Addressing Entity Translation Problem via Translation Difficulty and Context Diversity

Tian Liang[1*] **Xing Wang**[2†] **Mingming Yang**[2] **Yujiu Yang**[1†] **Shuming Shi**[2] **Zhaopeng Tu**[2]
[1]Shenzhen International Graduate School, Tsinghua University [2]Tencent AI Lab
{liangt21@mails,yang.yujiu@sz}.tsinghua.edu.cn
{brightxwang,shanemmyang,shumingshi,zptu}@tencent.com

Abstract

Neural machine translation (NMT) systems often produce inadequate translations for named entities. In this study, we conducted preliminary experiments to examine the factors affecting the translation accuracy of named entities, specifically focusing on their translation difficulty and context diversity. Based on our observations, we propose a novel data augmentation strategy to enhance the accuracy of named entity translation. The main concept behind our approach is to increase both the context diversity and translation probability for the targeted named entity pair. To achieve this, we construct additional samples for named entities that exhibit high translation difficulty or low context diversity and use the augmented training data to re-train the final translation model. Furthermore, we propose an entity-aware machine translation metric that prefers the translation output to generate more accurate named entities. Our experimental results demonstrate significant improvements over the baseline in terms of general translation performance and named entity translation accuracy across various test sets, such as WMT news translation and terminology test sets.

1 Introduction

Neural machine translation (NMT) that leverages the sequence-to-sequence learning paradigm to transform a source sentence into a target sentence has made significant advancements in recent years. However, previous studies have demonstrated that the translation of low-frequency named entities (e.g., person name, location, organization), which plays a central role in improving the customer experience for commercial translation systems, continues to pose a significant challenge for NMT systems (Koehn and Knowles, 2017; Yan et al., 2018).

Numerous techniques have been proposed to improve the accuracy of named entity translation. These techniques include 1) substituting entity translation in a post-processing step (Luong et al., 2015), 2) utilizing neural translation memory (Wang et al., 2017; Gu et al., 2018; He et al., 2021), 3) employing named entity tag embedding (Ugawa et al., 2018a; Moussallem et al., 2019), 4) applying lexically constrained decoding (Hokamp and Liu, 2017a; Post and Vilar, 2018a; Wang et al., 2022a), and 5) utilizing an interactive translation mechanism (Weng et al., 2019; Xiao et al., 2022). Most of these techniques explicitly decouple the named entity translation from the process of translating the entire sentence.

Along this line, we decompose the sentence into the named entity and its context. We performed preliminary experiments on large-scale in-house data, which allowed us to identify two factors that impact the translation accuracy of named entities:

- *translation difficulty*: entities comprising high-frequency words are more easily translated than those composed of low-frequency words, since NMT models tend to struggle with translating low-frequency words (Koehn and Knowles, 2017).
- *context diversity*: diverse and rich context information can prevent overfitting and boost the generalization capability of NMT models for named entities.

Based on the above observations, we propose a novel data augmentation strategy to improve the translation accuracy of named entities. Specifically, given bilingual entity pairs, we first compute the translation difficulty and context diversity scores based on the training corpus. Next, we select the bilingual context from the bilingual sentence pair and combine it with the bilingual entity pair to create synthetic parallel samples. The basic idea is to

[*]Work done during internship at Tencent AI Lab.
[†]Xing and Yujiu are co-corresponding authors.

Findings of the Association for Computational Linguistics ACL 2024, pages 11628–11638

generate additional synthetic samples for the entity pair that exhibits high translation difficulty or low context diversity. Consequently, these synthetic parallel samples are used to boost the context diversity and the translation probability associated with the named entity pair. Finally, we use the augmented training data to re-train the final translation model.

We also propose a lexical-based entity-aware machine translation metric to evaluate the quality of translations containing named entities more effectively. The proposed metric assigns a high score for the translation outputs with greater accuracy in named entity translation. We adopt the widely-used BLEU (Post, 2018) and ChrF++ (Popović, 2017) to our entity-aware machine translation metric and find that the proposed entity-aware entity metric correlates better with human judgment than the original metrics.

To evaluate the effectiveness of our data strategy, we conducted extensive machine translation experiments on WMT20 (Barrault et al., 2020) news translation and WMT21 (Alam et al., 2021) terminology translation with general decoding (without lexical constraint). Experimental results demonstrate that the proposed approach outperforms several strong baselines in terms of BLEU and named entity translation accuracy. Furthermore, our strategy can be applied to lexically constrained decoding scenarios. We implement the state-of-the-art template-based approach (Wang et al., 2022a) on top of our strategy and find that our strategy can improve the context translation performance while maintaining the entity translation accuracy.

Our main contributions are:

- We conduct experiments on an in-house test set to identify two key factors influencing the accuracy of named entity translation: translation difficulty and context diversity.
- We propose a novel data augmentation strategy to improve the accuracy of named entity translation. Our approach involves generating synthetic samples for entity pairs with high translation difficulty or low context diversity.
- The strength of our data augmentation method lies in providing theoretically guided augmentation strategies and maintaining a balance between computational cost and performance improvement.
- We propose an entity-aware machine translation metric that prefers the translation output to generate more accurate named entities.

2 Preliminary Experiment

To understand the main characteristics of entity translation in NMT, we performed preliminary experiments on large-scale in-house data [1] to empirically demonstrate that the accuracy of entity translation is strongly correlated to two of its attributes, namely, translation difficulty and contextual diversity.

Notations. Let $\mathcal{C} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^{N}$ denotes the authentic parallel data, where $\mathbf{x}^i$ and $\mathbf{y}^i$ are source and target sentences, and N denotes the total number of sentence pairs. Let $\mathcal{E} = \{(\mathbf{u}_j, \mathbf{v}_j)\}_{j=1}^{M}$ denotes the collection of named entity pairs, where M is the number of entity pairs. Furthermore, $M_{\mathcal{X}\mapsto\mathcal{Y}}$ refers to a well-trained translation model that trained on $\mathcal{C}$.

Translation Difficulty. Translation difficulty refers to how difficult the level of complexity is associated with translating an entity pair using a translation model without any contextual information. We employ normalized sentence-level probability as the metric for calculating the translation difficulty score to measure this difficulty score.

$$T(u, v) = 1 - p_{M_{\mathcal{X}\mapsto\mathcal{Y}}}(v|u) \quad (1)$$

where the $p_{M_{\mathcal{X}\mapsto\mathcal{Y}}}(v|u)$ is the sentence-level probability normalized by the length of target entity phrase v.

A high translation difficulty score indicates that it is hard to generate the entity translation correctly for the translation model $M_{\mathcal{X}\mapsto\mathcal{Y}}$.

Context Diversity. Context diversity refers to the entropy of diverse contexts associated with the entity pair in the training corpus. In accordance with Mikolov et al. (2013), we employ the surrounding words in the source sentences to represent the context and calculate the context diversity score for the entity pair based on the entropy of the context.

$$C(u, v) = -\sum_{h=0}^{H} p(c_h|u, v) \log p(c_h|u, v) \quad (2)$$

[1] The built in-house data will be publicly available at https://github.com/Skytliang/EntityTranslation.

Testset	Sentence	Sentences with Entity	Entity	Entity Occurrence
WMT 2020	2,000	696	340	1,143
In-House	101,515	101,515	20,303	125,544

Table 1: The statistics of entities in the WMT2020 and In-house test sets.

where $p(c_h|u,v)$ is the estimated probability of the context c_h of the entity pair (u,v), and is computed as follows,

$$p(c_h|u,v) = \frac{count(c_h|u,v)}{\sum_{h=0}^{H} count(c_h|u,v)} \quad (3)$$

where $count(c_h|u,v)$ denotes the frequency of the context c_h appears around the entity pair (u,v) in training corpus, and H is the number of context.

2.1 Experimental Setup

Data. We conducted experiments on the WMT20 Chinese-to-English translation task. We first utilized TexSmart (Liu et al., 2021)[2] to recognize Chinese named entities and used unsupervised neural aligner Mask-Align (Chen et al., 2021a)[3] to obtain word alignments. Then, we collect bilingual named entity pairs using the source Chinese entities and word alignment information to build the named entity dictionary D.

To ensure the quality of the named entity dictionary D, we excluded entity pairs that appeared less than five times in the training corpus. We observed that most dictionary entity pairs (83.4%) constituted one-to-one translations, meaning each named entity had only one translation. This preliminary work only considers the entity pair in one-to-one translation mode.

Due to the insufficiency of named entities in the general news translation test set (only 34.8% of the WMT20 Chinese-to-English translation test sentences contained named entities), we constructed a new entity translation test set for the preliminary experiment to evaluate entity translation performance more accurately. Specifically, for each entity pair in the dictionary D, we randomly selected five bilingual sentence pairs containing the entity pair from our in-house data. Entity pairs for which five test samples could not be collected were excluded from the test set.

We show the data statistics of WMT20 and in-house test set in Table 1. The in-house data offers three key advantages: 1) An extensive collection of sentences; 2) A diverse range of entities, ensuring each sentence contains entities; and 3) A comprehensive evaluation system, with each entity being present in five test sentences, facilitates to assess the translation accuracy in various contexts.

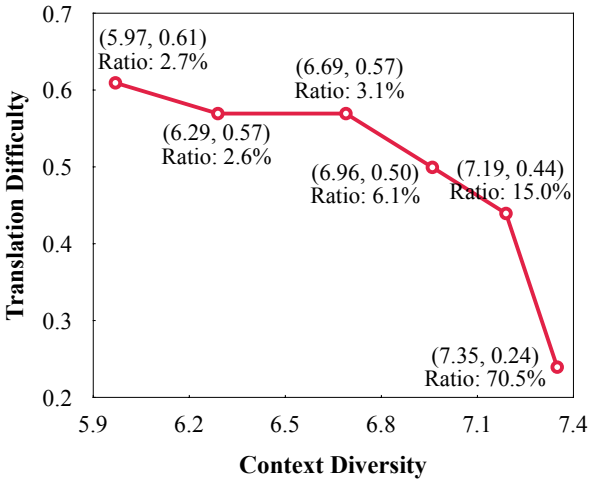

Figure 1: The attribute value of the centroid and the percentage of entity pairs for each group are presented. The points, arranged in a left-to-right order, represent the groups in terms of their accuracy 0%,20%,...,100%.

Feature Computation. For each entity pair in the dictionary D, we first employed a Transformer-Big model (Vaswani et al., 2017) to train the translation model $M_{\mathcal{X}\mapsto\mathcal{Y}}$. Subsequently, this translation model was utilized to calculate the translation difficulty score. Then, we followed Mikolov et al. (2013) to set the context window size to 3 (i.e., concatenating the three preceding tokens and three succeeding tokens to denote the context c_h) to calculate the context diversity score.

Entity Translation Accuracy. We use Exact-Match Accuracy (Anastasopoulos et al., 2021) to measure entity translation accuracy. This metric considers only those entity translations that appear in the output translation as correct predictions.

2.2 Result

We computed the translation accuracy for each pair of entities, representing it as a percentage (i.e., 0%,

[2]https://ai.tencent.com/ailab/nlp/texsmart/en/
[3]https://github.com/THUNLP-MT/Mask-Align

20%, ..., 100%) and categorized the entities into six groups based on this criterion. For better visualization, we calculated the centroid of each group using Euclidean distance. Figure 1 illustrates the attribute values of the centroid, along with the corresponding percentage of entity pairs for each group.

Our findings indicate that 1) most entity pairs exhibit a high translation accuracy, 85.5% (the right two points in Figure 1, 15.0% + 70.5% = 85.5%) of entity pairs have an accuracy rate higher than 80%. 2) The accuracy of entity translation predominantly depends on both the difficulty of translation and the diversity of the context. Entity pairs with challenging translations and limited contextual variation (represented as sample points in the upper left area of Figure 1) demonstrate lower accuracy.

Our work aims to identify the named entity pairs with low translation accuracy and improve entity translation performance. As shown in Figure 1, our preliminary experiment revealed that entity pairs exhibiting high translation difficulty scores and low context diversity scores generally possess low translation accuracy. In Section 3, we will apply a novel data augmentation technique to improve the translation accuracy of these particular entity pairs.

3 Methodology

In this section, we present a novel data augmentation strategy to improve the accuracy of named entity translation. The translation model training processing of our strategy includes following steps:

(1) **Entity-Pair Dictionary Collection** We use the named entity toolkit and the unsupervised word alignment toolkit to obtain the bilingual entity pair dictionary. It should be noted that This step can be skipped if an existing entity-pair dictionary is available.

(2) **Translation Model Training** We use the conventional Transformer model to train the translation model, which will be used to compute the normalized sentence-level probability for the named entity pair.

(3) **Feature Computation** For each entity pair in the entity-pair dictionary, we use the translation model to compute the difficult score as described in Eq.(1) and we follow Mikolov et al. (2013) to set the context window size to 3 to calculate the context diversity score as described in Eq.(2).

Name	#Training	#Test
News Zh-to-En	20,000,000	2,000
Terminology En-to-Fr	15,523,986	2,100
Terminology En-to-Zh	19,637,866	2,100

Table 2: The statistics of the data used in our translation experiments.

(4) **Augmentation Size Computation** We use translation difficulty and context diversity scores to compute the augmentation factor r for each entity pair,

$$r(u,v) = \frac{T(u,v)}{C(u,v)} \quad (4)$$

We calculate the augmentation size $N_{aug} = r(u,v) * N$ for the entity pair (u,v), where N is a pre-define hyper-parameter.

(5) **Data Agumentation** We randomly select N_{aug} sentence pairs from the training data and replace its entity pair with the augmented entity pair (u,v) to construct the synthetic parallel samples. This replacement is limited to entity pairs with identical entity tags to ensure semantic consistency.

(6) **Final Translation Model Training** We combine the original training data with the augmented data and train the final translation model from scratch.

4 Experiment

4.1 Experimental Setup

Data. We conducted experiments on Chinese-to-English translation and used WMT20 news translation test sets to evaluate the proposed approach. Additionally, we utilized the WMT21 terminology translation (Alam et al., 2021) for targeted evaluation. We follow the approach outlined by Wang et al. (2022a) to handle entity pairs in the one-to-many translation mode. The statistics of the data are described in Table 2.

Comparison Systems. We use the following approaches as baseline methods in comparison.

- BASELINE We use the Transformer-Big (Vaswani et al., 2017) to train the baseline system.

	Source	Target
ORIGINAL SAMPLE	bushi fangwen moxige	Bush visits Mexico
PLACEHOLDER	PER fangwen moxige	PER visits Mexico
TERMMIND	$ bushi / Bush $ fangwen moxige	Bush visits Mexico
TEMPLATE	C_0 bushi $ C_0 X_0 $ X_0 fangwen moxige	C_0 Bush $ C_0 Y_0 $ Y_0 visits Mexico
OUR	bushi he shalong juxing huitan	Bush held a talk with Sharon

Table 3: Examples for various named entity translation approaches. The target named entity pairs are underlined word "bushi" "Bush".

- PLACEHOLDER (Luong et al., 2015) that employs entity tags as placeholders to replace the named entity pairs within the training data. PLACEHOLDER adopts a post-processing step to recover the named entities.
- TERMMIND (Wang et al., 2021) that adopts the code-switching strategy to incorporate named entity translation into the source sentence. TERMMIND promotes the translation model to copy the named entity translation to the target sentence.
- TEMPLATE (Wang et al., 2022a) that rearranges the target sentence of named entities and context through a template in the training data. As for generation, TEMPLATE uses the named entities as the constrained prefix to decode the context part.

System Training. We used the open-source toolkit Fairseq (Ott et al., 2019) to implement the model. Specifically, we chose TRANSFORMER-BIG as our model, which consisits of an encoder of 6 layers and a decoder of 6 layers. We followed the settings in the original works to train the models. In brief, we trained the TRANSFORMER model for 25K steps with 131K (4096×32) tokens per batch. We warmed up it for 10K steps and decayed the learning rate based on the inverse square root of the update number. We used 8 Nvidia V100 GPUs to conduct the experiments and averaged the last five checkpoints to achieve strong performance.

Table 3 shows the data examples for the listed approaches. We adopted the same training settings but different data construction methods for these approaches. In our augmentation step, we obtained context templates while constructing the dictionary D as introduced in Section 2.1. It is important to note that these two steps are performed simultaneously, and excessive additional computing resources are not required.

To make a fair comparison, we conduct experiments with the same bilingual dictionary D for all systems. For PLACEHOLDER, we add a post-processing step to the BASELINE system. For TERMMIND and TEMPLATE, we follow Wang et al. (2021) and Wang et al. (2022a) to train the models. For the proposed data augmentation approach, we use the BASELINE to compute the normalized sentence-level probability for the translation difficulty $T(u, v)$. In addition, we set the pre-define hyper-parameter N in Section 3 to 350.

Evaluation. For the automatic evaluation, we used SacreBLEU (Post, 2018), NIST (Doddington, 2002), COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) for the sentence evaluation. we followed Alam et al. (2021) to adopt Exact-Match Accuracy (Acc), Window Overlap (Overlap), and Terminology-biased TER (1-TERm) to perform named entity evaluation.

We also performed a human evaluation to assess the quality of the translated outputs. We randomly chose 200 sentences from the test set and tasked the annotators with comparing our system against the baseline to determine which system output exhibited superior quality.

Furthermore, we proposed an entity-aware machine translation metric that prefers the translation output to generate more accurate named entities, and used the proposed entity-aware metric to evaluate the sentence quality.

4.2 Result

Table 4 presents the experimental results for Chinese-to-Enlgish translation. In the general decoding scenario, where lexical constraints are not applied, all systems significantly improve entity translation accuracy over the baseline. In the lexically constrained decoding scenario, PLACEHOLDER achieves the highest extract-match accuracy, with a score of 95.46 in the WMT20 test

Method	Sentence Evaluation				Entity Evaluation		
	BLEU	NIST	COMET	BLEURT	Acc	Overlap	1-TERm
General Decoding							
BASELINE	27.93	9.62	32.30	59.52	81.58	41.08	42.31
TERMMIND	27.25	9.60	30.38	58.99	87.47	39.91	41.46
OUR	**27.95**	**9.63**	**32.90**	**59.70**	**90.23**	**42.13**	**42.59**
Lexically Constrained Decoding							
PLACEHOLDER	27.34	9.61	31.63	59.25	95.46	40.58	42.32
TEMPLATE	27.07	9.60	32.01	59.44	98.68	40.58	41.67
+OUR	**28.30**	**9.65**	**33.44**	**59.97**	**98.89**	**42.77**	**42.99**

Table 4: Translation performance and entity translation accuracy of WMT20 Chinese-to-English translation.

Method	WMT 2020		
	BASELINE	OUR	COMPARABLE
ANNOTATOR 1	64	**98**	38
ANNOTATOR 2	49	**103**	48
TOTAL	113	**201**	86

Table 5: Human Evaluation of Chinese-to-English translation. Intra-agreement Scott's Pi: 62.68%.

Method	English-to-French				English-to-Chinese			
	BLEU	Acc	Overlap	1-TERm	BLEU	Acc	Overlap	1-TERm
General Decoding								
BASELINE	45.31	84.99	32.45	56.73	39.08	64.36	35.61	43.49
TERMMIND	44.39	90.10	32.88	57.13	38.69	79.86	37.66	44.85
OUR	**45.48**	**94.60**	**33.24**	**57.97**	**39.44**	**84.86**	**41.14**	**46.43**
Lexically Constrained Decoding								
PLACEHOLDER	44.23	96.22	33.29	57.12	38.75	97.02	39.06	45.88
TEMPLATE	44.62	99.35	33.87	58.23	40.32	100.00	42.81	49.40
+OUR	**45.94**	**99.53**	**34.31**	**58.86**	**40.67**	**100.00**	**43.28**	**50.04**

Table 6: Translation performance and Terminology Translation performance on WMT21 Terminology testsets.

set. OUR achieves the best overall performance by enhancing entity translation accuracy while preserving the BLEU score.

We conducted a human evaluation of the WMT20 Chinese-to-English translation. Table 5 provides a comprehensive account of the human evaluation. The human evaluation results further support the notion that the proposed strategy yields translation output of superior quality compared to the baseline system.

On the WMT21 terminology translation task, as shown in Table 6, our approach also outperforms the baseline system, TERMMIND and TEMPLATE in both English-to-French and English-to-Chinese translation tasks, demonstrating the effectiveness and universality of the proposed approach.

4.3 Ablation Study

In the ablation study, we analyze the impact of Translation Difficulty (Factor T) and Context Diversity (Factor C) as described in section 2.

As shown in Table 7, both Factor T and C make significant contributions to the translation accuracy of entities. We find that the impact of factor C on performance is greater. In other words, entities are coupled with certain specific contexts and entity translation errors usually occur when encountering other less common contexts. In our understanding,

Method	Sentence Evaluation				Entity Evaluation		
	BLEU	NIST	COMET	BLEURT	Acc	Overlap	1-TERm
General Decoding							
OUR	27.95	9.63	32.90	59.70	90.23	42.13	42.59
W/O FACTOR T	28.10	9.64	32.65	59.57	84.05	41.39	42.37
W/O FACTOR C	27.86	9.62	32.71	59.63	82.82	41.32	42.40
Lexically Constrained Decoding							
TEMPLATE + OUR	28.30	9.65	33.44	59.97	98.89	42.77	42.99
W/O FACTOR T	27.47	9.61	32.45	59.51	97.85	42.18	42.36
W/O FACTOR C	27.33	9.61	32.24	59.46	98.71	41.86	42.00

Table 7: Ablation study of WMT20 Chinese-to-English translation.

we consider T to be the translation difficulty of the entity itself. For example, some entities can be translated literally and are easily learned by the model even if they do not appear frequently.

However, for some non-literal and challenging entities, we aim to enrich their context to help translate them better. We regard T as an internal factor of entities and C as an external factor. Previous work only considers lexical constraints while ignoring the causes of entity translation errors, which is also the motivation of our work.

5 Entity-aware Machine Translation Evaluation

We propose a simple and effective entiy-aware evaluation metric for assessing machine translation quality. The main idea involves considering the translation accuracy of the entities during the calculation of the evaluation score, using both the translation output and the reference translation. In particular, when determining the value, if an entity matches the translation output, its weight will be correspondingly increased.

We apply the entiy-aware evaluation strategy on the SacreBLEU (Post, 2018) and ChfF+ (Popović, 2017). We assign an entity-aware weight of 4, which entails calculating the values of the n-grams (words or characters) by proportionally increasing the weights of the values present in the matching entities. We conducted experiments WMT19 metric tasks[4] on four high-resource pairs (from Chinese/zh, German/de, Russian/ru, Finnish/fi to English/en) and three low-resource pairs (from Gujarati/gu, Kazakh/kk, Lithuanian/lt to English/en). To make a fair comparison, we adopt the tokenizer_13a[5] to tokenize the text and apply the Stanford NER system[6] to tokenized text.

Table 8 lists the system-level human correlation results on WMT19 metric tasks. We observe that our entity-aware evaluation metric achieves better human correlation compared to the corresponding baseline system for most high-resource language pairs, with the exception of Finnish.

For the experiments in Table 4 and Table 6, we apply the entity-aware SacreBLEU and ChfF+ to the high-resource pairs (Chinese-to-English, English-to-French and English-to-Chinese) and observe the proposed approach OUR achieves improvements over the baseline system in terms of the entity-aware evaluation metric.

6 Related Work

Decoding Algorithm One line of approaches to lexically constrained NMT focuses on modifying the decoding algorithm to impose lexical constraints. Hokamp and Liu propose *grid beam search* (GBS) algorithm, which takes target-side pre-specified translation as lexical constraints at each decoding step. However, the decoding speed of GBS scales linearly with the number of constraints. To reduce the computation complexity, Post and Vilar propose the *dynamic beam allocation* (DBA) method, which limits the decoding complexity by dynamically providing a fixed size of beam to the decoder. VDBA (Hu et al., 2019) gives a fast version of DBA and supports batched decoding. A potential issue with these methods is the lack of consideration for translation fidelity, as there is no indication of a matching source for each

[4]https://www.statmt.org/wmt19/metrics-task.html

[5]https://github.com/mjpost/sacrebleu/blob/master/sacrebleu/tokenizers/tokenizer_13a.py

[6]https://nlp.stanford.edu/software/CRF-NER.shtml

Method	High resource					Low resource				Avg
	zh–en	de–en	ru–en	fi–en	avg	gu–en	kk–en	lt–en	avg	
SacreBLEU	80.7	79.4	81.3	98.5	85.0	97.5	91.2	96.7	95.1	89.3
+Entity	**83.2**	**85.5**	**85.4**	98.5	**88.2**	97.2	88.3	**97.2**	94.2	**90.8**
chrF+	85.1	86.0	87.8	99.2	89.5	96.1	76.9	93.4	88.8	89.2
+Entity	**87.6**	**89.0**	**89.4**	96.3	**90.6**	92.8	**82.1**	**95.2**	**90.0**	**90.3**

Table 8: WMT19 system-level human correlation (Pearson), for to English systems. We follow Agrawal et al. (2021) to remove the outlier systems.

pre-specified translation. Hasler et al. use decoder attentions to match target-side constraints and their corresponding source words. There are also several other constrained decoding algorithms that utilize word alignments to impose constraints(Song et al., 2020; Chen et al., 2021b). Although the alignment-based decoding methods are faster than previous works, they remain significantly slower compared to the standard beam search algorithm.

Annotating Strategy Another line of studies focus on annotating named entities to provide type and annotating information. Li et al. introduces more linguistic features by inserting special tokens before and after entities in the source sentence. Other researchers (Ugawa et al., 2018b; Modrzejewski et al., 2020) add entity embeddings when encoding the source sentence, helping models differentiate constrained and unconstrained tokens. Dinu et al. use source factors successfully to enforce terminology. Xie et al. attach entity classifiers to the transformer and design an adaptive loss function for named entities. To make NMT models better learn from and cope with lexical constraints, Wang et al. propose to leverage attention modules to explicitly integrate vectorized lexical constraints. Although these works enhance the representation of sentences, they still face a scalability challenge as they attempt to adapt the architecture of NMT models for this task.

Lexical Constraint Moreover, researchers have shown the benefit of editing the training data to induce constraints. Luong et al. use annotated *unk* tags to present the out-of-vocabulary (OOV) words in training corpora, where the correspondence between source and target *unk* symbols are obtained from word alignment. Output *unk* tags are replaced by the corresponding target constraints in a post-processing stage. Further research (Crego et al., 2016; Yan et al., 2019) extends *unk* tags symbols to specific symbols that can present name entities. Due to the loss of word meaning when representing them with placeholder tags, these methods may hinder next-token prediction in the generation process. Target Lemma Annotation (TLA)(Bergmanis and Pinnis, 2021) annotates source entity words with their target language lemmas instead of meaningless symbols to fix the problem of word meaning loss for placeholders. TermMind (Wang et al., 2021), TADA (Ailem et al., 2021) and Kakao (Bak et al., 2021) use a similar method to TLA besides minor difference. To enhance the proportion of constrained target words' occurrences, Code-Switching (Song et al., 2019) directly replaces the source words with their target translations, allowing the model to learn lexicon translations by copying source-side target words. However, it also introduces many target-side tokens and expands the source-side vocabulary. Zeng et al. add target constraints as a prefix of the decoder input to fix the problem of larger vocabulary size for Code-Switching. Further, (Wang et al., 2022a) proposes a template-based translation framework to handle constraints, achieving high match accuracy while maintaining the inference speed.

7 Conclusion

We demonstrate that the translation accuracy of named entities is influenced by both their level of translation difficulty and the diversity of contextual information. We propose a novel data augmentation strategy that boosts the translation probability and the context diversity of the targeted named entity pair. Experimental results show that the proposed strategy outperforms the baseline in terms of general translation performance and entity translation accuracy. We also propose an entity-aware machine translation metric that prefers the translation output to generate more accurate named entities.

Limitations

A limitation of this work is that our method can not cope with one-to-many (entity with more than one translation candidate) constraints. Moreover, efficiently applying the proposed data augmentation strategy to large-scale training data remains a future challenge.

References

Sweta Agrawal, George Foster, Markus Freitag, and Colin Cherry. 2021. Assessing reference-free peer evaluation for machine translation. In *NAACL*.

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Lingua custodia's participation at the WMT 2021 machine translation using terminologies shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 799–803, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the wmt shared task on machine translation using terminologies. In *WMT*.

Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, et al. 2021. On the evaluation of machine translation for terminology consistency. *arXiv*.

Yunju Bak, Jimin Sun, Jay Kim, Sungwon Lyu, and Changmin Lee. 2021. Kakao enterprise's WMT21 machine translation using terminologies task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 804–812, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Chi Chen, Maosong Sun, and Yang Liu. 2021a. Maskalign: Self-supervised neural word alignment. In *ACL*.

Guanhua Chen, Yun Chen, and Victor OK Li. 2021b. Lexically constrained neural machine translation with explicit alignment guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12630–12638.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems.

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.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *AAAI*.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *ACL-IJCNLP*.

Chris Hokamp and Qun Liu. 2017a. Lexically constrained decoding for sequence generation using grid beam search. In *ACL*.

Chris Hokamp and Qun Liu. 2017b. 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, Vancouver, Canada. Association for Computational Linguistics.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained

decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *WMT*.

Zhongwei Li, Xuancong Wang, Ai Ti Aw, Eng Siong Chng, and Haizhou Li. 2018. 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.

Lemao Liu, Haisong Zhang, Haiyun Jiang, Yangming Li, Enbo Zhao, Kun Xu, Linfeng Song, Suncong Zheng, Botong Zhou, Dick Zhu, et al. 2021. Texsmart: A system for enhanced natural language understanding. In *ACL-IJCNLP*.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *ACL*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv*.

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.

Diego Moussallem, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan. 2019. Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of the 10th international conference on knowledge capture*, pages 139–146.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL (Demonstrations)*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *WMT*.

Matt Post and David Vilar. 2018a. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *NAACL*.

Matt Post and David Vilar. 2018b. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *EMNLP*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.

Kai Song, Kun Wang, Heng Yu, Yue Zhang, Zhongqiang Huang, Weihua Luo, Xiangyu Duan, and Min Zhang. 2020. Alignment-enhanced transformer for constraining nmt with pre-specified translations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8886–8893.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018a. Neural machine translation incorporating named entity. In *COLING*.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018b. 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.

Ke Wang, Shuqin Gu, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. 2021. Termmind: Alibaba's wmt21 machine translation using terminologies task submission. In *WMT*.

Shuo Wang, Peng Li, Zhixing Tan, Zhaopeng Tu, Maosong Sun, and Yang Liu. 2022a. A template-based method for constrained neural machine translation. In *EMNLP*.

Shuo Wang, Zhixing Tan, and Yang Liu. 2022b. 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.

Xing Wang, Zhaopeng Tu, Deyi Xiong, and Min Zhang. 2017. Translating phrases in neural machine translation. In *EMNLP*.

Rongxiang Weng, Hao Zhou, Shujian Huang, Lei Li, Yifan Xia, and Jiajun Chen. 2019. Correct-and-memorize: Learning to translate from interactive revisions. In *IJCAI*.

Yanling Xiao, Lemao Liu, Guoping Huang, Qu Cui, Shujian Huang, Shuming Shi, and Jiajun Chen. 2022. Bitiimt: A bilingual text-infilling method for interactive machine translation. In *ACL*.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Machine Learning*, pages 1–23.

Jinghui Yan, Jiajun Zhang, JinAn Xu, and Chengqing Zong. 2018. The impact of named entity translation for neural machine translation. In *China Workshop on Machine Translation*.

Jinghui Yan, Jiajun Zhang, JinAn Xu, and Chengqing Zong. 2019. The impact of named entity translation for neural machine translation. In *Machine Translation: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings 14*, pages 63–73. Springer.

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Xu Tan, Tao Qin, and Tie yan Liu. 2023. Extract and attend: Improving entity translation in neural machine translation.