Extract and Attend: Improving Entity Translation in Neural Machine Translation

Zixin Zeng[1]*, **Rui Wang**[2], **Yichong Leng**[3], **Junliang Guo**[2], **Xu Tan**[2], **Tao Qin**[2], **Tie-yan Liu**[2]
[1] Peking University, [2] Microsoft Research Asia
[3] University of Science and Technology of China
[1]1800016623@pku.edu.cn
[2] {ruiwa,junliangguo,xuta,taoqin,tyliu}@microsoft.com
[3]lyc123go@mail.ustc.edu.cn

Abstract

While Neural Machine Translation (NMT) has achieved great progress in recent years, it still suffers from inaccurate translation of entities (e.g., person/organization name, location), due to the lack of entity training instances. When we humans encounter an unknown entity during translation, we usually first look up in a dictionary and then organize the entity translation together with the translations of other parts to form a smooth target sentence. Inspired by this translation process, we propose an Extract-and-Attend approach to enhance entity translation in NMT, where the translation candidates of source entities are first extracted from a dictionary and then attended to by the NMT model to generate the target sentence. Specifically, the translation candidates are extracted by first detecting the entities in a source sentence and then translating the entities through looking up in a dictionary. Then, the extracted candidates are added as a prefix of the decoder input to be attended to by the decoder when generating the target sentence through self-attention. Experiments conducted on En-Zh and En-Ru demonstrate that the proposed method is effective on improving both the translation accuracy of entities and the overall translation quality, with up to 35% reduction on entity error rate and 0.85 gain on BLEU and 13.8 gain on COMET.

1 Introduction

Neural machine translation (NMT) automatically translates sentences between different languages, which has achieved great success (Bahdanau et al., 2015; Sutskever et al., 2014; He et al., 2016; Song et al., 2019; Wang et al., 2021). Most current works consider to improve the overall translation quality. However, the words in a sentence are not equally important, and the translation accuracy of named entities (e.g., person, organization, location) largely affects user experience, an illustration of which is shown in Table 1. Unfortunately, the translation accuracy of named entities in a sentence is not quite good with current NMT systems (Hassan et al., 2018; Läubli et al., 2020) due to the lack of training instances, and accordingly more effort is needed.

Recalling the process of human translation, when encountering an unknown entity in a sen tence, humans look up the translation of the entity in mental or external dictionaries, and organize the translation of the entity together with the translations of other parts to form a smooth target sentence based on grammar and language sense (Gerver, 1975; Cortese, 1999). As the original intention of neural networks is to mimic the human brain, the human translation process is also an important reference when dealing with entities in NMT. However, none of the previous works on improving the entity translation in NMT consider both steps in human translation: 1) some works annotate the types and positions of the entities without using the dictionary (Li et al., 2018b; Modrzejewski et al., 2020); 2) some works first extract the entity translations from a dictionary (Wang et al., 2017) or an entity translation model (Li et al., 2018a; Yan et al., 2019; Li et al., 2019), and then directly use them to replace the corresponding entities in the translated sentence via post-processing, which only takes the first step of human translation and may affect the fluency of the target sentence; 3) a couple of works use data augmentation or multi-task training to handle the entities in NMT (Zhao et al., 2020a; Hu et al., 2022), which do not explicitly obtain the translation for each entity as the first step in human translation.

Inspired by the human translation process, we propose an Extract-and-Attend approach to improve the translation accuracy of named entities in NMT. Specifically, in the "Extract" step, translation candidates of named entities are extracted by first detecting each named entity in the source sentence and then translating to target language

*This work was completed at Microsoft. Corresponding authors: Rui Wang and Xu Tan ({ruiwa,xuta}@microsoft.com).

Source	北岛的绘画展在巴黎地平线画廊开幕。
Reference	Bei Dao's painting exhibition opens at Horizon Gallery in Paris.
Output 1	North Island's painting exhibition opens at Horizon Gallery in Paris.
Output 2	Bei Dao's picture exhibition opens on Horizon Gallery in Paris.

Table 1: Illustration of entity translation in a sentence, where "北岛" in Chinese can be either a person name or an island. Both outputs 1 and 2 have two different words with red color compared with the reference sentence, while output 2 with correct translation on the entity "北岛" is much better.

based on the dictionary. Considering that some types of entities (e.g. person names) have relatively high diversity and low coverage in dictionaries, we also develop a transliteration[1] pipeline to handle the entities uncovered by the dictionary. In the "Attend" step, the extracted candidates are added to the beginning of the decoder input as a prefix to be attended to by the decoder via self-attention. The Extract-and-Attend approach enjoys the following advantages: 1) the translation candidates of the named entities are explicitly extracted and incorporated during translation, which provides specific references for the decoder to generate the target sentence; 2) the extracted candidates are incorporated via self-attention instead of hard replacement, which considers the context of the whole sentence and leads to smooth outputs. The main contributions of this paper are summarized as follows:

- We propose to mimic the human translation process when dealing with entities in NMT, including extracting the translations of entities based on dictionary and organizing the entity translations together with the translations of other parts to form a smooth translation.

- Accordingly, we propose an Extract-and-Attend approach to improve the quality of entity translation in NMT, which effectively improves the translation quality of the named entities.

- Experiments conducted on En-Zh and En-Ru demonstrate that the proposed Extract-and-Attend approach significantly reduces the error rate on entity translation. Specifically, it reduces the entity error rate by up to 35% while also improving BLEU by up to 0.85 points and COMET up to 13.8 points.

2 Related Work

To improve the entity translation in NMT, some works focus on annotating named entities to provide type and position information. For example, the inline annotation method (Li et al., 2018b) inserts special tokens before and after the entities in the source sentence. The source factor method (Ugawa et al., 2018; Modrzejewski et al., 2020) adds entity type embeddings to the tokens of the entities in the encoder. Xie et al. (2022) attach entity classifiers to the encoder and decoder. One main challenge when dealing with entities is that the entities are quite diverse while the corresponding data is limited compared to the large number of entities. Dictionaries are important supplements to the limited data on entities, which are not utilized in these works.

With the help of bilingual dictionaries, one common approach to improve the entity translation in NMT is to first extract the translation of source entities based on a dictionary (Wang et al., 2017) or an entity translation model (Li et al., 2018a; Yan et al., 2019; Li et al., 2019), and then locate and replace the corresponding tokens in the target sentence via post-processing. However, such approach only takes the first step of human translation (i.e., extracting the entity translations), since the entity translations are inserted to the target sentence by hard replacement, which affects the fluency of the target sentence. Moreover, this approach is sensitive to the inaccurate predictions made by NER (Modrzejewski et al., 2020).

Recently, some works take advantage of additional resources (e.g., dictionary) via data augmentation or multi-task training to improve the translation quality on entities. Zhao et al. (2020b) augment the parallel corpus based on paired entities extracted from multilingual knowledge graphs, while DEEP (Hu et al., 2022) augments monolingual data with paired entities for a denoising pretraining task. The entity translation can also be enhanced by multi-task training with knowledge reasoning (Zhao et al., 2020a) and integrating lexical constraints (Wang et al., 2022). These methods don't look up translation candidates in bilingual

[1] Transliteration is to convert between languages while keeping the same pronunciation (Karimi et al., 2011).

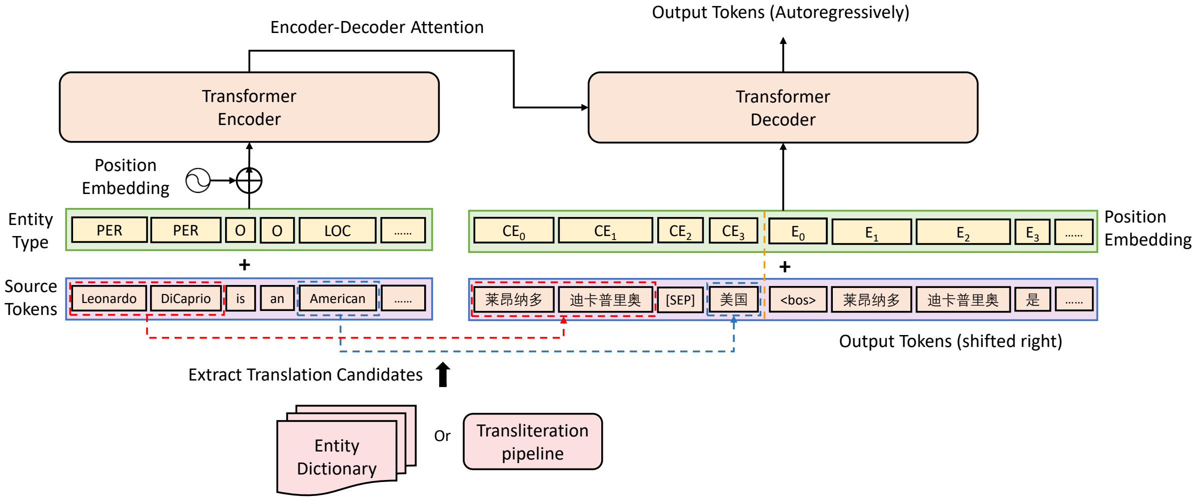

Figure 1: Extract-and-Attend approach, where the translation candidates are extracted and added as a prefix of the decoder input. Entity type embeddings are added to the source input (e.g., 'PER' for person names, 'LOC' for locations and 'O' for other tokens other than entities). Independent position embeddings are used for the translation candidates and the shifted output tokens (i.e., 'CE' for translation candidates and 'E' for output tokens).

dictionares during inference. Considering that entities are quite diverse, providing specific translation candidates from dictionary may further improve the quality of entity translation.

Bilingual dictionaries are also utilized for improving translation quality on rare words or domain-specific terminology. One common approach is to augment training data with pseudo parallel sentences generated based on the dictionary (Zhang and Zong, 2016; Nag et al., 2020; Zhao et al., 2020b; Peng et al., 2020). Some works adjust the output probabilities over the vocabulary in the decoder according to the dictionary (Arthur et al., 2016; Zhao et al., 2018; Zhang et al., 2021). Zhong and Chiang (2020) attach the definitions of the rare words in the dictionary to enhance the rare word translation. Similarly, Dinu et al. (2019) and Exel et al. (2020) proposed to inject terminology by replacing or inserting translations inline in the source sentence. Though the human translation process when encountering an unknown rare word/terminology or entity is the same, we argue that the two-step human translation process is more suitable for entities. This is because rare words can be polysemous and require context-based disambiguation; on the other hand, each entity is usually linked with a single sense after controlling for entity type. Accordingly, retrieved translations of entities are less ambiguous than other words. On the contrary, domain-specific terminology always has a single sense which has little relevant to context, and thus it is usually with much higher accuracy to identify the terminologies in the domain-specific sentences than entities. Another uniqueness of entities is that some entities are translated by the same rule, which makes it possible to generalize to unseen entities. For example, when translating the names of Chinese people from Chinese to English, Pinyin[2] is commonly used.

3 Improving Entity Translation in NMT

Inspired by the translation process of humans when encountering an unknown entity, where the translation of the entity is extracted from a dictionary and then organized with the translations of other parts to form a fluent target sentence, we propose an Extract-and-Attend approach. Specifically, we first extract the translation candidates of the entities in the source sentence, and then attend the translation candidates into the decoding process via self-attention, which helps the decoder to generate a smooth target sentence based on the specific entity translations. An overview of the proposed Extract-and-Attend approach is shown in Fig. 1, where a Transformer-based (Vaswani et al., 2017) encoder-decoder structure is adopted. Specifically, to extract the translation candidates, entities in the source sentence are first detected based on NER (Li et al., 2020), then the translation candidates are obtained from a bilingual dictionary. Considering that some types of named entities (e.g., person

[2] https://en.wikipedia.org/wiki/Pinyin

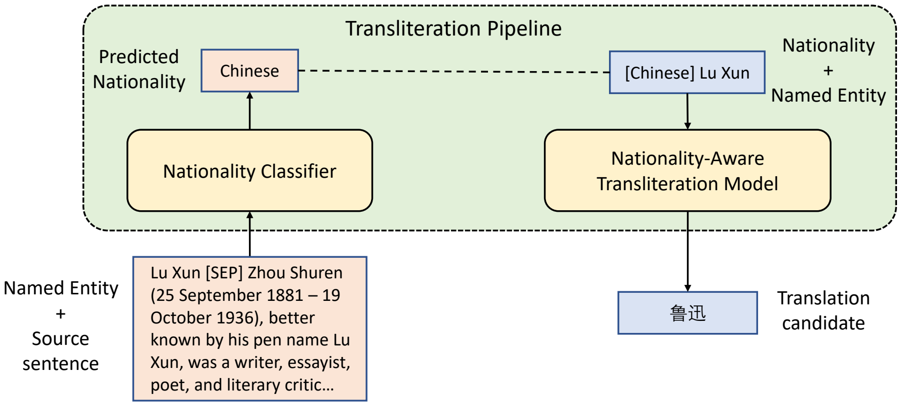
Transliteration Pipeline
Predicted Nationality
Chinese
[Chinese] Lu Xun
Nationality + Named Entity
Nationality Classifier
Nationality-Aware Transliteration Model
Named Entity + Source sentence
Lu Xun [SEP] Zhou Shuren (25 September 1881 – 19 October 1936), better known by his pen name Lu Xun, was a writer, essayist, poet, and literary critic...
鲁迅
Translation candidate

Figure 2: Transliteration pipeline.

names) are quite diverse and the coverage in the dictionary of such entities is limited, we also develop a transliteration pipeline to handle entities uncovered by the dictionary. To make the decoder attend to the translation candidates, we add the translation candidates in order as a prefix of the decoder input. In the following sections, we will provide the details of “Extract” and “Attend”.

3.1 Extracting Translation Candidates

Extracting the translation candidates for entities in the source sentence provides explicit references when generating the target sentence in NMT. There are two steps when extracting the entity translation candidates, where the entities in the source sentences are first detected by NER and then translated to the target language. If the entity is found in the bilingual dictionary, we retrieve its translation(s). Although there may be multiple translation candidates for one entity, the entity usually links to a single sense after disambiguating by entity type, and the multiple candidates in the dictionary for one named entity are commonly all correct. For example, “John Wilson” can be translated to “约翰·维尔逊” or “约翰·威尔森”. During training, we consider the one with shortest Levenshtein distance[3] compared to the ground truth translation to encourage the decoder to copy the given candidate. During inference, considering that only the source sentence is available, we select the one with highest frequency in the training set.

The coverage in the dictionary is limited for some types of entities (e.g, person names). Meanwhile, a large number of named entities (e.g., person names and some of locations) are translated by transliteration (i.e., translated according to the pronunciations). Accordingly, we consider to use transliteration to handle such entities if they are uncovered by the dictionary. Transliteration in different countries often follow different rules. For example, names of Chinese persons are transliterated into English via Pinyin, while names of Korean persons are often transliterated via McCune-Reischauer[4]. Current transliteration models (Kundu et al., 2018; Karimi et al., 2011; Le et al., 2019) do not consider different nationalities for a single language pair, which is an important cause for transliteration errors. Considering this, we develop a nationality-aware transliteration pipeline, which consists of a nationality classifier and a nationality-aware transliteration model. As shown in Fig. 2, the nationality classifier takes the source entity and source sentence as input, and predicts the nationality of the entity. Then, the nationality tag is concatenated with the entity and translated by the word-level transliteration model.

3.2 Attending to Translation Candidates

We consider to let the decoder attend to the extracted translation candidates via self-attention, which has shown to be more effective in improving entity translation compared to alternative designs (see Section 5.3). Accordingly, we concatenate extracted candidate translations with “[SEP]” and place it before the “<bos>” token of the decoder input. In order to identify the alignments between the translation candidates and the corresponding entities in the source sentence, we add entity type embeddings to word embeddings of the entities in the source sentence as (Modrzejewski et al., 2020), and concatenate the corresponding translation candidates in the same order as they are in the source

[3] https://en.wikipedia.org/wiki/Levenshtein_distance

[4] https://en.wikipedia.org/wiki/McCune-Reischauer

sentence. We demonstrate that our model can correctly align the entities and the corresponding translation candidates in Appendix A.1 via case study. We use independent position embeddings for the translation candidates and the target sentence as shown in Fig. 1. The loss on the tokens of translation candidates is ignored. In this way, the decoder can attend to the translation candidates through the attention mechanism in the decoder, which helps improve the performance of the model on translating entities.

4 Experimental Settings

In this section, we describe experimental settings including datasets, model configurations, the evaluation criterion and baselines.

4.1 Datasets

We conduct experiments on English-Chinese (En-Zh) and English-Russian (En-Ru) translation. We chose language pairs so that the source and target languages come from different scripts[5], because cross-script entity translation is more challenging. Following Modrzejewski et al. (2020), three types of named entities are considered, i.e., person name, organization and location. Note that the proposed framework is not limited to the three types and can be applied to other entities (e.g., domain entities).

Entity dictionary. Entity pairs and corresponding nationality information are obtained from two multilingual knowledge graphs (i.e., DBPedia and Wikidata). For En-Ru, we extract 401K, 175K and 50K pairs of PER, LOC and ORG entities respectively. For En-Zh, we extract 338K, 200K, 38K pairs of PER, LOC and ORG entities respectively. Besides, we increase the coverage of the entity dictionary by mining entity pairs from parallel data. First, we use spaCy NER models[6] to recognize entities from parallel sentences, then use awesome-align (Dou and Neubig, 2021) to align the source and target tokens and extract the corresponding translations. Infrequent entity pairs or empty alignment results are filtered out. Specifically, we obtain 179K person names, 51K locations, and 63K organizations for En-Ru, and 152K person names, 32K locations, and 39K organizations for En-Zh.

Dataset for transliteration pipeline. Most person names and part of locations can be translated by transliteration. Because the dictionary has relatively high coverage for location entities, we train the transliteration pipeline based on parallel person names, and use it for both person names and unseen locations. To train the nationality classifier, we extract English biographies from DBPedia and link them to the entity dictionary, which are translated into Chinese and Russian with custom NMT models. In total, we collect 54K sentences with person names and nationalities, where 48.2K, 1.5K and 3.9K of them are used as training set, validation set and test set, respectively. We also merge countries that share the same official language (e.g. USA and UK), and regard the nationalities with fewer than 1000 examples as "Other". For the nationality-aware transliteration model, the paired person names with nationality information from the collected entity dictionary are used. For En-Zh, 316K, 5K, and 17K are used as training set, validation set and test set respectively, and for En-Ru, 362K, 13K, 26K are used as training set, validation set and test set respectively. Besides, we also collect common monolingual person names from various databases[7], and create pseudo entity pairs via back translation (Sennrich et al., 2016). In total, 10K, 1.6M and 560K entities are collected for English, Chinese and Russian respectively.

Dataset for NMT model. The training data is obtained from UN Parallel Corpus v1.0 and News Commentary Corpus v15[8]. The test data is constructed by concatenating test sets of the WMT News Translation Task (2015-2021) and deduplicating samples. Dataset statistics are shown in Table 2. For En-Zh, there are 6.6K PER entities, 4.4K ORG entities and 1.9K LOC entities. For En-Ru, there are 4.9K PER entities, 2.5K ORG entities and 1.2K LOC entities. We use Moses[9] to tokenize English and Russian corpus, and perform word segmentation on Chinese corpus with jieba[10]. We perform joint byte-pair encoding (BPE) by subwordnmt[11] with a maximum of 20K BPE tokens.

[5] English uses Latin script, Chinese uses Logographic script, and Russian uses Cyrillic script.

[6] https://pypi.org/project/spacy/

[7] https://namecensus.com/
http://www.openkg.cn/dataset/cndbpedia
https://github.com/wainshine/Chinese-Names-Corpus
https://github.com/datacoon/russiannames

[8] Available at https://www.statmt.org/wmt20/translation-task.html

[9] https://github.com/moses-smt/mosesdecoder

[10] https://pypi.org/project/jieba/

[11] https://github.com/rsennrich/subword-nmt

Languages	#Sentences	#Tokens
En-Ru	23.5M/25K/13K	726M/456K/227K
En-Zh	16.2M/20k/19k	487M/512K/503K

Table 2: Statistics of NMT datasets (Train/Val/Test).

4.2 Model Configurations and Training Pipeline

The nationality classifier is fine-tuned from pretrained BERT checkpoint (base, cased) available on HuggingFace[12]. Both the NMT model and the nationality-aware transliteration model use Transformer base architecture (Vaswani et al., 2017) with 6-layer encoder and decoder, hidden size as 512 and 8 attention heads.

4.3 Evaluation Criterion and Baselines

To evaluate the overall translation quality, we compute BLEU and COMET (Rei et al., 2020) scores[13]. To evaluate the translation quality on entities, we consider using error rate of entity translation as the evaluation criterion. Following Modrzejewski et al. (2020), we evaluate entity error rate by recognizing named entities from the reference sentence, and then checking occurrence in the output sentence, where it is regarded as error if it does not occur.

We compare our Extract-and-Attend approach with the following baselines[14]:

- *Transformer.* The Transformer model is directly trained on parallel corpus.

- *Transformer with Dictionary.* The entity dictionary is directly added to the parallel corpus to train a transformer model.

- *Replacement.* After identifying entities in the source sentence with NER and aligning them with target tokens, the corresponding tokens are replaced by translation candidates.

- *Placeholder (Yan et al., 2019; Li et al., 2019).* It first replaces the entities in the source sentence with placeholders based on NER and then restores the placeholders in the output sentence with the extracted translation candidates.

- *Annotation (Modrzejewski et al., 2020).* Entity type embeddings are added to the original word embeddings for the tokens of entities in the source sentence.

- *Multi-task (Zhao et al., 2020a)* It improves the entity translation in NMT by multi-task learning on machine translation and knowledge reasoning.

5 Experimental Results

In this section, we demonstrate the effectiveness of the proposed Extract-and-Attend approach by comparing it with multiple baselines. We also conduct experiments to verify the design aspects of "Extract" and "Attend".

5.1 Main Results

BLEU, COMET and entity error rates of the Extract-and-Attend approach with the baselines are shown in Table 3 and Table 4, where the proposed approach consistently performs the best on all the metrics and language pairs. From the results, it can be observed that: 1) The proposed method reduces the error rate by up to 35% and achieves a gain of up to 0.85 BLEU and 13.8 COMET compared to the standard Transformer model; 2) Compared with the annotation method (Modrzejewski et al., 2020), which annotates the entities in the source sentence based on NER without incorporating any additional resources (e.g., dictionary), the proposed Extract-and-Attend approach takes advantage of the entity dictionary and nationality-aware transliteration pipeline, and reduces the entity error rate by up to 26% while achieving up to 0.77 points gain on BLEU and 3.0 points on COMET; 3) Compared with the replacement and placeholder (Yan et al., 2019; Li et al., 2019) methods, the Extract-and-Attend approach is more robust to NER errors (see A.3) than hard replacement and reduces the error rate by up to 16% while gaining up to 2.1 BLEU and 7.2 COMET; 4) Compared to the multi-task (Zhao et al., 2020a) method, the Extract-and-Attend approach explicitly provides the translation candidates when decoding, which reduces the entity error rate by up to 35% and improves BLEU by up to 0.8 points and COMET up to 4.4 points. We also provide the error rates for different entity types in Appendix A.2, and analyze the effect of dictionary coverage in Appendix A.4

Entity error rates calculated according to Section 4.3 may incur false negative errors, which has

[12] https://huggingface.co/bert-base-uncased

[13] The wmt22-comet-da model is used to calculate COMET scores

[14] The entity resources used in Transformer with Dictionary, Replacement and Placeholder are obtained as is Section 3.1

Model	$En \rightarrow Ru$		$Ru \rightarrow En$		$En \rightarrow Zh$		$Zh \rightarrow En$	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Transformer	31.83	52.2	34.63	54.0	26.32	34.8	27.45	41.5
Transformer w/ Dictionary	31.85	53.6	34.67	56.1	26.36	38.1	27.49	43.2
Replacement	30.52	55.2	32.01	56.7	25.92	41.4	27.21	45.0
Placeholder	31.88	57.6	34.72	59.1	26.41	42.9	27.50	47.2
Annotation	31.91	59.4	34.84	60.5	26.44	45.8	27.73	48.0
Multi-task	31.88	57.8	34.76	60.3	26.38	45.0	27.64	47.4
Extract & Attend (ours)	**32.68**	**62.2**	**35.41**	**63.5**	**26.79**	**48.6**	**27.98**	**50.1**

Table 3: BLEU and COMET scores on WMT newstest. BLEU and COMET scores are statistically higher than baselines across all language pairs with 95% statistical significance (Koehn, 2004).

Model	$En \rightarrow Ru$	$Ru \rightarrow En$	$En \rightarrow Zh$	$Zh \rightarrow En$
Transformer	60.0	51.3	42.7	41.0
Transformer w/ Dictionary	59.2	50.4	42.1	40.6
Replacement	49.6	49.8	29.5	28.9
Placeholder	49.7	49.3	28.6	27.9
Annotation	43.2	44.5	37.4	30.0
Multi-task	58.9	50.0	42.4	40.4
Extract & Attend (ours)	**42.7**	**41.6**	**27.7**	**27.5**

Table 4: Error rates (%) on WMT newstest.

two main causes. First, as noted by Modrzejewski et al. (2020), it is common for NER models to make erroneous predictions. Second, there may be multiple correct translations for one entity, but the ones different from that in the reference sentence are regarded as errors. For example, BMA (British Medical Association) can either be copied in the target sentence, or translated into its Chinese form "英国医学会". Therefore, we also perform human evaluation wmttest150 (see Table 5), where 150 sentence pairs with entities are randomly sampled from the $En \rightarrow Zh$ test set. Compared to automatic evaluation results in Table 4, entity error rates based on human evaluation become lower after eliminating the false negatives, while the relative performance of different models remain almost consistent. Therefore, though there are false negatives in the automatic evaluation as in Section 4.3, it is still a valid metric for evaluating entity translation. Moreover, we observe that the Extract-and-Attend approach performs the best on all three entity types and reduces the total error rate by 32%.

Model	PER	ORG	LOC	Total
Transformer	26.4	14.3	13.4	17.9
Transformer w/ Dictionary	25.8	13.6	13.4	16.7
Replacement	19.8	12.4	12.6	14.8
Placeholder	18.9	12.4	10.9	13.6
Annotation	17.9	11.4	10.9	13.3
Multi-task	21.7	12.4	12.6	15.5
Extract & Attend (ours)	**16.0**	**11.4**	**9.2**	**12.1**

Table 5: Human evaluation of entity error rates (%) on wmttest150 for $En \rightarrow Zh$.

5.2 Analysis on Extracting

To investigate the effectiveness of our transliteration pipeline, we implement a variant denoted as Extract-and-Attend (w/o Transliteration), in which we only extract translation candidates covered by the dictionary. From Table 6, we can see that the translation quality of person names is significantly improved, reducing the error rate by 37%; transliteration is also effective for locations, reducing the error rate by 9%. Overall, the transliteration model improves BLEU by 0.33 and COMET by 4.1.

Model	BLEU	COMET	PER	LOC
Extract & Attend (with Transliteration)	26.79	48.6	25.6	31.6
Extract & Attend (w/o Transliteration)	26.46	46.5	40.8	34.8

Table 6: BLEU, COMET and error rates (%) for $En \rightarrow Zh$.

Considering that different transliteration rules may be applied for different countries, we propose to incorporate nationality information during transliteration. To evaluate the effectiveness of utilizing the nationality information in the transliteration pipeline, we compare the performance of the proposed nationality-aware transliteration pipeline with the transliteration model trained on paired entities without nationality information. As shown in Table 7, adding nationality information during transliteration consistently improves transliteration quality across all language pairs, and is most helpful for $Zh \rightarrow En$, where the transliteration accuracy is improved by 9%.

Transliteration	$En \rightarrow Ru$	$Ru \rightarrow En$	$En \rightarrow Zh$	$Zh \rightarrow En$
Nationality-aware	79	85	95	97
w/o Nationality	74	82	90	88

Table 7: Accuracy of transliteration (%).

5.3 Analysis on Attending

We also conduct experiments to evaluate the effect of attending translation candidates in the encoder compared to the decoder. Similar to Zhong and Chiang (2020), we append translation candidates to the source tokens, where the position embeddings of the translation candidates are shared with the first token of the corresponding entities in the source sentence. Relative position embeddings denoting token order within the translation candidate are also added. As shown in Table 8, adding the translation candidates to the decoder is better than adding to the encoder. Intuitively, attending to translation candidates in the encoder may incur additional burden to the encoder to handle multiple languages.

Model	BLEU	COMET	Error rate
Extract & Attend (Decoder)	26.79	48.6	27.7
Extract & Attend (Encoder)	26.56	46.2	29.8

Table 8: BLEU, COMET and error rates (%) for $En \rightarrow Zh$.

Some entities have multiple translation candidates in the entity dictionary. To study whether to provide multiple candidates for each named entity, we extract up to three candidates from the entity dictionary. To help the model distinguish different candidates, we use a separator between candidates of the same entity, which is different from the one used to separate the candidates for different entities. Table 9 shows that adding multiple translation candidates slightly reduces the translation quality in terms of BLEU, COMET and entity error rate. Intuitively, all the retrieved translation candidates for an entity are typically correct, and using one translation candidate for each entity provides sufficient information.

Model	BLEU	COMET	Error rate
Extract & Attend (single candidate)	26.79	48.6	27.7
Extract & Attend (multiple candidates)	26.75	47.9	27.8

Table 9: BLEU, COMET and error rates (%) for $En \rightarrow Zh$.

5.4 Inference Time

Extracting translation candidates requires additional inference time, including the delays from NER and transliteration pipeline. Specifically, the average inference time for standard Transformer, Replacement, Placeholder, Annotation, Multi-task and our method are 389ms, 552ms, 470ms, 416ms, 395ms, 624ms[15].

6 Conclusion

In this paper, we propose an Extract-and-Attend approach to improve the translation quality in NMT systems. Specifically, translation candidates for entities in the source sentence are first extracted, and then attended to by the decoder via self-attention. Experimental results demonstrate the effectiveness of the proposed approach and design aspects. Knowledge is an important resource to enhance the entity translation in NMT, while we only take advantage of the paired entities, nationality and biography information. In the future work, it is interesting to investigate how to make better use of the knowledge, which can be obtained from knowledge graphs and large-scale pre-trained models. Besides, the proposed Extract-and-Attend approach also has some limitations. First, our method requires additional entity resources, which may be difficult to obtain for certain language pairs. With the development of multilingual entity datasets like Paranames (Sälevä and Lignos, 2022), we are optimistic such resources will be more accessible in the near future. Second, as demonstrated in Section 5.4, extracting translation candidates increases inference time. Due to space limitation, more limitations are discussed in Appendix A.6.

[15] Evaluated on a P40 GPU with batch size of 1, other experimental settings same as Section 4

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Giuseppina Cortese. 1999. Cognitive processes in translation and interpreting. joseph h. danks, gregory m. shreve, stephen b. fountain, and michael k. mcbeath (eds.). london: Sage, 1997. pp. 294. *Applied Psycholinguistics*, 20(2):318–327.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

David Gerver. 1975. A psychological approach to simultaneous interpretation. *Meta: Translators' Journal*, 20:119–128.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. DEEP: DEnoising entity pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1753–1766, Dublin, Ireland. Association for Computational Linguistics.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Computing Surveys (CSUR)*, 43(3):1–46.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Soumyadeep Kundu, Sayantan Paul, and Santanu Pal. 2018. A deep learning based approach to transliteration. In *Proceedings of the seventh named entities workshop*, pages 79–83.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human-machine parity in language translation. *J. Artif. Intell. Res.*, 67:653–672.

Ngoc Tan Le, Fatiha Sadat, Lucie Menard, and Dien Dinh. 2019. Low-resource machine transliteration using recurrent neural networks. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–14.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Xiaoqing Li, Jinghui Yan, Jiajun Zhang, and Chengqing Zong. 2018a. Neural name translation improves neural machine translation. In *China Workshop on Machine Translation*, pages 93–100. Springer.

Xiaoqing Li, Jinghui Yan, Jiajun Zhang, and Chengqing Zong. 2019. Neural name translation improves neural machine translation. In *Machine Translation*, pages 93–100, Singapore. Springer Singapore.

Zhongwei Li, Xuancong Wang, Ai Ti Aw, Eng Siong Chng, and Haizhou Li. 2018b. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46, Melbourne, Australia. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *International Conference on Language Resources and Evaluation*.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating external annotation to improve named entity translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51, Lisboa, Portugal. European Association for Machine Translation.

Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Jonne Sälevä and Constantine Lignos. 2022. Paranames: A massively multilingual entity name corpus. *arXiv preprint arXiv:2202.14035*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. *arXiv preprint arXiv:2107.04239*.

Shuo Wang, Zhixing Tan, and Yang Liu. 2022. Integrating vectorized lexical constraints for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7063–7073, Dublin, Ireland. Association for Computational Linguistics.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for wmt17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Mach. Learn.*, 111(3):1181–1203.

Jinghui Yan, Jiajun Zhang, JinAn Xu, and Chengqing Zong. 2019. The impact of named entity translation for neural machine translation. In *Machine Translation*, pages 63–73, Singapore. Springer Singapore.

Jiajun Zhang and Chengqing Zong. 2016. Bridging neural machine translation and bilingual dictionaries.

Tong Zhang, Long Zhang, Wei Ye, Bo Li, Jinan Sun, Xiaoyu Zhu, Wen Zhao, and Shikun Zhang. 2021. Point, disambiguate and copy: Incorporating bilingual dictionaries for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3970–3979, Online. Association for Computational Linguistics.

Yang Zhao, Yining Wang, Jiajun Zhang, and Chengqing Zong. 2018. Phrase table as recommendation memory for neural machine translation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4609–4615. International Joint Conferences on Artificial Intelligence Organization.

Yang Zhao, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020a. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020b. Knowledge graphs enhanced neural machine translation. In *IJCAI*, pages 4039–4045.

Xing Jie Zhong and David Chiang. 2020. Look it up: Bilingual and monolingual dictionaries improve neural machine translation. *arXiv preprint arXiv:2010.05997v2*.

A Appendix

A.1 Case Study

We also conduct a case study on the $En \rightarrow Zh$ test set to demonstrate the capability of our model when handing multiple entities in a sentence. As shown in Table 10, the outputs of our model normally has correct alignments between the translations and the corresponding entities in the source sentence. Besides, the baseline model has a strong tendency to copy unfamiliar entities in the source sentence, while our model can alleviate this problem and encourage the translation model to incorporate proper transliteration.

Source	Simone , Gabby and Laurie all took the same path as Aly and Madison to make the Olympic team .
Reference	西蒙、加布丽埃勒和劳瑞进入奥运代表队的途径跟阿里及麦迪逊一样。
Baseline	Simone、Gabby和劳瑞进入奥运代表队的途径跟Aly及麦迪逊一样。
Ours	西蒙、加比和劳瑞进入奥运代表队的途径跟阿里及麦迪逊一样。
Source	Lomachenko defends his belt against Miguel Marriaga on Saturday night at 7 on ESPN .
Reference	在周六晚上7点的ESPN比赛中,洛马琴科战胜了米格尔·马里亚加，保全了他的地位。
Baseline	Lomachenko 在周六晚上7点在ESPN上为Miguel Marriaga辩护。
Ours	洛马琴科周六晚7点在ESPN对阵米格尔-玛利亚加的比赛中卫冕他的腰带。
Source	iCloud ’ s main data center at Gui-An New Area will be the first data center Apple has set up in China . On completion , it will be used to store the data of Apple users in China .
Reference	iCloud贵安新区主数据中心也将是苹果公司在中国设立的第一个数据中心项目，项目落成后，将用于存储中国苹果 用户的数据。
Baseline	iCloud在桂安新区的主要数据中心将是苹果在中国建立的第一个数据中心。完成后，它将用于存储中国苹果用户的数据。
Ours	iCloud在贵安新区的主要数据中心将是苹果在中国建立的第一个数据中心。完成后，它将用于存储中国苹果用户的数据。

Table 10: Examples of $En \rightarrow Zh$ entity translation. Entities are underlined.

A.2 Error Rates by Entity Type

To alleviate the problem of false errors caused by NER, We aggregate across all language pairs and calculated the average error rate for each type of entity. From Table 11, it is shown that our method outperforms all baselines for PER, ORG and LOC entities.

Model	PER	ORG	LOC
Transformer	50.4	42.4	37.5
Transformer w/ Dictionary	49.8	41.7	37.2
Replacement	35.2	38.9	35.2
Placeholder	34.5	39.0	33.9
Annotation	35.7	40.2	34.1
Multi-task	49.2	41.4	37.6
ours	29.9	38.1	33.4

Table 11: Error rates (%) on WMT newstest by entity type.

A.3 Robustness against NER errors

To test the robustness against NER errors, we filter the samples in which incorrect candidates are collected, which can result from NER errors and transliteration errors. Compared to the Transformer baseline, in 32% of the cases, the extract and attend method is misguided by the incorrect candidates, while for the replacement and placeholder approaches 100% of the cases is misguided. Accordingly, our method is arguably more robust against NER errors.

A.4 Analysis of Dictionary Coverage

To analyze the performance of our approach on domains not well covered by the dictionary, we evaluate our approach and baselines on OpenSubtitles dataset (Lison and Tiedemann, 2016). Because there is no official test set for this dataset, we randomly sample 10K En-Zh sentence pairs. There are 3.6K PER entities, 1.1K ORG entities and 1.1K LOC entities in this test set. Compared to the dictionary coverage of 32.4% for WMT newstest, the dictionary coverage is only 15.2% for the Opensubtitles test set. The overall entity error rates are shown in Table 12. Our results show that even when the coverage of the entity dictionary is relatively low, the proposed Extract-and-Attend framework achieves consistent improvement in entity error rates compared to alternative methods.

Model	Error Rate(%)
Transformer	29.6
Transformer w/ Dictionary	29.2
Replacement	26.8
Placeholder	26.3
Annotation	27.9
Multi-task	28.2
ours	**24.9**

Table 12: Entity error rates (%) on OpenSubtitles test set for $En \rightarrow Zh$.

A.5 Comparison with VecConstNMT

Some researchers have proposed VecConstNMT to mine and integrate lexical constraints from parallel corpora, which can potentially improve entity translation quality (Wang et al., 2022). We compare our method with VecConstNMT on $En \rightarrow Zh$ and $Zh \rightarrow En$. For $En \rightarrow Zh$, and the results are shown in Table 13. Possible reasons that our method outperforms their method include: (1) our method uses additional resources such as dictionaries (2) a relatively small portion of lexical constraints are related to entity translation.

Model	$En \rightarrow Zh$	$Zh \rightarrow En$
VecConstNMT	31.8	28.1
ours	27.7	27.5

Table 13: Error rates (%) on WMT newstest.

A.6 Extended discussion of limitations

Though errors caused by NER are alleviated by attending to the translation candidates via self-attention, the quality of the extracted translation candidates is still affected by NER accuracy and dictionary coverage, and higher quality of translation candidates normally leads to better performance. Another issue worth noting is the evaluation criterion for entity translation. As mentioned in Section 5.1, automatically calculating the error rate on entities based on NER and the reference sentence incurs false negative errors, and better criteria to evaluate the translation quality of entities are needed. What's more, in this paper we assume transliteration rules are the same for regions using the same language and assume that nationality is the same as language of origin, which may be inappropriate in some rare cases. Last but not least, considering that languages may have their own uniqueness, experiments on other language pairs are still needed.

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.