Challenges of Neural Machine Translation for Short Texts

Yu Wan*
NLP^{2}CT Lab, University of Macau
nlp2ct.ywan@gmail.com

Baosong Yang**
Alibaba Group
yangbaosong.ybs@alibaba-inc.com

Derek Fai Wong**
NLP^{2}CT Lab, University of Macau
derekfw@um.edu.mo

Lidia Sam Chao
NLP^{2}CT Lab, University of Macau
lidiasc@um.edu.mo

Liang Yao
Alibaba Group
yaoliang.yl@alibaba-inc.com

Haibo Zhang
Alibaba Group
zhanhui.zhb@alibaba-inc.com

Boxing Chen
Alibaba Group
boxing.cbx@alibaba-inc.com

Short texts (STs) present in a variety of scenarios, including query, dialog, and entity names. Most of the exciting studies in neural machine translation (NMT) are focused on tackling open problems concerning long sentences rather than short ones. The intuition behind is that, with respect to human learning and processing, short sequences are generally regarded as easy

∗ This research was accomplished when Yu Wan was interning at Alibaba DAMO Academy.
∗∗ Baosong Yang and Derek Fai Wong are co-corresponding authors.

Submission received: 29 April 2021; revised version received: 14 December 2021; accepted for publication: 17 December 2021.

https://doi.org/10.1162/COLI_a_00435

examples. In this article, we first dispel this speculation via conducting preliminary experiments, showing that the conventional state-of-the-art NMT approach, namely, TRANSFORMER (Vaswani et al. 2017), still suffers from over-translation and mistranslation errors over STs. After empirically investigating the rationale behind this, we summarize two challenges in NMT for STs associated with translation error types above, respectively: (1) the imbalanced length distribution in training set intensifies model inference calibration over STs, leading to more over-translation cases on STs; and (2) the lack of contextual information forces NMT to have higher data uncertainty on short sentences, and thus NMT model is troubled by considerable mistranslation errors. Some existing approaches, like balancing data distribution for training (e.g., data upsampling) and complementing contextual information (e.g., introducing translation memory) can alleviate the translation issues in NMT for STs. We encourage researchers to investigate other challenges in NMT for STs, thus reducing ST translation errors and enhancing translation quality.

1. Introduction

Short texts (STs) refer to examples that contain fewer tokens within sequences.[1] Translating STs from one language to another plays a crucial part in natural language processing (NLP) scenarios, including the modeling of query (Huang et al. 2016; Song, Kim, and Park 2017; Saleh and Pecina 2020; Bi et al. 2020; Yao et al. 2020a, 2020b), dialogue (Wang et al. 2017; Liu et al. 2018), title (Kreutzer et al. 2018; Karakanta, Dehdari, and van Genabith 2018; Darwish and Sayaheen 2019; Etchegoyhen and Gete 2020; Banar, Daelemans, and Kestemont 2020), entity name (Jiang et al. 2007; Zhao et al. 2020), and text matching (Chen et al. 2020; Lyu et al. 2021).

With the length of one sentence being short, fewer tokens are involved to form the completeness of its semantic. With the perspective of human linguistic intuition, STs are generally engaged with fewer combinations of lexical components. Consequently, they are regarded as more easily learned and translated compared with longer sentences (Le, Martinez, and Matsumoto 2017), as the corresponding complexity of examples decreases exponentially with their length (Jiang et al. 2015). This intuition is widely mentioned in related research with respect to neural machine translation (NMT, Kocmi and Bojar 2017; Hasler et al. 2017; Liu et al. 2019) and results in the lack of exploration on NMT for STs.

However, the speculation "STs are easy for NMT learning," which originates from linguistics, is hardly well supported empirically. Modern NMT engines still give inappropriate translations when input source sentences become shorter (see Table 1). Additionally, based on experimental results derived from existing studies (Cho et al. 2014; Bahdanau, Cho, and Bengio 2015; Toral and Sánchez-Cartagena 2017; Neishi and Yoshinaga 2019), the translation quality over STs is hardly improved compared with sequences with other length ranges. To confirm this, we represent an empirical study in Figure 1. As shown, we compare the translation quality between phrase-based statistical machine translation (PBSMT, Koehn 2004) and NMT (Vaswani et al. 2017) over different length buckets.[2] Compared with PBSMT, NMT consistently reduces the translation error rates across all length buckets, showing its dominant performance.

1 Following previous studies (Bahdanau, Cho, and Bengio 2015; Zhang et al. 2016; Murray and Chiang 2018), sequences that contain no more than 10 tokens are treated as STs in this research.

2 Experimental settings are concluded in Section 4.3.

Table 1
Zh⇒En ST translation examples. For each case, HYP1 denotes the translation using website engine (`https://fanyi.baidu.com`), and HYP2 is derived via trained machine translation system on WMT'17 Zh⇒En dataset. Chinese word "帮" (help) is directly translated, leading the phrase "帮...劝架" (stop . . . from fighting) mistranslated. For the second case, the word "劝架" (not fighting) is under-translated, and the order of clause is better located at the start of output.

Case 1										
SRC	他	从不	帮	我们	劝架	。				
HYP1	He	never	helps	us	argue	.				
HYP2	He	never	helps	us	out	.				
REF	He	never	stops	us	from	fighting	.			
Case 2										
SRC	他	从不	帮	我们	劝架	除非	他	很	闲	。
HYP1	He	never	helps	us	unless	he	is	free	.	
HYP2	He	never	helps	us	unless	he	is	idle	.	
REF	Unless	idle	,	he	never	stops	us	from	fighting	.

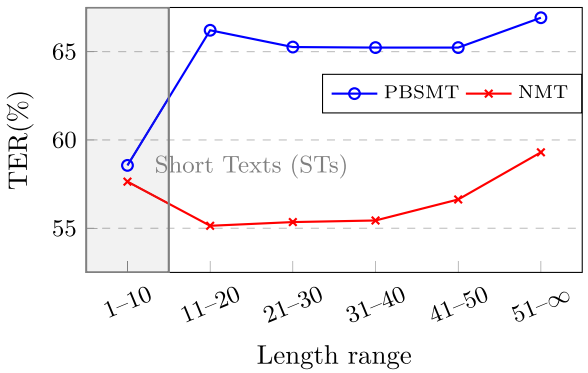


Figure 1
Comparison on TER scores over each bucket of WMT14 En⇒De dev set with phrase-based statistical machine translation (PBSMT, Koehn, Och, and Marcu 2003) and neural machine translation (NMT, Vaswani et al. 2017) model. Lower score is better. NMT significantly improves translation quality compared with PBSMT except ST set. Interestingly, PBSMT gives the best performance over STs compared with other buckets, whereas NMT does not.

Nevertheless, the performance gap over short sequences (bucket $[1, 10]$) is relatively marginal compared with other buckets. More interestingly, consistent with previous research (Toral and Sánchez-Cartagena 2017), the PBSMT model gives the lowest error rate when translating STs, whereas the NMT model performs best over sequences with medium lengths (bucket $[11, 20]$).

In this article, we promote our research on investigating why the NMT model does not achieve excellent translation quality over STs, as well as promising approaches to improve corresponding translation accuracy. First, we categorize translation errors (Snover et al. 2006; Tu et al. 2016) into four types—**over-translation**, **under-translation**, **mistranslation**, and **misorder**—to arrange the comparison of translation

quality with a more precise and concrete manner. By conducting experiments over WMT'14 English⇒German (En⇒De) and WMT'17 Chinese⇒English (Zh⇒En) MT tasks, we have derived interesting findings and conclusions as follows:

- *Data imbalance*: We first observe that, with the sequence length being longer, the ratio of over-translation errors generated by NMT model tends to decrease first, and then increase until the longest range. We assume that this phenomenon is created by the imbalanced distribution of training data with respect to sequence length. By quantifying the inference miscalibration (Wang et al. 2020) over all length buckets, we find that the length range whose distribution dominates among all length buckets is engaged with a higher level of miscalibration when decoding. We find that balancing the data used when training, for example, upsampling the training data with a smaller portion of the entire set (Hendrycks et al. 2018), can relieve the model inference miscalibration, thus improve NMT translation quality over STs. This verifies that the over-translation issue over STs is caused by data distribution.
- *Contextual information*: We also find that a higher ratio of mistranslation errors is assigned with shorter sequences. We presume that the reason behind this phenomenon remains the insufficiency of contextual information, as STs contain less semantic information due to their short lengths. With the help of model uncertainty measurement (Dong, Quirk, and Lapata 2018), we demonstrate that NMT has a higher level of model uncertainty over STs. We empirically prove that providing additional contextual semantics for the NMT model, for example, leveraging translation memory (Cao and Xiong 2018), is helpful for reducing model uncertainty, thus can increase the output quality when translating STs.

These findings verify that, for modern NMT approaches, STs may not be learned and translated as easily as we intuitively thought. In this research, to our best knowledge, we are the first to investigate the shortcomings when an NMT model translates STs empirically. We hope that this work can appeal to other NLP researchers, as well as increase their interest on further investigating the topic of NMT for STs. We are also excited to see if the NLP community can further investigate other possible reasons why translation quality of NMT for ST is unsatisfying, as well as propose other promising solutions to pursue better ST translation quality.

2. Related Work

NMT has proven its superior translation quality in recent years (Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015; Vaswani et al. 2017; Ott et al. 2018; Nguyen et al. 2020). The distribution of training data is always of vital importance in NMT model training and analysis. Some length ranges represent only a small part of the training examples, meaning that model training is less frequent for sentences of these corresponding lengths. This thus degrades the accuracy of generated outputs. In this section, we give an overview of existing studies that are related to such an idea.

2.1 Model Training on Sequence Length

On one hand, many studies were motivated by the following human intuition: Longer sequences are more difficult to learn. As modeling those sequences is a challenge due to the long-term dependency problem (Sutskever, Vinyals, and Le 2014; Cho et al. 2014; Liu et al. 2019), enhancing the ability of processing long sentences can also improve the overall translation quality. Sutskever, Vinyals, and Le (2014) and Cho et al. (2014) proposed to use recurrent neural networks (RNNs, Elman 1990) with gated mechanism, that is, long short-term memory (LSTM, Hochreiter and Schmidhuber 1997) and gated recurrent unit (GRU, Cho et al. 2014), to help preserve the semantic information at longer distance; Bahdanau, Cho, and Bengio (2015) used the attention mechanism where target representation can leverage the semantic information from the source side for reference; Luong, Pham, and Manning (2015) proposed the local attention module, which constrains the source of attentive representations with a fixed window size; and Vaswani et al. (2017) introduced TRANSFORMER to utilize the token wise semantic with a sentence-level attention paradigm, which can also be implemented with self-attention at the source or target side only.

Additionally, some model training accelerations are inspired by this intuition. Curriculum learning (Bengio et al. 2009), which trains the model by mimicking the learning process of humans, suggests arranging the training data in an easy-to-hard order. After designing the required curricula carefully, easy examples are applied to form the "warm-up" phase, heuristically accelerating the convergence of the model. In NMT, Kocmi and Bojar (2017) proposed gathering the examples with similar sequential lengths to accelerate model training; Platanios et al. (2019) used two metrics, namely, word rarity and sequence length, for identifying the data difficulty to fix up the training order of examples; Zhang et al. (2018) further investigated the effectiveness of involved curriculum learning approaches of different curricular designs, confirming the effectiveness of short sequences being the easiest curricula for model training.

To conclude, the studies above are all derived from the speculation that longer examples are harder for both human and NMT models, and shorter examples can ease human learning and NMT model training.

2.2 Model Analyses on Sequence Length

On the other hand, many studies investigated the performance of NMT, as well as PBSMT models, over various sets of length ranges. Cho et al. (2014) showed that a RNN model performs worse over short and long sentences, whereas PBSMT is more stable when handling these cases; Bahdanau, Cho, and Bengio (2015) showed that a RNN model with attention mechanism can ameliorate the translation quality of long sequences, whereas the gap of performance over short sequences is rather marginal compared with an RNN model without the attention mechanism; Toral and Sánchez-Cartagena (2017) investigated the translation quality between PBSMT and NMT systems at the level of character; Neishi and Yoshinaga (2019) showed that TRANSFORMER translation quality over the bucket of examples with given range is impacted by the unevenness of corresponding data distribution.

2.3 Our Work

Different from previous research, we focus on NMT for ST scenarios only. In this article, we take the analysis further by categorizing translation errors in a more fine-grained

manner. We also apply two metrics, namely, inference miscalibration (Wang et al. 2020) and model uncertainty (Gal and Ghahramani 2016), to analyze model performance with newer perspectives aside from overall translation quality.

3. Background

In this section, we briefly introduce NMT, and, more importantly, the model architecture of the TRANSFORMER model.

3.1 A Brief Introduction to NMT

NMT aims to build a deep neural network that accepts a sequence $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_I\}$ from the source language and generates its corresponding result $\mathbf{T} = \{\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_J\}$ on the target side automatically. Specifically, the state-of-the-art NMT model architecture TRANSFORMER (Vaswani et al. 2017) follows the paradigmatic design of an autoencoder, where the encoder is in charge of encoding source tokens $\mathbf{S}$ to contextual representation $\mathbf{C} \in \mathcal{R}^{I \times d_m}$. Here I and d_m are the length of source sequence and embedding dimensionality, respectively.

Multihead attention modules play a crucial role in the TRANSFORMER model. The inputted representations are linearly transformed into query $\mathbf{Q} \in \mathcal{R}^{H \times L_1 \times d_h}$, key $\mathbf{K} \in \mathcal{R}^{H \times L_2 \times d_h}$, and value representation $\mathbf{V} \in \mathcal{R}^{H \times L_2 \times d_h}$; H and d_h represent the number of heads and the dimensionality in each head, respectively. For the h-th ($1 \leqslant h \leqslant H$) head, the attention network first calculates the attention energy to describe how the l_1-th ($1 \leqslant l_1 \leqslant L_1$) query vector $\mathbf{Q}_{h,l_1,\cdot} \in \mathcal{R}^{d_h}$ should attend to all key vectors $\mathbf{K}_{h,\cdot,\cdot} \in \mathcal{R}^{L_2 \times d_h}$, and derives the output representation by weighted-sum over all value vectors $\mathbf{V}_{h,\cdot,\cdot} \in \mathcal{R}^{L_2 \times d_h}$:

$$\mathbf{A}_{h,l_1,\cdot} = \text{softmax}(\frac{\mathbf{Q}_{h,l_1,\cdot}\mathbf{K}^{\top}_{h,\cdot,\cdot}}{\sqrt{d_m}}) \in \mathcal{R}^{L_2} \tag{1}$$

$$\mathbf{H}_{h,l_1,\cdot} = \sum_{k=1}^{L_2} \mathbf{A}_{h,l_1,k}\mathbf{V}_{h,k,\cdot} \in \mathcal{R}^{d_h} \tag{2}$$

In total, three types of attention networks are involved in the conventional TRANSFORMER model:

- The encoder self-attention networks (enc-SANs), which globally gather source semantics from all positions ($L_1 = L_2 = I$);
- The decoder self-attention networks (dec-SANs), which collect semantics at target to comply with the auto-regressive process ($L_1 = L_2 = J$);
- The encoder-decoder cross-attention networks (CANs), which devote into aligning two semantic spaces ($L_1 = J, L_2 = I$).

The learning objective is to maximize the probability of generating sequence $\mathbf{T}$ following a teacher forcing paradigm (Sutskever, Vinyals, and Le 2014):

$$\text{P}(\mathbf{T}) = \prod_{j=1}^{J} \text{P}(\mathbf{T}_j | \mathbf{T}_{<j}, \mathbf{C}, \theta) \tag{3}$$

Table 2
Number of sentences and tokens over WMT14 En⇒De and WMT17 Zh⇒En datasets. Results are conducted on source side at the level of subword units. M: million. For En⇒De experiments, we concatenate official released subsets from `newstest2013` to `newstest2016` for further analyses. As for Zh⇒En, the test set is collected among `newstest2017` to `newstest2020`.

	En⇒De		**Zh⇒En**	
	# Sents	# Tokens	# Sents	# Tokens
Train	4.56M	132M	20.6M	485M
Dev	3,000	74,957	2,002	56,374
Test	11,171	286,290	9,982	302,077
ST	1,273	9,807	717	5,418

where θ is the set of model parameters. For inference, the probability of generated sequence $\tilde{\mathbf{T}} = \{\tilde{\mathbf{T}}_1, \tilde{\mathbf{T}}_2, \cdots, \tilde{\mathbf{T}}_{\tilde{J}}\}$ is the accumulative multiplicative production based on the output of the model itself. When decoding at the $\tilde{j}$-th ($1 \leqslant \tilde{j} \leqslant \tilde{J}$) step, the output of model $\tilde{\mathbf{T}}_{<\tilde{j}}$ is used as the generated output to guide the model inference.

4. Preliminary Materials

In this section, we first describe the MT tasks involved for analysis and data preprocessing. Then, we introduce the experimental details and our method for further analysis that categorizes translation errors into four types. After counting the ratio of each translation error type, we directly check the effectiveness over baseline model training, as well as see that simply tuning the hyperparameters for decoding does not significantly improve the translation quality over STs. Last, we observe two interesting phenomena by counting the ratio of all error types over all length ranges, raising our interest in further investigating the rationale behind them.

4.1 Dataset

In this research, we choose two MT tasks for analysis: WMT'14 English⇒German (En⇒De) and WMT'17 Chinese⇒English (Zh⇒En), containing around 4.50 million and 20.1 million training examples, respectively. All datasets are tokenized, truecased,[3] and segmented into subword units by 32k byte-pair encoding (BPE) merging operations (Sennrich, Haddow, and Birch 2016). In addition, we use Jieba segmentation[4] to tokenize the Chinese datasets to maintain better performance, and BPE merging steps are learned separately. Note that, to preserve the quality of STs for analyses, as well as prevent data leaking and domain mismatching between training and test datasets, we collect WMT test sets from adjacent years for further analyses, resulting in 1,273 and 717 ST sentence pairs for En⇒De and Zh⇒En task, respectively (see Table 2).

For efficiency of model training, we remove all sequences that contain over 256 tokens in either language. Our experiments are conducted with the `fairseq`[5] open-source toolkit.

3 `https://github.com/moses-smt/mosesdecoder`.
4 `https://github.com/fxsjy/jieba`.
5 `https://github.com/pytorch/fairseq`.

4.2 Types of Translation Errors

As the BLEU metric (Papineni et al. 2002) can hardly reveal to what degree each type of translation error affects the MT model (Snover et al. 2006), and it prefers longer sequences for observing more *n*-grams and penalizes short sequences heavily (Och 2003; Nakov, Guzmán, and Vogel 2012), we propose to use more fine-grained statistics for further analyses. Inspired by the TER metric (Snover et al. 2006) and related research about translation error analysis (Tu et al. 2016), we directly involve all subentries of edit moves in TER: insertion, deletion, substitution, and shift, and assign them with four typical types of translation errors, respectively:

- **Over-translation** (Over.): candidate contains token(s) excluded from references;
- **Under-translation** (Under.): candidate fails to generate tokens for sufficiency;
- **Mistranslation** (Mistr.): candidate token is inaccurate compared with reference;
- **Misorder** (Misor.): candidate token (or a text span) is located in an incorrect position;

The reason behind this is that the listed translation error types above are highly related to each type of edit movement. Specifically, insertion means that some tokens are inserted into ground truth to obtain a translation candidate, which is identical to the case of over-translation; deleting several tokens from ground truth leads to under-translation issues; substituting candidate tokens in ground truth to obtain candidate results in mistranslation errors; and shifting tokens in ground truth to other positions is identical to misorder problems.

Then, to evaluate the translation quality with multiple aspects of translation errors, as well as fairly compare performance across different length buckets, here we calculate the ratio of each error type by normalizing the number of error cases with the number of tokens in the reference set:

$$r_t = \frac{\text{\# of edit moves for type } t}{\text{\# of tokens in reference}} \tag{4}$$

where $t \in \{\text{Over.}, \text{Under.}, \text{Mistr.}, \text{Misor}\}$. Note here that some types of errors can be transited into other types, for example, a mistranslation case is equivalent to the union of an over-translation and an under-translation case. In this research, we only consider the least number of required edits to recover from hypothesis to reference, as in Snover et al. (2006).

4.3 Hyperparameter Setting

For NMT approaches, we train both baseline and our models following the TRANSFORMER-*Base* setting (Vaswani et al. 2017; Ott et al. 2018). Each training mini-batch consists of approximately 32,768 source tokens. We use the learning rate schedule (Vaswani et al. 2017) where the linear warm-up phase takes 12k steps, reaching its maximum value at 0.0008 (Ott et al. 2018). Checkpoints are stored after every 1k steps, and

the whole training process takes 150k and 200k steps in total for En⇒De and Zh⇒En, respectively. We apply dropout regularization with a ratio of 0.1 for embedding layers, attention weights, ReLU activations, and residual connections. All NMT experiments are conducted on 4 Nvidia V100 GPUs.

For model inference, we apply beam size 4 and decoding alpha 0.6 (Vaswani et al. 2017) over the En⇒De dev set. As for the Zh⇒En task, the beam size and decoding alpha are 10 and 1.5, respectively. For each machine translation task, we conduct 5 independent runs, and leverage the checkpoints with best performance on the dev set for further analysis. As a result, our reimplementation gives 28.01 and 24.27 BLEU scores on average over the En⇒De `newstest2014` and Zh⇒En `newstest2017` test set, respectively. Compared with the reimplementations of other research (Vaswani et al. 2017; Wang et al. 2019; Ghazvininejad et al. 2019), our results show consistently higher BLEU scores, thus confirming the appropriateness of our further analyses.

As for PBSMT, we follow Koehn, Och, and Marcu (2003) and apply minimum risk training (Och 2003). The hyperparameters of log-linear models are automatically tuned over the dev set.

4.4 Model Performance

One may ask whether the NMT for ST can be simply improved by carefully tuning the hyperparameters for inference, that is, beam size and decoding alpha (Wu et al. 2016). Specifically, the former is in charge of controlling the number of top candidates, and the latter is used to regulate the penalty of candidate length. To check whether these two hyper-parameters can significantly improve ST translation quality, we conduct two series of experiments by tuning them respectively. As shown in Figure 2a, although larger beam size can decrease the ratio of mistranslation and over-translation cases, it significantly increases the under-translation errors. This is similar to previous findings that, with larger beam size, NMT prefers to generate shorter translations, leading to worse translation quality (Koehn and Knowles 2017; Yang, Huang, and Ma 2018; Murray and Chiang 2018). As to decoding alpha (Figure 2b), the ratio of each type of erroneous case does not vary much. Overall, tuning both hyperparameters does not improve the general translation quality. These experimental results dispel our concern, revealing that NMT for ST is hardly improved by simply tuning hyperparameters during inference.

4.5 Translation Error Analysis

To check why the performance of NMT over STs is unsatisfied, we directly accumulate different types of translation errors over the NMT baseline model. Following the categories of translation error types in Section 4.2, we construct error ratios of all types over each bucket in Figure 3. We can see that as sequences become longer, under-translation and misorder errors increase slightly. However, two phenomena in this figure raise our interests:

- 1) *More over-translation errors are observed over the ST set than other buckets;*
- 2) *Mistranslation errors show a slight trend of increasing as the sequences become shorter.*

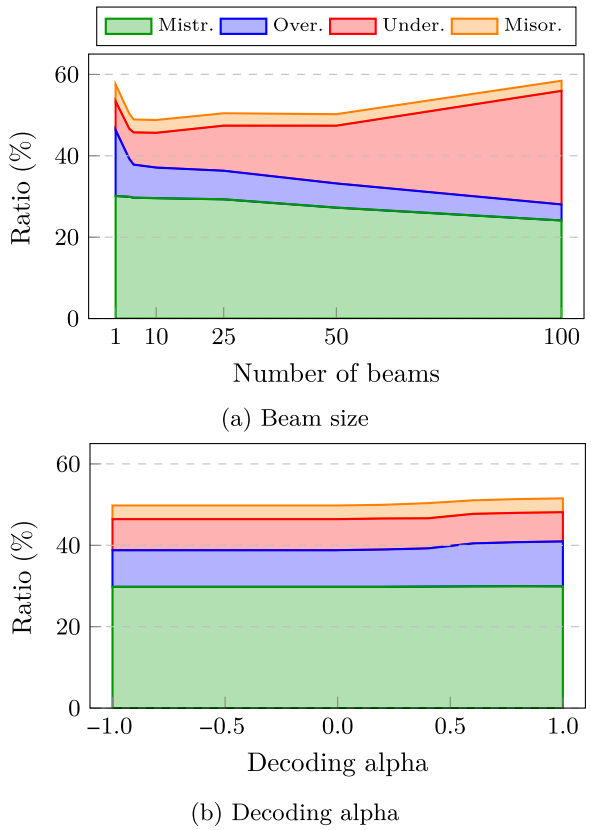


Figure 2
Comparison on different translation error types with multiple settings of beam size and decoding alpha over En⇒De ST set. For different beam sizes, we check results by setting it as 1, 4, 5, 10, 25, and 100, respectively. As to decoding alpha, it varies from −1.0 to 1.0, with an interval being 0.2. As seen, larger beam size decreases mistranslation (Mistr.) and over-translation (Over.) issues, whereas it significantly increases the under-translation (Under.) problem. Additionally, larger decoding alpha value increases the ratio of over-translation cases, while other types show marginal turbulance. Tuning both hyper-parameters merely affects the number of misorder cases (Misor.). Aside from the default setting, carefully tuning both hyper-parameters cannot reduce the summation of all kinds of translation errors. The same conclusion can be derived from Zh ⇒ En.

The former observation demonstrates that the NMT model tends to generate longer translation candidates when given STs, leading to massive over-translation errors. We think the reason lies in the distribution of data, where short sequences are rarely trained compared with examples from other length ranges (He et al. 2016). The latter observation reveals that the NMT model cannot provide adequate accurate translations over STs, unlike other buckets. We believe this is related to the insufficiency of contextual information, where the model is troubled by a limitation on the number of tokens engaged within ST examples. Translating STs is more difficult because the contextual information is not adequate to generate precise predictions.

Following these two ideas, we propose to investigate the reasons behind poor NMT performance over STs with (1) Imbalance of data distribution, and (2) Insufficiency of contextual information.

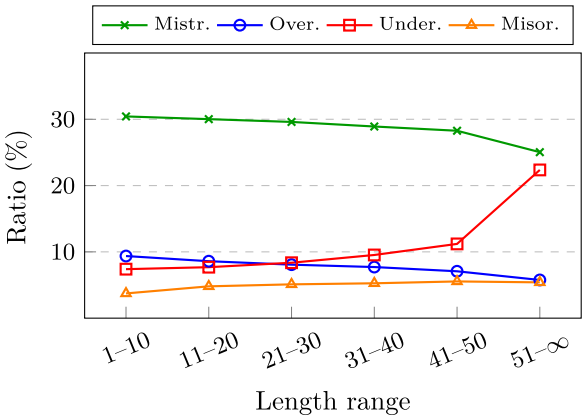


Figure 3
Translation error ratio (%) of each bucket from the En⇒De test set by NMT model. For longer sequences, the percentage of under-translation errors (Under.) varies marginally, and the model gives more misorder errors (Misor.). However, STs involve more over-translation (Over.) and mistranslation (Mistr.) errors.

5. Imbalance of Data Distribution

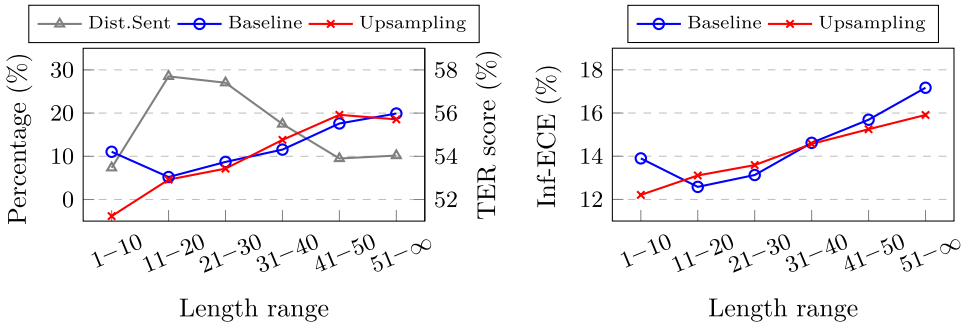
The distribution of the training dataset has a transforming influence over model inference (He et al. 2016; Shen et al. 2016; Wang et al. 2018). Generally, most of the learned positional information is relevant to the bucket, which possesses a dominant proportion over the whole training set. In this section, we investigate how the imbalance of data distribution affects NMT performance over STs.

5.1 Inference Miscalibration

To measure how critically the NMT model suffers from imbalance of data, we propose to explore the translation quality with model miscalibration (Guo et al. 2017; Müller, Kornblith, and Hinton 2019; Wang et al. 2020). Specifically, this represents the gap between model accuracy and confidence, which is also tightly aligned with exposure bias (Ranzato et al. 2016; Wang et al. 2020) in sequence learning. To quantify how a model is miscalibrated during inference, we apply the inference expected calibration error (Inf-ECE, Wang et al. 2020) which is calculated at the token level:

$$\text{Inf-ECE} = \sum_{n=1}^{N} \frac{|\mathcal{D}_n|}{N} |acc(\mathcal{D}_n) - conf(\mathcal{D}_n)| \tag{5}$$

where N is the number of partitions in total, $\mathcal{D}_n$ is the n-th partitioned bin containing the samples whose confidence scores lie in corresponding range, and $|\mathcal{D}_n|$ is the number of samples in $\mathcal{D}_n$. The accuracy score *acc* denotes the ratio of candidate tokens that are not required to be modified (i.e., those cases excluded from translation errors), and the confidence score *conf* is equivalent to the predicted probability of a candidate token (Wang et al. 2020). By averaging the differential values between accuracy and confidence scores, this gap reveals how miscalibrated the model is during inference (Müller, Kornblith, and Hinton 2019).



(a) Model performance and data distribution (b) Model Inf-ECE

Figure 4
Left: Model performance following baseline (Baseline) and the upsampling setting (Upsampling), as well as data distribution at the sentence-level (Dist.Sent). Right: Inference Expected Calibration Error (Inf-ECE, Wang et al. 2020) over En⇒De dev buckets following baseline (Baseline) and the upsampling setting (Upsampling). Lower TER score denotes better model performance, and higher Inf-ECE denotes more severe inference miscalibration. For the baseline model, it gives better performance over the bucket composed of more training examples, and shows higher inference miscalibration. Upsampling those buckets with small proportions significantly alleviates Inf-ECE, where miscalibration issue of translating short texts (STs) decreases the most. ST-NMT benefits most from upsampling. Similar trends can also be observed over the Zh⇒En dev set.

Following Wang et al. (2020), we set the number of partitions N as 10, and conduct the Inf-ECE over each bucket in Figure 4b. We can see that, as the sequences become longer, Inf-ECE decreases first, then increases until the bucket composed of the longest sequences. Interestingly, we find that the model Inf-ECE is highly related to model performance, where a higher Inf-ECE level leads to worse translation performance (TER score in Figure 4a). This confirms our hypothesis, that *the imbalanced distribution of STs leads to a rather high inference miscalibration, thus determines worse model performance when translating STs.*

5.2 Balancing Data Sampling in Training

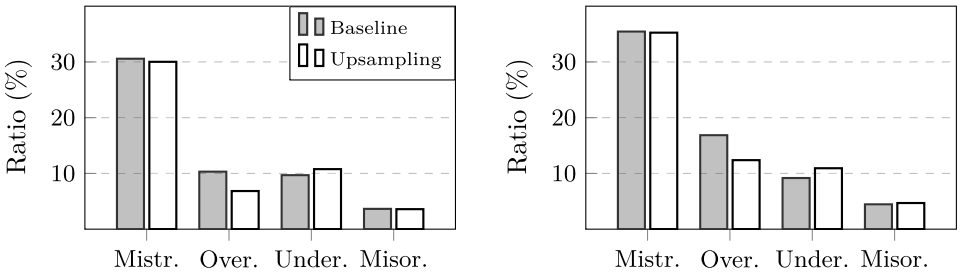
Following the conclusion above, balancing the distribution of training data is likely a promising way to alleviate inference miscalibration for NMT over STs. We thus check this method by upsampling the examples whose bucket contributes a lower proportion over the entire training set in experiments. Specifically, within each training epoch, the model is required to gain the same number of updates over different buckets of training examples (Dong, Quirk, and Lapata 2018; Schuster et al. 2019). As a consequence, examples from the same bucket that contributes less to the whole training set are repeatedly sampled for training within each epoch until it reaches the balance of sampling.[6]

As seen in Figure 4b, with balanced data sampling, the Inf-ECE values over all buckets decrease significantly, among which Inf-ECE for STs decreases the most. As

6 Although it is more reasonable to achieve this balance via counting the number of trained tokens, our token-level experiments give worse translation quality over STs. After conducting statistical tests, we find that STs contribute a lower proportion of training data at the token-level than sentence-level, thus easily leading to the over-fitting problem.

Table 3
BLEU (%) score, TER (%) score, and human evaluation (HE) over short texts. BLEU and TER scores are conducted over 5 independent runs, and HE score is conducted via the checkpoint with intermediate performance. ↑/↓: higher/lower is better. Upsampling (Up.) can increase ST translation quality.

	En⇒De		**Zh⇒En**		
	BLEU↑	TER↓	BLEU↑	TER↓	HE↑
Baseline	30.51 ± 0.17	54.17 ± 0.16	19.49 ± 0.13	65.71 ± 0.21	2.58
Up.	**32.13** ± 0.19	**51.26** ± 0.15	**20.44** ± 0.14	**63.14** ± 0.18	**3.11**



(a) En⇒De (b) Zh⇒En

Figure 5
Ratio of each translation error type over En⇒De and Zh⇒En short text (ST) set. For both En⇒De and Zh⇒En tasks, upsampling decreases the ratio of mistranslation (Mistr.) and over-translation (Over.) errors, whereas it increases the under-translation (Under.) cases. The ratio of misorder (Misor.) errors marginally fluctuates.

a consequence, NMT performance over STs is significantly improved, whereas the performances of other buckets fluctuate (Figure 4a). As shown in Table 3, the upsampling method shows significant improvements over baseline over both tasks. Additionally, we also randomly selected 100 translated candidates from models trained with both the default setting and the upsampling strategy. We collect human assessments from 10 volunteers, who are asked to provide an integer score from 1 to 4 to describe how well the translation output expresses the reference for each data item, resulting in 3.11 versus 2.58 on average, respectively. Those results verify our hypothesis about the imbalance of data distribution.

To investigate further, we calculate the ratio of each translation error type from baseline and our approaches in Figure 5. As shown, for both tasks, upsampling can mitigate the mistranslation issue, whereas it also raises more from under-translation. More importantly, the upsampling method significantly reduces over-translation cases. Integrating this with the findings in Figure 4, we thus conclude that *balancing the data distribution can relieve model miscalibration at inference, thus improving the ST translation quality on reducing over-translation cases*.

6. Insufficiency of Contextual Information

Although we see that decreasing miscalibration can reduce over-translation errors, the portion of mistranslation errors changes marginally after introducing training data

upsampling. This indicates that mistranslation cases are mainly caused by some other reasons aside from miscalibration. In fact, as the length of sentence becomes shorter, the sequence will carry fewer tokens for processing. Elaborating the semantic information of STs for translation is rather difficult compared with those examples with adequate tokens. As a consequence, such a lack of contextual information aggravates semantic disambiguation (Brunner et al. 2020; Xu et al. 2021a) and cross-lingual alignments (Cao and Xiong 2018) in the NMT model. We think such a constraint is the reason why translating STs generates more mistranslation errors than other buckets (Figure 3). In this section, we continue our analyses on the insufficiency of contextual information in ST scenarios of NMT.

6.1 Model Uncertainty

To investigate how the quantity of contextual information influences the translation quality of STs, we propose to use model uncertainty (Gal and Ghahramani 2016; Wang et al. 2019) to identify to what degree the model is uncertain about its own output. Such an uncertainty estimation is implemented using Monte Carlo Dropout Sampling (MCDS, Gal and Ghahramani 2016; Dong, Quirk, and Lapata 2018), where each sampling time is arranged after partially disabling model parameters randomly with a dropout mask (Srivastava et al. 2014). The uncertainty value represents how much the model hesitates when generating translated candidates.

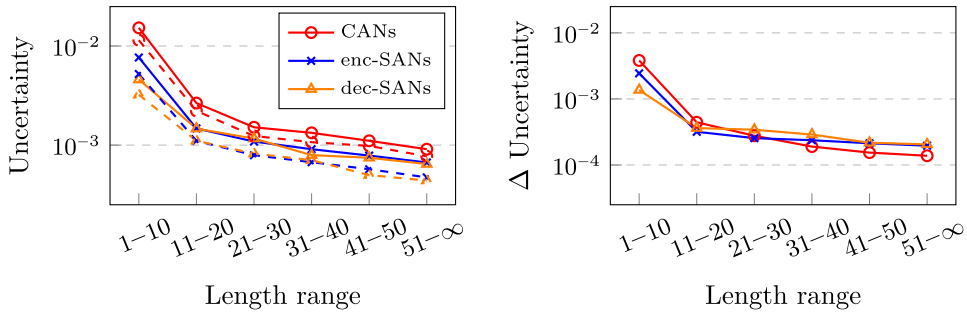
After taking a glimpse of all modules in the conventional TRANSFORMER model, it is easy to see that the attention networks are the only ones where semantic processing is arranged across positions inside the sequence. The other modules, such as positional encoding and feedforward networks, gather the representations in a position-wise manner, where the output at one specific position is only contributed to by the input at its corresponding place. We speculate that MCDS, by applying dropout over all model parameters (Wang et al. 2019; Zhou et al. 2020), cannot appropriately mimic the insufficiency of contextual information. Here we propose to strictly arrange MCDS over three types of attention networks (Section 3.1): enc-SAN, dec-SAN, and CAN, respectively. Specifically, the MCDS is applied over the attention weights, thus it can check how the model is affected by the insufficiency of contextual information. Aside from Equation 2, the output of the attention network is then redefined as follows:

$$\mathbf{u} \sim \text{Bernoulli}(p) \in \{0, 1\}^{L_2} \tag{6}$$

$$\tilde{\mathbf{A}}_{h,l_1,\cdot} = \text{dropout}(\mathbf{A}_{h,l_1,\cdot}, \mathbf{u}) \in \mathcal{R}^{L_2} \tag{7}$$

$$\tilde{\mathbf{H}}_{h,l_1,\cdot} = \sum_{k=1}^{L_2} \tilde{\mathbf{A}}_{h,l_1,k} \mathbf{V}_{h,k,\cdot} \in \mathcal{R}^{d_h} \tag{8}$$

where $\mathbf{u} \in \mathcal{R}^I$ denotes the dropout mask that follows the Bernoulli distribution with probability p, $\tilde{\alpha} \in \mathcal{R}^I$ contains the attention weights after applying dropout, and p is the ratio of disabled neurons (Srivastava et al. 2014). The dropout masks among multiple MCDS operations are randomly generated for each tensor, resulting in the fact that the sampled probabilities of target sequence P($\mathbf{T}$) fluctuates. For these sets of probabilities that show a higher level of turbulence, the model tends to become more uncertain over corresponding examples (Dong, Quirk, and Lapata 2018; Xiao and Wang 2019; Zhou et al. 2020). In this test, we simply assign the model uncertainty over the corresponding



(a) Uncertainty value (b) Uncertainty Gap

Figure 6
Model uncertainty values of NMT models with/without translation memory (up) and their differences (down) over each bucket from the En⇒De dev set. On the left, solid and dashed lines denote the uncertainty values of baseline model and translation memory (TM) model, respectively. The uncertainty value is consistently larger and the sequence lengths are shorter. As can be seen, short texts (STs) are severely hindered by insufficiency of contextual information. After applying TMs in training, the uncertainty values all decrease, with that of the ST set decreasing most. Similar trends can be observed over Zh⇒En dev buckets.

sequence with the variance of sampled probabilities by MCDS. To avoid underflow issues, we use negative log-likelihood probabilities (Wang et al. 2019; Wan et al. 2020b) instead. As well as giving a more fair comparison across different buckets of various lengths, we use the sequence length of each sentence J to normalize the values:

$$u_m = \sigma(-\frac{1}{J}\sum_{j=1}^{J}\log \mathrm{P}(\mathbf{T}_j|\mathbf{T}_{<j}, \mathbf{C}, \theta, m, p))_{k=1}^{K} \tag{9}$$

where $\sigma(\cdot)$ denotes the variance of input tensors, $m \in \{\text{enc-SAN, dec-SAN, CAN}\}$ denotes the module set applied with MCDS, and K is the number of sampling times. Following this paradigm, a larger u_m value denotes a higher level of model uncertainty over the corresponding sequence $\mathbf{T}$. This design results in discrimination showing that a higher model uncertainty value indicates a more severe impact of the insufficiency of contextual information on translation quality.

Using this design, we calculate the averaged model uncertainty value over examples from each length bucket. Following a previous study (Xiao and Wang 2019), we set the dropout ratio p as 0.1. Additionally, although a higher number for sampling K denotes a more accurate estimation of variance, it requires more time to accomplish the computation. After conducting preliminary experiments and carefully tuning, we determine to collect model output probabilities by $K = 5$ times, as setting this value larger does not result in an obvious change in variance values.

Results are shown in Figure 6. As can be seen, the uncertainty values derived by randomly disabling CAN attention weights are the highest across all example buckets. We believe this highlights the intuition that the semantic transformation from source to target is the most crucial among NMT model functionalities. Also, shorter sequences have a higher degree of uncertainty, indicating that they suffer heavily from insufficient contextual information. We thus conclude that *comparing other examples containing more*

Table 4
BLEU (%) score, TER (%) score, and human evaluation (HE) over short texts. ↑/↓: higher/lower is better. BLEU and TER scores are conducted over 5 independent runs, and HE score is conducted via the checkpoint with intermediate performance. Translation memory (TM) improves ST translation quality.

	En⇒De		**Zh⇒En**		
	BLEU↑	TER↓	BLEU↑	TER↓	HE↑
Baseline	30.51 ± 0.17	54.17 ± 0.16	19.49 ± 0.13	65.71 ± 0.21	2.58
TM	**31.41** ± 0.20	**53.24** ± 0.16	**20.16** ± 0.20	**63.76** ± 0.24	**2.92**

tokens, translating STs with the NMT model is more easily affected by the inadequacy of contextual information.

6.2 Complementing Contextual Information

As the NMT for ST is highly troubled by the lack of contextual information, we believe that incorporating additional information for ST translation is potentially helpful. Translation memory (TM), which addresses our aim, has been shown to be an efficient way to relieve the insufficiency of contextual information (Cao and Xiong 2018; Eriguchi, Rarrick, and Matsushita 2019; Kim, Tran, and Ney 2019). Specifically, TM is derived from extra examples that are similar to the inputted sentence pair. It augments the integrity of contextual information by providing extra reliable information from TM, thus can ease NMT model learning (Cao and Xiong 2018).

To check whether TM helps NMT for ST, we conduct experiments by incorporating TM into NMT model training. Empirically, following recent studies (Cao and Xiong 2018; Kim, Tran, and Ney 2019), we first derive the TM of each training example by calculating the similarity of two sequences based on token-level Levenshtein distance:

$$\mathrm{Sim}(\mathbf{S}, \hat{\mathbf{S}}) = 1 - \frac{Levenshtein(\mathbf{S}, \hat{\mathbf{S}})}{max(J, \hat{J})} \tag{10}$$

where $\hat{\mathbf{S}}$ is the example different from input sequence $\mathbf{S}$ at source side, and $\hat{J}$ is the length of $\hat{\mathbf{S}}$. Here the Levenshtein distance measurement $Levenshtein(\cdot, \cdot)$ gives the number of required edit moves when transforming one sequence to another at the token level. Additionally, to normalize the difference between $\mathbf{S}$ and $\hat{\mathbf{S}}$, we use the maximum value of sequence lengths to normalize the score. Therefore, for those sequence pairs whose lexical choices and token orders have little overlap, the normalized Levenshtein distance is large, thus giving a lower similarity score. When implementing such an idea, for simpler and faster TM searching, we restrict the TM candidates inside the same document where $\mathbf{S}$ is located.

As seen in Figure 6, after introducing TM into NMT training, the model uncertainty value of short sequences also downgrades. Specifically, as the sequences become shorter, the drop of model uncertainty becomes rather larger. This indicates that complementing contextual information can decrease the level of model uncertainty, which benefits NMT for ST the most. As seen in Table 4, translation errors over STs drop significantly with given TM, indicating that the additional contextual information is constructive for ST

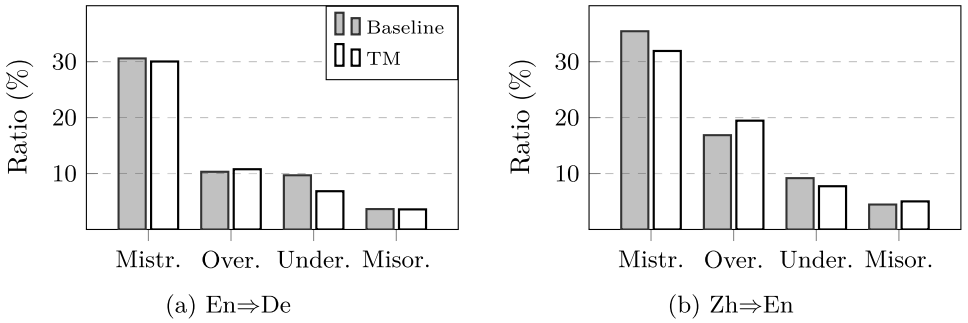


Figure 7
Ratio of each translation error type over En⇒De and Zh⇒En short text (ST) set. Translation memory (TM) significantly mitigates mistranslation (Mistr.) and under-translation (Under.). However, the model is hampered by more over-translation (Over.) cases. The ratio of misorder (Misor.) errors marginally fluctuates.

translation scenarios. This demonstrates that the sufficiency of contextual information is beneficial to translation quality.

To further investigate which part and to what degree TM improves NMT for ST translation performance, we calculate the ratio of all translation error types of the outputs from the NMT model trained with TM. As shown in Figure 7, TM significantly reduces the mistranslation cases, as well as under-translation errors. However, it creates more over-translation issues compared to baseline. These results demonstrate that *providing contextual information can significantly reduce model uncertainty over ST, and the NMT model gains significant improvement of translation quality over STs when incorporating TM.*

7. Conclusion

In this study, we investigated the challenges of neural machine translation over short texts. To obtain more fine-grained analyses, we categorize translation errors into four types, namely, mistranslation, over-translation, under-translation, and misorder (Section 4.2). Our contributions are as follows:

- Based on empirical analyses, we bring out two challenges in NMT for ST, which may encourage the NLP community to pay more attention to such scenarios (Section 4.5);
- By identifying inference miscalibration, we confirm that imbalanced data distribution leads to higher miscalibration, thus raising more over-translation errors for NMT for ST (Section 5.1);
- By quantifying the model uncertainty, we verify that the lack of contextual information leads to higher uncertainty, which results in more mistranslation errors for NMT for ST (Section 6.1);
- Based on the above findings, we also suggest two potential directions, namely, balancing the sampling of training data (Section 5.2) and complementing contextual information (Section 6.2), which can alleviate the over-translation and mistranslation issue in NMT for ST, respectively.

We believe that some existing NMT research topics, for example, title translation and entity translation, are also included in the scenario of translating STs. In this research, we conducted experiments on the news domain, and an ST set for analytic experiments mainly composed of titles and entities. Considering that many scenarios heavily rely on understanding and processing titles/entities (e.g., Web site search, information retrieval, keyword search, and cross-lingual search), we believe that investigating the shortcomings of title/entity translation is another interesting research topic. We welcome more reports on this topic in the future.

Additionally, as our work mainly explores the challenges of NMT for ST, following the findings of this research, we believe that some approaches related to data distribution imbalance and contextual information insufficiency may also help improve NMT for ST translation quality. These approaches are mainly engaged in potential research directions such as data augmentation (Sugiyama and Yoshinaga 2019; Li and Specia 2019; Wan et al. 2020a), localness modeling (Yang et al. 2018; Shaw, Uszkoreit, and Vaswani 2018; Wu et al. 2019; Xu et al. 2019), and contextual representation enhancement (Voita, Sennrich, and Titov 2019; Maruf, Martins, and Haffari 2019; Yang et al. 2019; Zhang et al. 2021; Xu et al. 2021b). We leave these promising approaches as potential solutions to ameliorate NMT for ST performance, as well as other aspects of translation quality estimation for NMT for ST interpretability, to future work.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund, Macau SAR (grant no. 0101/2019/A2), and the Multi-year Research Grant from the University of Macau (grant no. MYRG2020-00054-FST), National Key Research and Development Program of China (No. 2018YFB1403202), and Alibaba Group through Alibaba Research Intern Program.

The authors would like to thank all the anonymous reviewers and the chief editor for their insightful comments. Special thanks to Xingzhang Ren and Jun Xie, who offered valuable comments during paper revision.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Banar, Nikolay, Walter Daelemans, and Mike Kestemont. 2020. Neural machine translation of artwork titles using iconclass codes. In *LaTeCH-CLfL@COLING*, pages 42–51.

Bengio, Yoshua, Jerome Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*, pages 41–48. https://doi.org/10.1145/1553374.1553380

Bi, Tianchi, Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Constraint translation candidates: A bridge between neural query translation and cross-lingual information retrieval. In *eCom@SIGIR*.

Brunner, Gino, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *ICLR*.

Cao, Qian and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In *EMNLP*, pages 3042–3047. https://doi.org/10.18653/v1/D18-1340

Chen, Lu, Yanbin Zhao, Boer Lyu, Lesheng Jin, Zhi Chen, Su Zhu, and Kai Yu. 2020. Neural graph matching networks for Chinese short text matching. In *ACL*, pages 6152–6158. https://doi.org/10.18653/v1/2020.acl-main.547

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST@EMNLP*, pages 103–111. https://doi.org/10.3115/v1/W14-4012

Darwish, Ibrahim and Bilal Sayaheen. 2019. Manipulating titles in translation. *Journal of Educational and Social Research*, 9(3):239. https://doi.org/10.2478/jesr-2019-0042

Dong, Li, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural

semantic parsing. In *ACL*, pages 743–753. https://doi.org/10.18653/v1/P18-1069

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211. https://doi.org/10.1207/s15516709cog1402_1

Eriguchi, Akiko, Spencer Rarrick, and Hitokazu Matsushita. 2019. Combining translation memory with neural machine translation. In *WAT@EMNLP-IJCNLP*, pages 123–130. https://doi.org/10.18653/v1/D19-5214

Etchegoyhen, Thierry and Harritxu Gete. 2020. To case or not to case: Evaluating casing methods for neural machine translation. In *LREC*, pages 3752–3760.

Gal, Yarin and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059.

Ghazvininejad, Marjan, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP-IJCNLP*, pages 6111–6120. https://doi.org/10.18653/v1/D19-1633

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*, pages 1321–1330.

Hasler, Eva, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. 2017. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45:221–235. https://doi.org/10.1016/j.csl.2016.12.001

He, Wei, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with SMT features. In *AAAI*, pages 151–157.

Hendrycks, Dan, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, pages 10477–10486.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Huang, Guangpu, Arseniy Gorin, Jean-Luc Gauvain, and Lori Lamel. 2016. Machine translation based data augmentation for Cantonese keyword spotting. In *ICASSP*, pages 6020–6024. https://doi.org/10.1109/ICASSP.2016.7472833

Jiang, Lu, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. 2015. Self-paced curriculum learning. In *AAAI*, pages 2694–2700.

Jiang, Long, Ming Zhou, Lee-Feng Chien, and Cheng Niu. 2007. Named entity translation with web mining and transliteration. In *IJCAI*, pages 1629–1634.

Karakanta, Alina, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189. https://doi.org/10.1007/s10590-017-9203-5

Kim, Yunsu, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *DiscoMT@EMNLP*, pages 24–34. https://doi.org/10.18653/v1/D19-6503

Kocmi, Tom and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *RANLP*, pages 379–386. https://doi.org/10.26615/978-954-452-049-6_050

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *NMT@ACL*, pages 28–39. https://doi.org/10.18653/v1/W17-3204

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*. https://doi.org/10.21236/ADA461156

Kreutzer, Julia, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? In *NAACL-HLT (Industry Papers)*, pages 92–105. https://doi.org/10.18653/v1/N18-3012

Le, An Nguyen, Ander Martinez, and Yuji Matsumoto. 2017. Improving sequence to sequence neural machine translation by utilizing syntactic dependency information. In *IJCNLP*, pages 21–29.

Li, Zhenhao and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *W-NUT@EMNLP*, pages 328–336. https://doi.org/10.18653/v1/D19-5543

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Zhongxin, Xin Xia, Ahmed E. Hassan, David Lo, Zhenchang Xing, and Xinyu

Wang. 2018. Neural-machine-translation-based commit message generation: How far are we? In *ASE*, pages 373–384. https://doi.org/10.1145/3238147.3238190

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421. https://doi.org/10.18653/v1/D15-1166

Lyu, Boer, Lu Chen, Su Zhu, and Kai Yu. 2021. LET: Linguistic knowledge enhanced graph transformer for Chinese short text matching. In *AAAI*, pages 13498–13506.

Maruf, Sameen, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *NAACL-HLT*, pages 3092–3102. https://doi.org/10.18653/v1/N19-1313

Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *NeurIPS*, pages 4696–4705.

Murray, Kenton and David Chiang. 2018. Correcting length bias in neural machine translation. In *WMT*, pages 212–223. https://doi.org/10.18653/v1/W18-6322

Nakov, Preslav, Francisco Guzmán, and Stephan Vogel. 2012. Optimizing for sentence-level BLEU+1 yields short translations. In *COLING*, pages 1979–1994.

Neishi, Masato and Naoki Yoshinaga. 2019. On the relation between position information and sentence length in neural machine translation. In *CoNLL*, pages 328–338. https://doi.org/10.18653/v1/K19-1031

Nguyen, Xuan-Phi, Shafiq R. Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: An elegant strategy for neural machine translation. In *NeurIPS*, pages 10018–10029.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167. https://doi.org/10.3115/1075096.1075117

Ott, Myle, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *WMT*, pages 1–9. https://doi.org/10.18653/v1/W18-6301

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318. https://doi.org/10.3115/1073083.1073135

Platanios, Emmanouil Antonios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *NAACL:HLT*, pages 1162–1172. https://doi.org/10.18653/v1/N19-1119

Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Saleh, Shadi and Pavel Pecina. 2020. Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In *ACL*, pages 6849–6860. https://doi.org/10.18653/v1/2020.acl-main.613

Schuster, Sebastian, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *NAACL-HLT*, pages 3795–3805. https://doi.org/10.18653/v1/N19-1380

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162

Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *NAACL-HLT*, pages 464–468. https://doi.org/10.18653/v1/N18-2074

Shen, Shiqi, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *ACL*, pages 1683–1692. https://doi.org/10.18653/v1/P16-1159

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *ATMA*, pages 223–231.

Song, Hyun-Je, A.-Yeong Kim, and Seong-Bae Park. 2017. Translation of natural language query into keyword query using a RNN encoder-decoder. In *SIGIR*, pages 965–968. https://doi.org/10.1145/3077136.3080691

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Sugiyama, Amane and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *DiscoMT@EMNLP*, pages 35–44. https://doi.org/10.18653/v1/D19-6504

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Toral, Antonio and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *EACL*, pages 1063–1073. https://doi.org/10.18653/v1/E17-1100

Tu, Zhaopeng, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *ACL*, pages 76–85. https://doi.org/10.18653/v1/P16-1008

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Voita, Elena, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *ACL*, pages 1198–1212. https://doi.org/10.18653/v1/P19-1116

Wan, Yu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Haihua Du, and Ben C. H. Ao. 2020a. Unsupervised neural dialect translation with commonality and diversity modeling. In *AAAI*, pages 9130–9137. https://doi.org/10.1609/aaai.v34i05.6448

Wan, Yu, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020b. Self-paced learning for neural machine translation. In *EMNLP*, pages 1074–1080. https://doi.org/10.18653/v1/2020.emnlp-main.80

Wang, Longyue, Jinhua Du, Liangyou Li, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Semantics-enhanced task-oriented dialogue translation: A case study on hotel booking. In *IJCNLP*, pages 33–36.

Wang, Shuo, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *EMNLP-IJCNLP*, pages 791–802. https://doi.org/10.18653/v1/D19-1073

Wang, Shuo, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *ACL*, pages 3070–3079. https://doi.org/10.18653/v1/2020.acl-main.278

Wang, Xinyi, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: An efficient data augmentation algorithm for neural machine translation. In *EMNLP*, pages 856–861. https://doi.org/10.18653/v1/D18-1100

Wu, Felix, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *ICLR*.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiao, Yijun and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *AAAI*, pages 7322–7329. https://doi.org/10.1609/aaai.v33i01.33017322

Xu, Linlong, Baosong Yang, Xiaoyu Lv, Tianchi Bi, Dayiheng Liu, and Haibo Zhang. 2021a. Leveraging advantages of interactive and non-interactive models for vector-based cross-lingual information retrieval. *arXiv preprint arXiv:2111.01992*.

Xu, Mingzhou, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. 2021b. Document graph for neural machine translation. In *EMNLP*, pages 8435–8448. https://doi.org/10.18653/v1/2021.emnlp-main.663

Xu, Mingzhou, Derek F. Wong, Baosong Yang, Yue Zhang, and Lidia S. Chao. 2019. Leveraging local and global patterns for self-attention networks. In *ACL*, pages 3069–3075. https://doi.org/10.18653/v1/P19-1295

Yang, Baosong, Jian Li, Derek F. Wong, Lidia S. Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *AAAI*, pages 387–394. https://doi.org/10.1609/aaai.v33i01.3301387

Yang, Baosong, Zhaopeng Tu, Derek F. Wong, Fandong Meng, Lidia S. Chao, and Tong Zhang. 2018. Modeling localness for self-attention networks. In *EMNLP*, pages 4449–4458. https://doi.org/10.18653/v1/D18-1475

Yang, Yilin, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *EMNLP*, pages 3054–3059. https://doi.org/10.18653/v1/D18-1342

Yao, Liang, Baosong Yang, Haibo Zhang, Boxing Chen, and Weihua Luo. 2020a. Domain transfer based data augmentation for neural query translation. In *COLING*, pages 4521–4533. https://doi.org/10.18653/v1/2020.coling-main.399

Yao, Liang, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020b. Exploiting neural query translation into cross lingual information retrieval. In *eCom@SIGIR*.

Zhang, Biao, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *EMNLP*, pages 521–530. https://doi.org/10.18653/v1/D16-1050

Zhang, Long, Tong Zhang, Haibo Zhang, Baosong Yang, Wei Ye, and Shikun Zhang. 2021. Multi-hop transformer for document-level machine translation. In *NAACL: HLT*, pages 3953–3963. https://doi.org/10.18653/v1/2021.naacl-main.309

Zhang, Xuan, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J. Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*.

Zhao, Yang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graphs enhanced neural machine translation. In *IJCAI*, pages 4039–4045. https://doi.org/10.24963/ijcai.2020/559

Zhou, Yikai, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *ACL*, pages 6934–6944. https://doi.org/10.18653/v1/2020.acl-main.620