NMT system of Section 3.3 can still achieve performance gains of 0.3 BLEU points. From Table 1, we also observed that the proposed systems of Section 3.2 achieved a better performance than the system of Section 3.3.

Furthermore, we quantitatively compared our study with the work of Luong et al. [5]. As the result shown in Table 1, compared with the NMT system with PosUnk model that is proposed as the best model by Luong et al. [5], the proposed NMT system achieves performance gains of 1.1 BLEU points and 0.2 RIBES points when the output translations are produced by NMT decoding and SMT technical term translation described in Section 3.2.

In this study, we also conducted two types of human evaluation according to Nakazawa et al. [6]: pairwise evaluation and JPO adequacy evaluation[5]. 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 systems achieved the best performance for both pairwise evaluation and JPO adequacy evaluation.

5 Conclusion

In this paper, we proposed an NMT method capable of translating patent sentences with a large vocabulary of technical terms by training an NMT system on a bilingual corpus, wherein technical terms are replaced 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. As future work, we will also evaluate the present study by reranking translations from both the n-best SMT translations and n-best NMT translations, where the translation with the highest average of SMT score and NMT score is expected to be an effective translation.

References

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

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

[3] 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. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

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

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

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

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

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

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

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