Named Entity-Factored Transformer for Proper Noun Translation

Kohichi Takai†‡ Gen Hattori‡ Akio Yoneyama‡
Keiji Yasuda†‡ ♦ Katsuhito Sudoh† Satoshi Nakamura†

†NARA Institute of Science and Technology ‡KDDI Research, Inc ♦MINDWORD, Inc
{ ko-takai , ge-hattori , yoneyama }@kddi-research.jp
{ ke-yasuda , sudoh , s-nakamura }@dsc.naist.jp

Abstract

Subword-based neural machine translation is almost free from out-of-vocabulary (OOV) words. However, it does not always work well for composing proper nouns. We propose a method to use Named Entity (NE) features with Factored Transformer for accurate proper noun translation. The NE features are extracted from NE recognition on input sentences. Our experimental results showed the proposed method outperformed the baseline subword-based Transformer in BLEU and proper noun translation accuracy.

1 Introduction

Recent advances in neural machine translation (NMT) have made machine translation (MT) systems useful in practical applications. However, translation of proper nouns still remains difficult in spite of its importance in practice. Proper nouns are sometimes processed as out-of-vocabulary (OOV) words in MT systems due to the limitation of the vocabulary size and data sparseness. Approaches for proper noun translation can be divided roughly into two approaches: the use of hand-crafted bilingual lexicon as the external knowledge and the use of subwords.

The former approach uses a bilingual proper noun dictionary to translate proper nouns. Okuma et al (2008) proposed replacement-based proper noun translation. Their method uses a bilingual dictionary whose entries are associated with proper noun classes to replace a proper noun with another surrogate proper noun that frequently appears in the training corpus. Another method called lexically constrained decoder (LCD) (Hokamp et al., 2017) guarantees that proper nouns are translated into the target language sentence constrained by a bilingual dictionary (Chen et al., 2020, Chousa et al., 2021). It extends the beam search algorithm to find the hypothesis that contains all of the proper nouns (Hokamp et al., 2017). The dictionary-based approach works well only if the proper nouns to be translated are included in the bilingual dictionary and requires efforts for developing the dictionary.

In NMT, the subword-based approach is widely used. Sennrich et al (2016) proposed the use of subwords to decompose a word into shorter units. The method decreases the number of OOV words and keeps the translation quality if input sentences include OOV words. However, the subword-based NMT does not always work on a proper noun translation due to wrong compositions of subword translations.

In this paper, we propose a method for NMT focusing on the proper noun translation using Factored Transformer with named entity (NE) features. The proposed method only uses a parallel corpus and an NE recognition (NER) model as external knowledge.

2 Related Work

2.1 Factored NMT

Factored NMT (García-Martínez et al., 2016) integrates linguistic information into an NMT decoder. It decomposes morphological and grammatical features of a word into factors. Jordi et al. (2019) proposed Factored Transformer as an extension of Transformer (Vaswani et al., 2017) for low-resource NMT. The outputs from its subword and factor embedding layers are combined.

2.2 Named Entity Recognition

NER identifies and classifies proper nouns in a sentence. Recent studies in NER use neural networks as well. Huang et al., (2015) used Long Short-Term Memory (LSTM) and Conditional Random Field (CRF). Arkhipov et al., (2019) used BERT (Devlin et al., 2018) for NER. BERT utilizes

Proceedings of the 18th International Conference on Natural Language Processing, pages 7–11
Silchar, India. December 16 - 19, 2021.

a multilayer bidirectional transformer encoder which can learn deep bi-directional representations and can be fine-tuned for various NLP tasks later.

With respect to named entities, the use of a class-based language model was proposed to solve the problem of data sparseness in the field of automatic speech recognition (ASR) research field (Yamamoto et al., 1999, 2004). This idea was extended to MT (Tonoik et al., 2005, Yasuda et al., 2017). This approach improved the translation performance for unknown and low-frequency words by using high-frequency surrogate words in the same category.

The main focus of this paper is proper noun translation in the subword-based NMT using Factored Transformer and NE features from NER without a bilingual dictionary.

3 Proposed Method

We propose the use of NE features as linguistic factors of Factored Transformer for accurate proper noun translation.

3.1 Named Entity Feature Vector

NE features obtained from NER on a source language sentence are injected into the embedding or encoder layer, as a factor. We can use two types of NE features: a one-hot NE vector and an NE probability distribution vector. We extract the NE features from an input source language sentence by the following steps.

1. Apply word segmentation into an input source language sentence.
2. Apply subword segmentation onto the word-segmented input.
3. Apply part-of-speech (POS) tags to the word-segmented input using a POS tagger.
4. Recognize NE in the POS-tagged input sentence to obtain one-hot NE vectors or NE probability distribution vectors.
5. Align those NE feature vectors with the subwords composing corresponding words.

Here, a one-hot NE vector represents a 1-best NE category while an NE probability distribution vector represents the ambiguity of NE categories.

3.2 Factored Transformer Architecture

We propose a Factored Transformer model that uses two factors: subwords and NE features. We present two types of NE features and the two model variants in factor-injecting layers, as shown in Fig.1.

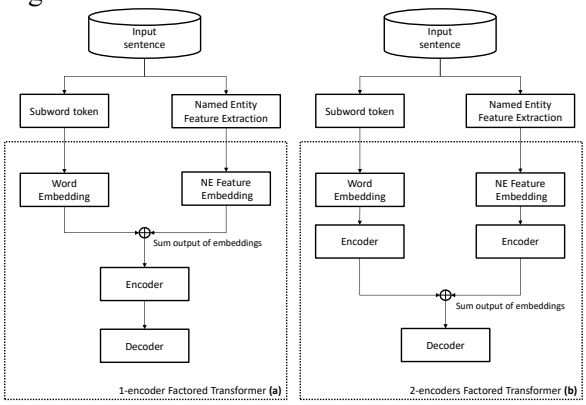

Figure 1: 1-encoder (a) and 2-encoders models (b)

1-encoder model (Fig. 1(a)):
Each factor has its own embedding layer. The embedding vectors are summed up together with the corresponding positional encoding vector and sent to the following encoder layer. The rest of the model remains unchanged from the *vanilla* Transformer.

2-encoders model (Fig. 1 (b)):
Each factor has its own encoder in addition to the embedding layer. The outputs from the encoders are summed up and used as encoder outputs. The rest of the model remains unchanged.

4 Experimental Settings

We conducted Japanese-to-English and English-to-Japanese MT experiments to compare the performance of the proposed method with a standard NMT method.

4.1 Named Entity Recognition Model

For the Japanese-to-English experiments, we used an NER model based on a pre-trained BERT model and fine-tuned it using Japanese NER training data generated by using the method presented by Takai et al., (2018). Table 1 shows the detailed parameter settings of the NER model.

parameter	mini batch size	epoch	optimizer
BERT-NER	**32**	**4**	**Adam**

Table 1: Detail of NER Hyperparameter

For the English-to-Japanese experiments, we used the NER module of Stanza[1]. In the experiments, the Japanese NER model had 33 categories, and the English NER model had 77 categories.

As a bilingual corpus of NER training data on automatic construction method (Takai et al., 2018) in Japanese-to-English experiments, we used 10 million Japanese and English part sentences of JParaCrawl. We extracted sentence pairs including proper nouns by using tagger and POS. Here, we used Sudachi[2] for Japanese morphological analysis to find proper nouns. We chose 1,000 sentence pairs containing proper nouns for the NER training, based on the sentence pair scores (Morishita et al., 2020).

4.2 MT Models

We used Transformer and Factored Transformer models for NMT, with 6-layer encoders and decoders. The configurations of the models and their training were mostly the same as those of the *vanilla Transformer*, but we used different settings on the hyperparameters as shown in the following Table 2.

directions	max token size	max epoch
J-E	7,300	60
E-J	7,300	33

Table 2: Details of NMT hyperparameters

We used SentencePiece (Kudo et al., 2018) with a subword unigram model for the subword tokenization. We used Sudachi and Moses[3] as Japanese and English POS taggers.

4.3 Training and Dev. Data for MT Models

Details of the corpus for the NMT models are shown in Table 3. For the Japanese-to-English experiments, we used a part of 10 million Japanese-to-English sentence pairs in JParaCrawl (Morishita et al., 2020) for the training of the NMT models. We chose 160,000 sentence pairs that contain proper nouns, have sentence pair scores higher than 0.786, and shorter than 250 subwords. For the English-to-Japanese experiments, we used all the 10 million sentence pairs in JParaCrawl as a training data set due to the effectiveness of the different conditions from the Japanese-to-English one: language pairs and amount of training data. WMT 2020 development set[4] was used as the development set for all the NMT models.

	direction	# of sentences	# of subwords	# of uniq subwords
Train	J-E	159,888	5,318,140	10,073
Dev		10,000	333,933	9,941
Train	E-J	10,116,570	332,520,888	47,087
Dev		1,998	65,649	6,873

Table 3: Details of corpus size

4.4 Evaluation Data

Details of the evaluation data are shown in Table 4.

For the Japanese-to-English, we used an evaluation dataset of 271 sentences containing a single proper noun. It was collected through field experiments with taxis in Japan and was translated manually. The data consisted of conversations between taxi drivers and travelers. For the English-to-Japanese task, we used WMT 2020 Test set.

direction	# of sentences	# of subwords	# of uniq subwords
J-E	271	4,258	646
E-J	1,000	32,696	5,171

Table 4: Details of evaluation data size

4.5 Compared Methods

We compared the following NMT models:

- Transformer (baseline)
- Proposed methods with the combination of the model architecture and the NE feature vector representations:
    - 1-encoder / 2-encoders
    - NE one-hot vector / NE probability distribution vector

4.6 Evaluation metrics

We used BLEU (Papineni et al., 2002) as a translation quality metric. We also evaluated proper noun translation accuracy (PRPacc); i.e., the percentage of proper noun words that correctly translated over the entire test set.

[1] http://nlp.stanford.edu/software/stanza/1.2.2/en/ner/ontonotes.pt

[2] https://github.com/WorksApplications/Sudachi

[3] http://www.statmt.org/moses/

[4] http://www.statmt.org/wmt20/translation-task.html

5 Results

Table 5 shows the results. In Japanese-to-English, the proposed 1-encoder models were worse than the baseline, but the 2-encoders models outperformed the baseline. The results by the 2-encoder model with NE probability distributions showed the best performance, outperformed the baseline by 9.6 points in PRPacc and 2.5 points in BLEU. In English-to-Japanese, however, the 1-encoder models outperformed the baseline. The improvement in BLEU and PRPacc was smaller than that in Japanese-to-English. This may be due to the difference in the training data sizes; the English-to-Japanese MT models were trained using the 60 times larger parallel corpus. Another possible reason is the difference in the degrees of difficulty in these domains; WMT News task would be more difficult than the taxi conversation.

With respect to proper noun translation, the lack of a specific treatment of proper noun translation in the baseline resulted in worse performance than the proposed method. Translation examples are shown in Table 5. The 2-encoders models worked well on two types of proper nouns: the non-compositional proper noun of Table 6 (1), and the combination with proper nouns and general noun of Table 6 (2). This result can be assumed that the factor of NE feature vector directly works on the proper noun translation better in the near decoder.

As shown in Table 6, 2-encoders with NE probability distributions have better performance than 2-encoders with NE one-hot vector performance. Expression of the ambiguity of proper nouns in the NE probability distributions method influent on not only proper noun translation but also the surrounding words of a proper noun.

NMT Model	NE Feature	PRPacc(%)		BLEU	
		J-E	E-J	J-A	E-J
vanilla (baseline)	-	56.1	46.5	11.4	17.5
1-encoder	One-hot	43.2	**50.1**	10.1	**18.8**
2-encoders		63.5	47.5	13.8	17.8
1-encoder	Probability distributions	53.5	49.5	10.9	18.4
2-encoders		**65.7**	46.7	**13.8**	17.6

Table 5:Proper noun accuracy and BLEU in J-E task / E-J task

6 Conclusions

We proposed a method to enhance accurate proper noun translation using subword-based NMT by Factored-Transformer and NE features. The NE feature vectors are injected into Factored Transformer model as factors together with subwords. In the Japanese-to-English experiments using a small bilingual training corpus, the proposed method using the best NE feature vector outperformed the baseline sub-word-based transformer model by more than 9.6 points in proper noun accuracy and 2.5 points in the BLEU score. It also showed some improvements in the English-to-Japanese experiments using a large-scale bilingual corpus.

In future work, we will work on automatic clustering of proper nouns instead of given NE categories.

(1) Input sentence: 山の上に**岐阜城**があります (**Gifu Castle** is on top of the mountain.)		
vanilla	-	there are castle on the mountains above the mountains.
1-encoder	One-hot	mount Huangshan is a mountains above the altitude.
2-encoders		there are **Gifu Castle** on the mountains of the mountains.
1-encoder	Probability distributions	In the mountains, **Gifu Castle** is located above the top.
2-encoders		there is **Gifu Castle** on the top of the mountain.
(2) Input sentence: この城は**豊臣秀吉**が作りました (This castle was built by **Toyotomi Hideyoshi**.)		
vanilla	-	this castle was created by an excellent Japanese castle.
1-encoder	One-hot	this castle was created by yoshino hideyoshi hideyoshinori.
2-encoders		this castle of this castle was created by **toyotomi hideyoshi**.
1-encoder	Probability distributions	this castle was created by minister **toyotomi hideyoshi.**
2-encoders		this castle was created by **toyotomi hideyoshi**.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. *Attention is all you need.* In Advances in neural information processing systems, pages 5998–6008.

Chris Hokamp and Qun Liu. 2017. *Lexically constrained decoding for sequence generation using grid beam search.* In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1535–1546.

Guanhua Chen, Yun Chen, Yong Wang and Victor O.K. Li. 2020. *Lexical-constraint-aware neural machine translation via data augmentation.* In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 3587–3593. International Joint Conferences on Artificial Intelligence Organization. Main track.

Hideo Okuma, Hirofumi Yamamoto and Eiichiro Sumita. 2008. *Introducing a translation dictionary into phrase-based smt.* The IEICE Transactions on Information and Systems 91-D pages 2051–2057.

Hirofumi Yamamoto, Hiroaki Kokubo, Genichiro Kikui, Yoshihiko Ogawa and Yoshinori Sagisaka. 2004. *Out-of-vocabulary word recognition with a hierarchical language model using multiple markov model.* In IEICE D-II 87(2), pages. 2104–2111.

Hirofumi Yamamoto and Yoshinori Sagisaka. 1999. *Multi-Class Composite N-gram based on connection.* In ICASSP, pages. 533-536

Huang Zhiheng, Xu Wei and Yu Kai. 2015. *Bidirectional LSTM-CRF models for sequence tagging.* arXiv preprint arXiv:1508.01991.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. *bidirectional transformers for language understanding.* arXiv:2004.08053v2

Jordi Armengol-Estapé, Marta R. Costa-jussà and Carlos Escolano 2020. *Enriching the Transformer with Linguistic Factors for Low-Resource Machine Translation.* arXiv preprint arXiv: arXiv:2004.08053v2.

Katsuki Chousa and Makoto Morishita. 2021. *Augmentation Improves Constrained Beam Search for Neural Machine Translation* The 8th Workshop on Asian Translation (WAT 2021)

Keiji Yasuda, Panikos Heracleous, Akio Ishikawa, Masayuki Hashimoto, Kazunori Matsumoto and Fumiaki Sugaya. 2017. *Building a location dependent dictionary for speech translation systems.* In CICLING.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).

Kohichi Takai, Gen Hattori, Keiji Yasuda, Panikos Heracleous, Akio Ishikawa, Kazunori Matsumoto and Fumiaki Sugaya. 2018. *Automatic Method to Build a Dictionary for Class-based Translation Systems* In CICLING.

Makoto Morishita, Jun Suzuki and Masaaki Nagata. 2020. *JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus.* In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages. 3603-3609

Masatsugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro and Satoshi Sato. 2005. *Translation estimation for technical terms using corpus collected from the web.* In: Proceedings of the Pacific Association for Computational Linguistics, pages. 325–331.

Mercedes García-Martínez, Loïc Barrault and Fethi Bougares. 2016. *Factored neural machine translation.* CoRR, abs/1609.04621.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov and Alexey Sorokin. 2019. *Tuning Multilingual Transformers for Named Entity Recognition on Slavic Languages* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) pages 89-93

Rico Sennrich, Barry Haddow and Alexandra Birch. 2016. *Neural machine translation of rare words with subword units.* In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pages.1715-1725

Taku Kudo and John Richardson. 2018. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.* In EMNLP demo, pages 66-71

.