Patent NMT integrated with Large Vocabulary Phrase Translation by SMT

Division of Intelligent Interaction Technologies, Faculty of Engineering, Information and Systems, University of Tsukuba

Takehito Utsuro

Takehito Utsuro is a professor at the Division of Intelligent Interaction Technologies, Faculty of Engineering, Information and Systems, University of Tsukuba, since 2012. His professional interests in natural language processing, Web intelligence, information retrieval, machine learning, spoken language processing, and artificial intelligence.

Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba

Zi Long

Zi Long is a student of doctor course in Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba.

Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba

Ryuichiro Kimura

Ryuichiro Kimura is a student of master course in Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba.

Division of Information Engineering, Faculty of Engineering, Information and Systems, University of Tsukuba

Mikio Yamamoto

Mikio Yamamoto is a professor at the Division of Information Enginieering, Faculty of Engineering, Information and Systems, University of Tsukuba, since 2008. His professional interests in natural language processing and machine translation.

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 article, we present 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 Bahdanau el al.[1], 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 Bahdanau et al.[1], 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)[3][4].

2 Neural Machine Translation

NMT uses a single neural network trained jointly to maximize the translation performance[8][1]. Given a source sentence x ($x=x_1, \cdots x_N$) and target sentence y ($y=y_1, \cdots y_M$), an NMT model uses a neural network to parameterize the conditional distributions

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

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

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

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

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

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 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", ….[1] In the step 1 of Figure 1, we align the source technical terms, which are automatically extracted from the source sentences, with their

1 In this work, we approximately regard all the Japanese compound nouns as source 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 target side, on the other hand, we regard target translations of extracted Japanese compound nouns as target technical terms, where we do not regard other target phrases as technical terms.

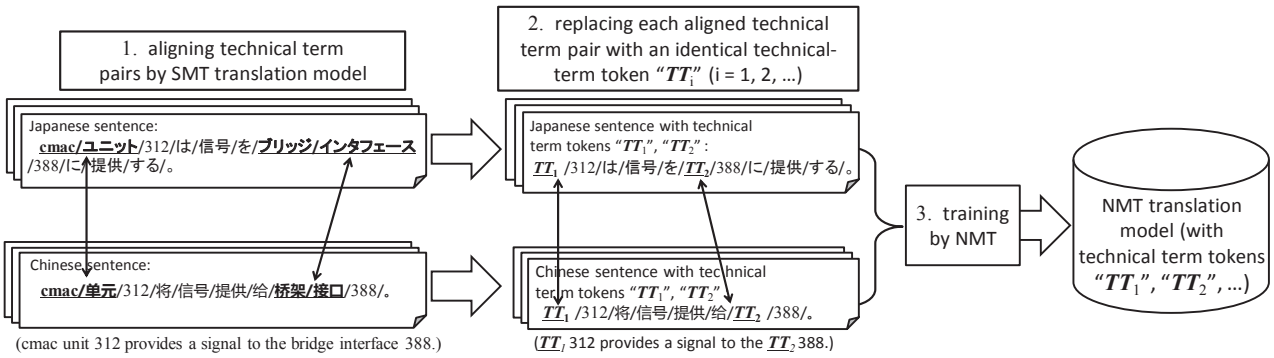

Figure 1　NMT training after replacing technical term pairs with tokens "TT_1", "TT_2", ⋯

target translations in the target sentences.[2] As shown in the step 2 of Figure 1, in each of source-target parallel patent sentence pairs, occurrences of technical term pairs $\langle t_S^1, t_T^1\rangle$, $\langle t_S^2, t_T^2\rangle$, ⋯, $\langle t_S^k, t_T^k\rangle$ 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 $\langle t_S^1, t_T^1\rangle$, $\langle t_S^2, t_T^2\rangle$, ⋯, $\langle t_S^k, t_T^k\rangle$ are numbered in the order of occurrence of source technical terms t_S^i (i=1, 2, ⋯, k) in each source sentence S_S. Here, note that in all the parallel sentence pairs $\langle S_S, S_T\rangle$, 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 source patent sentences S_S, the source technical term t_S^1 which appears earlier than other source 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, ⋯, k) tokens and obtain an NMT model in which the technical terms are represented as technical term tokens.[3]

3.2　NMT Decoding and SMT Technical Term Translation

Figure 2 illustrates the procedure for producing target translations via decoding the source sentence using the method presented in this aritcle. In the step 1 of Figure 2, when given an input source sentence, we first automatically extract the technical terms and replace them with the technical term tokens "TT_i" (i=1, 2, ⋯, k). Consequently, we have an input sentence in which the technical term tokens "TT_i" (i=1, 2, ⋯, k) represent the positions of the technical terms and a list of extracted source technical terms. Next, as shown in the step 2-N of Figure 2, the source sentence with technical term tokens is translated using the NMT model trained according to the procedure described in Section 3.1, whereas the extracted source technical terms are translated using an SMT phrase translation table in the step 2-S of Figure 2.[4] Finally, in the step 3, we replace the technical term tokens "TT_i" (i=1, 2, ⋯, k) of the sentence translation with SMT the technical term translations.

2　Details of the procedure of identifying technical term pairs in the bilingual corpus can be found in the work of Long et al.[3].

3　We treat the NMT system as a black box, and the strategy we present in this article could be applied to any NMT system. [7][1]

4　We use the translation with the highest probability in the phrase translation table. When an input source 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 target translation is generated compositionally from the constituents of source technical terms.

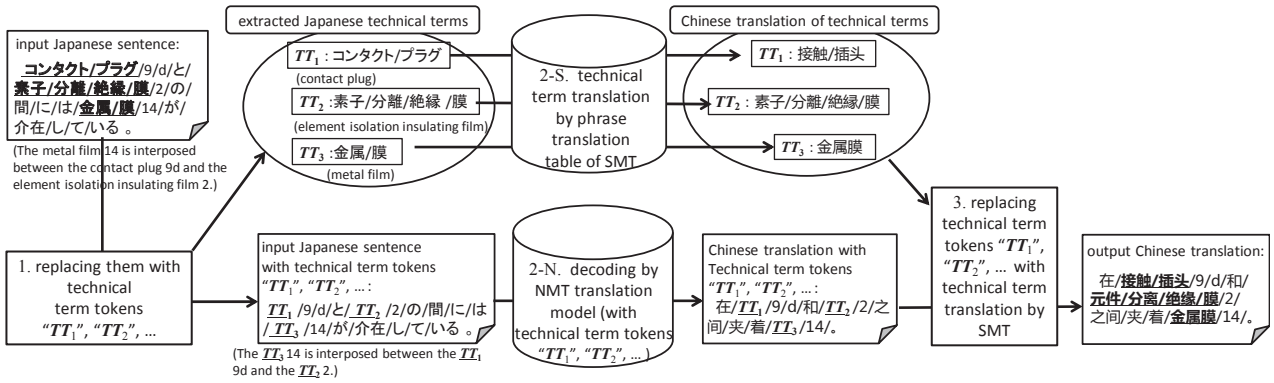

Figure 2　NMT decoding with technical term tokens "TT_i" (i=1, 2, ⋯, k) and SMT technical term

4 Evaluation

4.1 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 Uchiyama and Isahara[9] was applied[5] 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[6] with the morpheme lexicon IPAdic,[7] and the Chinese sentences were segmented into a sequence of words using the Chinese morphological analyzer Stanford Word Segment[10] 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 Chinese sentences contain fewer than 40 words.[8]

Japanese-English patent documents are provided in the NTCIR-7 workshop[11], 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 Uchiyama and Isahara[9] 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 with the morpheme lexicon IPAdic. Similar to the case of Japanese-Chinese patent

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

6　http://mecab.sourceforge.net/

7　http://sourceforge.jp/projects/ipadic/

8　It is expected that our NMT model can improve the baseline NMT without our technique when translating longer sentences that contain more than 40 morphemes / words. It is because the approach of replacing phrases with tokens also shortens the input sentences, expected to contribute to solving the weakness of NMT model when translating long sentences.

documents, 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.

4.2 Training and Test Sets

We evaluated the effectiveness of the NMT model presented in this article at translating parallel patent sentences described in Section 4.1. 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 datasets.

Table 1 Statistics of datasets

	training set	validation set	test set
ja ↔ ch	2,877,178	1,000	1,000
ja ↔ en	1,167,198	1,000	1,000

4.3 Training Details

For the training of the SMT model, including the word alignment and the phrase translation table, we used Moses[2], 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.[1]. 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.[8], 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.

4.4 Evaluation Results

We calculated automatic evaluation scores for the translation results using a metric called BLEU[7]. As shown in Table 2, we report the evaluation scores, on the basis of the translations by Moses[2], as the baseline SMT.[9] and the scores based on translations produced by the equivalent NMT system without our approach as the baseline NMT. As shown in Table 2, our NMT systems clearly improve the translation quality when compared with the baselines. When compared with the baseline SMT, the performance gain of our system is approximately 6.1 BLEU points when translating Japanese into Chinese and 8.4 BLEU when translating Japanese into English. When compared with the result of decoding with the baseline NMT, our NMT system achieved performance gains of 2.1 BLEU points when translating

9 We train the SMT system on the same training set and tune it with the validation set.

Table 2　Automatic evaluation results (BLEU)

System	ja → ch	ja → en
Baseline SMT[2]	52.5	32.3
Baseline NMT	56.5	39.9
NMT with PosUnk model[5]	56.9	40.1
NMT with technical term translation by SMT	58.6	40.7

Table 3　Human evaluation results [PE: Pairwise Evaluation (scores range from −100 to 100) and JAE: JPO Adequacy Evaluation (scores range from 1 to 5)]

System	ja → ch		ja → en	
	PE	JAE	PE	JAE
Baseline SMT[2]	−	3.5	−	3.1
Baseline NMT	23.0	4.2	21.0	3.9
NMT with technical term translation by SMT	30.5	4.3	29.5	4.0
NMT with PosUnk model[5]	37.0	4.5	33.5	4.1

Japanese into Chinese, and 0.8 BLEU points when translating Japanese into English.

Furthermore, we quantitatively compared our study with the work of Luong et al.[5]. As the result shown in Table 2, compared with the NMT system with PosUnk model that is proposed as the best model by Luong et al.[5], our NMT system achieves performance gains of 1.7 BLEU points when translating Japanese into Chinese and 0.6 BLEU points when translating Japanese into English.

In this study, we also conducted two types of human evaluation according to Nakazawa et al.[6]: pairwise evaluation and JPO adequacy evaluation.[10] Table 3 shows the results of the human evaluation for the baseline SMT, the baseline NMT, the NMT system with PosUnk[5] and our NMT system. We observed that our systems achieved the best performance for both pairwise evaluation and JPO adequacy evaluation.

10 https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/tokkyohonyaku_hyouka/01.pdf (in Japanese)

5 Conclusion

In this article, we presented 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 NMT system performs better than the phrase-based SMT system as well as the equivalent NMT system without our approach.

Reference

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

[2] 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.: Moses: Open Source Toolkit for Statistical Machine Translation, in Proc. 45th ACL, Companion Volume, pp. 177–180 (2007)

[3] Long, Z., Utsuro, T., Mitsuhashi, T., and Yamamoto, M.: Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation, in Proc. 3rd WAT, pp. 47–57 (2016)

[4] Long, Z., Kimura, R., Utsuro, T., Mitsuhashi, T., and Yamamoto, M.: Neural Machine Translation Model with a Large Vocabulary Selected by Branching Entropy, in Proc. MT Summit XVI, (2017)

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

[6] Nakazawa, T., Mino, H., Goto, I., Neubig, G., Kurohashi, S., and Sumita, E.: Overview of the 2nd Workshop on Asian Translation, in Proc. 2nd WAT, pp. 1–28 (2015)

[7] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, in Proc. 40th ACL, pp. 311–318 (2002)

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

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

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

[11] 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.

SMT による大語彙フレーズ翻訳との併用による ニューラルネットワーク機械翻訳（抄録）

Takehito Utsuro, Zi Long, Ryuichiro Kimura, Mikio Yamamoto

近年，従来の統計的機械翻訳（Statistic Machine Translation；SMT）に代わって，ニューラルネットワーク機械翻訳（Nerual Machine Translation；NMT）モデルが盛んに研究されている．NMT は，原言語文を固定長ベクトルへ写像し，その固定長ベクトルから目的言語文を生成するため，意味的要素の翻訳に非常に優れており，SMT を上回る翻訳精度を達成している．しかしながら，NMT の弱点の一つとして，扱える語彙に限りがある点が知られている．具体的には，扱う語彙のサイズの増加に伴い，NMT モデルの訓練および翻訳に要する時間が増す点が課題となっている．

NMT においては，語彙辞書に含まれていない単語は未知語トークンとして出力されるため，これが誤訳となる．そこで，これまでにも，NMT が扱える語彙の規模を拡大する方式について研究が行われてきた．文献 [5] では，訓練用対訳文における単語対応の情報に基づいて，語彙辞書に含まれていない未知語単語を，単語間の対応関係を特定できるトークンに置き換えた後，NMT の訓練を行う方式を提案した．この方式では，出力文に含まれたトークンから未知語が対応する原言語の単語を推定し，その訳語に置き換えることによって，NMT の出力文において出力可能となる語彙の規模を拡大した．しかし，文献 [5] の方式は，単語単位での語彙規模の拡大にとどまる点が弱点であった．この弱点のため，複合語によって構成される専門用語が多数含まれた特許文の翻訳精度の改善においては限界があった．

以上の背景のもとで，本稿では，特許文を対象としたニューラルネットワーク翻訳において，大規模専門用語語彙に対応する方式 [3][4] について述べる．本方式においては，訓練用対訳文において専門用語間の二言語対応の情報を収集し，二言語間で対応済みの専門用語対訳対を同一のトークンに置き換えた後，NMT の訓練を行う．本方式による特許文の翻訳時には，専門用語以外の部分に対しては，NMT モデルによる訳文生成がなされ，一方，専門用語部分に対しては，SMT モデルによる翻訳がなされる．本方式を用いない従来型の NMT モデルと，本方式との間で翻訳精度の比較を行った結果，本方式によって従来型の NMT モデルの翻訳精度が改善することができた．

謝辞

本稿で述べた研究においては，日本特許情報機構（Japio）より提供して頂いたパテントファミリーのデータを利用させて頂いた．関係各位に感謝の意を表す．