Towards Enhancing Faithfulness for Neural Machine Translation

Rongxiang Weng[1], **Heng Yu**[1], **Xiangpeng Wei**[2,3]*, **Weihua Luo**[1]
[1]Machine Intelligence Technology Lab, Alibaba Group, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, China
{wengrx,yuheng.yh}@alibaba-inc.com
weixiangpeng@iie.ac.cn
weihua.luowh@alibaba-inc.com

Abstract

Neural machine translation (NMT) has achieved great success due to the ability to generate high-quality sentences. Compared with human translations, one of the drawbacks of current NMT is that translations are not usually faithful to the input, e.g., omitting information or generating unrelated fragments, which inevitably decreases the overall quality, especially for human readers. In this paper, we propose a novel training strategy with a multi-task learning paradigm to build a faithfulness enhanced NMT model (named FENMT). During the NMT training process, we sample a subset from the training set and translate them to get fragments that have been mistranslated. Afterward, the proposed multi-task learning paradigm is employed on both encoder and decoder to guide NMT to correctly translate these fragments. Both automatic and human evaluations verify that our FENMT could improve translation quality by effectively reducing unfaithful translations.

1 Introduction

Neural machine translation (NMT) based on the *encoder-decoder framework* (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2014; Luong et al., 2015b) has obtained state-of-the-art performance on many language pairs (Wu et al., 2019; Wei et al., 2020). Various neural architectures have been explored for modeling NMT under this framework, such as recurrent neural network (RNN) (Bahdanau et al., 2014; Luong et al., 2015b, RNNSearch), convolutional neural network (CNN) (Gehring et al., 2016, ConvS2S) and self-attention network (Vaswani et al., 2017, Transformer). Compared with human translations or traditional statistical machine translation (SMT) (Koehn et al., 2007b; Chiang, 2007), NMT can generate high-quality sentences that are very close to natural language. However, it usually appears some parts (e.g., phrase) from input sentences cannot be correctly translated, leading to that the translation is inadequate for direct using in some scenarios. This phenomenon appeals that enhancing the faithfulness of translations is an important aspect for further improving NMT.

We summarize three possible causes for the unfaithfulness problem based on the encoder-decoder framework: 1). *Some parts from input sentences are hard to encode, and thus cannot be translated correctly.* 2). *The decoder cannot fetch the correct contextual representation from the encoder.* 3).*The dominant language model of NMT prompts the decoder generates common words to make sure outputs are fluent.* Several recent studies are proposed following one of the above perspectives and have achieved considerable effects. Zheng et al. (2019) proposed to divide the encoder output into past and future parts to fine-grained modeling contextual representation. Feng et al. (2020) proposed a faithfulness part to optimize the contextual representation before feeding into the decoder. Kong et al. (2019) proposed to use a coverage difference ratio metric as a reward to train NMT.

In this paper, we propose a novel training strategy with a multi-task learning paradigm, taking into account the use of real translations for building a faithfulness enhanced NMT (named FENMT). Firstly, we align source and target sentences in the training set. Then, at each training epoch, we sample a subset from the training set and translate source sentences by the NMT in the this set. For convenience, we simply define a mistranslated fragment is a continues segment from a target sentence which does not appear in the translation. So, we can collect mistranslated fragments by comparing the translation and reference, and get the corresponding source words by the alignment relationship.

*Work done during the internship at Alibaba Group.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2675–2684,
November 16–20, 2020.

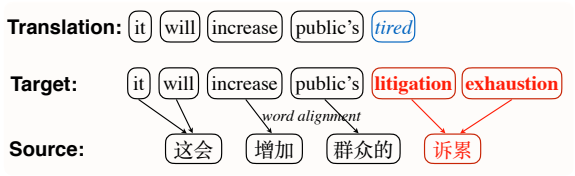

Figure 1: The case of collecting mistranslated fragments. "litigation exhaustion" is the mistranslated fragment and "累诉" is the corresponding source word.

After that, our multi-task learning paradigm (MTL) is incorporated into the training process to learn to correctly translate these mistranslated fragments. To make the most of the collected mistranslated fragments, the proposed MTL method considers all sides of the above hypotheses.

Specifically, we employ a *masked language model task* (Devlin et al., 2018) on the encoder side to infer the input words didn't be correctly translated. This task can enhance the ability of modeling the whole input sentence and give the decoder accurate and complete representations. On the decoder side, we use a *word alignment task* to improve the alignment accuracy of the *encoder-decoder attention* (or *cross-attention*) to help the decoder to capture correct contextual representation. Furthermore, along with the NMT objective, an auxiliary *max-margin objective* based on contrastive learning is introduced in all decoding time-steps. The goal of this task is to avert the tendency of translating frequent but unrelated words.

We implement the proposed approach based on Transformer (Vaswani et al., 2017) and evaluate it on WMT14 English to German (En→De), WMT17 Chinese to English (Zh→En) and WMT16 English to Romania (En→Ro) machine translation tasks. Both automatic and human evaluations show that the proposed FENMT could substantially improve the overall quality and faithfulness of translations.

2 The Proposed Approach

We will introduce the whole procedure of the proposed FENMT model based on the advanced Transformer (Vaswani et al., 2017). We firstly show the details of how to collect mistranslated fragments and the multi-task learning paradigm at section 2.1 and 2.2, respectively. Then, the overall training strategy of our approach is represented at section 2.3.

2.1 Collecting Mistranslated Fragments

Given a parallel training set $\mathcal{B}$, we achieve the alignment matrix set $\mathcal{A}$ through a word alignment model trained by the parallel training set, and get the phrase table $\mathcal{P}$ according to the word alignment (Koehn et al., 2003).

At the tth training epoch of NMT, we sample a subset $\mathcal{B}_t^{\mathrm{S}}$ from the $\mathcal{B}$. Given a sentence pair $\{\mathbf{x}, \mathbf{y}\}$ from the $\mathcal{B}_t^{\mathrm{S}}$, where $\mathbf{x}$ is the source sequence $(x_1, \cdots, x_i, \cdots, x_I)$ and $\mathbf{y}$ is the target sequence $(y_1, \cdots, y_j, \cdots, y_J)$, I and J are the length of $\mathbf{x}$ and $\mathbf{y}$, respectively. The alignment matrix $\mathbf{A} \in \mathbb{R}^{J \times I}$ of $\{\mathbf{x}, \mathbf{y}\}$ can be obtained from $\mathcal{A}$, in which $a_{j,i} = 1$ means y_j aligns x_i. We then translate the source sentence by $\hat{\mathbf{y}} = f_{\theta_{t-1}}(\mathbf{x})$, where $f_{\theta_{t-1}}(\cdot)$ is the NMT model, which parameters are θ and has been trained after $t-1$ epochs. $\hat{\mathbf{y}}$ is composed of $(\hat{y}_1, \cdots, \hat{y}_k, \cdots, \hat{y}_K)$, K is the sentence length.

We define that a fragment in $\mathbf{y}$ is mistranslated when it does not appear in $\hat{\mathbf{y}}$ but is contained in $\mathcal{P}$. Subsequently, we randomly sample consecutive parts from $\mathbf{y}$ included in $\mathcal{P}$ and compare them with $\hat{\mathbf{y}}$ to achieve mistranslated fragments. We denote a subsequence $\mathbf{y}^T$ of $\mathbf{y}$ containing all words mistranslated, $y_{t,j}$ is the tth word of $\mathbf{y}^T$ whose position in the $\mathbf{y}$ is j. Afterward, we can get the aligned source words of $\mathbf{y}^T$ by using the alignment relationship. For a word $y_{t,j}$, we collect source words when $a_{j,\cdot} = 1$. We denote the sequence having all aligned source words as $\mathbf{x}^M$, in which $x_{m,i}$ is the mth word of $\mathbf{x}^M$ whose position in the $\mathbf{x}$ is i. A shortly case is shown in Figure 1.

2.2 Multi-task Learning Paradigm

Masked language model task for the encoder. The first hypothesis mentioned above is that *the encoder cannot model mistranslated parts well*, which leads to the subsequent module cannot translate them correctly. Here, we introduce a *masked language model task* (MLM) to further model these source words. Specifically, before feeding into the decoder, we ask for the source representation predicts mistranslated words which are masked at the input sentence (see Figure 2).[1]

Formally, given the input sentence $\mathbf{x}$ from $\mathcal{B}_t^{\mathrm{S}}$ and the mistranslated subsequence $\mathbf{x}^M$. We define a sequence $\mathbf{x}^R$, which likes $\mathbf{x}$ but the words in the $\mathbf{x}^M$ will be replaced as a special <MASK> token with the probability of 80%, and as a random word

[1] We also implement the MLM task though randomly sampling tokens in sentences, but it doesn't work well.

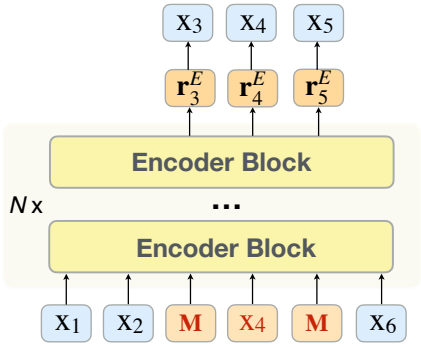

Figure 2: The overview of the masked language model task. M means the masked source word.

or keep unchanged with the probability of 10% individually. This procedure is the same as Devlin et al. (2018). The encoder with the MLM task maximizes the conditional probability defined as:

$$P(x_{m,i}|\mathbf{x}^R) = \text{softmax}(\text{FFN}(\mathbf{r}_i^E)), \quad (1)$$

$$\mathbf{R}^E = \text{Encoder}(\mathbf{x}^R), \quad (2)$$

where $\mathbf{R}^E \in \mathbb{R}^{l_{\text{input}} \times d_{\text{model}}}$ is the output hidden states of the encoder, and $\mathbf{r}_i^E \in \mathbb{R}^{d_{\text{model}}}$ is ith hidden state of $\mathbf{R}^E$. l_{input} is the length of the input sentence and d_{model} is the dimension of hidden state. Finally, the objective of the masked language model is

$$\mathcal{L}_{\text{M}} = -\mathbb{E}_{\mathbf{x}^R \in \mathcal{B}_t^{\text{S}}}[\mathbb{E}_{x_{m,i} \in \mathbf{x}^M}[\log P(x_{m,i}|\mathbf{x}^R)]]. \quad (3)$$

Word alignment task for the cross attention. After getting a better source contextual representation, i.e., the $\mathbf{R}^E$, *whether the decoder can get the correct representation for each output word* is another factor determining translation faithfulness. The cross-attention is the single connection between the encoder and decoder. A natural intuition is that improving the accuracy of cross-attention is helpful for getting faithful translations. Thus, we introduce a word alignment task for the cross-attention here (see Figure 3).

Specifically, given the target sentence $\mathbf{y}$, we define the cross-attention weight matrix as $\mathbf{C} \in \mathbb{R}^{J \times I}$, the vector $\mathbf{c}_j$ from $\mathbf{C}$ is the weight of jth decoder hidden state to the encoder representation. We then define the alignment label as $\mathbf{B} \in \mathbb{R}^{J \times I}$. Given the word y_j in the $\mathbf{y}^T$, the corresponding alignment label vector $\mathbf{b}_j$ is computed by:

$$\mathbf{b}_j = \text{softmax}(\mathbf{a}_j), \quad (4)$$

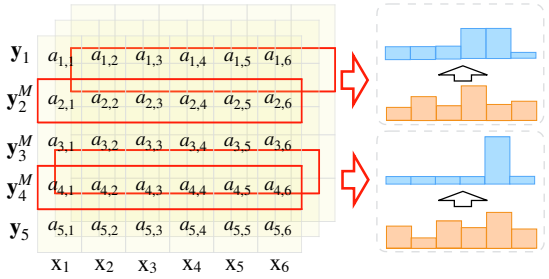

Figure 3: The overview of the word alignment objective. The word has superscript M is mistranslated.

where $\mathbf{a}_j$ is from the alignment matrix $\mathbf{A}$. Note that when using subword (Sennrich et al., 2015; Devlin et al., 2018) as input, alignment probability will be divided into the corresponding tokens equally (e.g., if a word is divided into two tokens, the probability for each token is 0.5).

Generally, the decoder has N block and the cross attention from each block has H heads (Vaswani et al., 2017). We randomly choose two heads at each blocks to employ the word alignment objective. We define the selected attention weight matrix set as $\mathcal{C} = \{\mathbf{C}_1, \cdots, \mathbf{C}_k, \cdots, \mathbf{C}_K\}$, where $K = 2 * N$. Then, the word alignment objective is

$$\mathcal{L}_{\text{A}} = -\mathbb{E}_{\mathbf{C}_k \in \mathcal{C}}[\mathbb{E}_{\mathbf{b}_j \in \mathbf{B}, \mathbf{c}_j \in \mathbf{C}_k}[\mathbf{b}_j \log \mathbf{c}_j]]. \quad (5)$$

This objective is used to guide the cross-attention to capture correct contextual information rather than only learn the word alignment information. So we only employing it on parts of attention head to avoid "*overfitting*" to the alignment task.

Max-margin task for the decoder. Empirically, the language model in current NMT is more stronger than the translation model, so the NMT model tends to translate common words even unrelated to the source sentence (Kong et al., 2019). Only using cross-entropy objective isn't enough to keep translations faithful. Here, we introduce a *max-margin objective* based on contrastive learning to suppress the tendency of NMT to generate common but unfaithful words.

Specifically, given the target sentence $\mathbf{y}$ and the translation $\hat{\mathbf{y}}$, the max-margin loss is defined as $\mathcal{L}_{\text{C}} = \sum_{j=1}^{J} \mathcal{L}_j$, where $\mathcal{L}_j$ is computed by

$$\mathcal{L}_j = \begin{cases} \max(0, mg - P(y_j|y_{1:j}, \mathbf{x}) \\ \quad + P(\hat{y}_j)|\hat{y}_{1:j}, \mathbf{x})) & , y_{t,j} \in \mathbf{y}^T \\ 0 & , y_{t,j} \notin \mathbf{y}^T \end{cases}$$

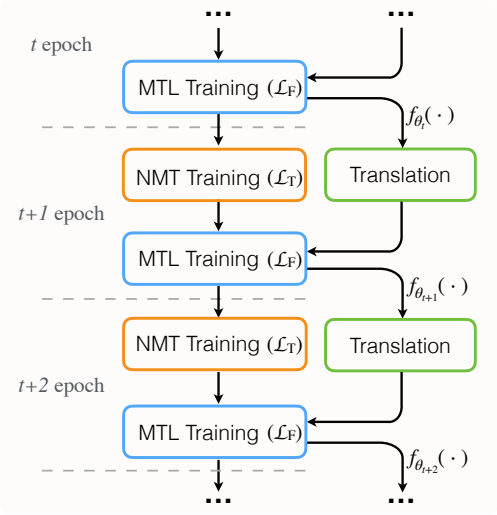

Figure 4: Overview of the training strategy of the proposed FENMT. MTL: multi-task learning.

where the mg is the margin, we empirically set to 0.2. The cross-entropy objective with this objective can prompt the decoder to translate fluent and faithful sentences.

2.3 The Overall Training Strategy

The standard NMT training objective is to minimize the negative log-likelihood by:

$$\mathcal{L}_{\mathrm{T}} = -\mathbb{E}_{\{\mathbf{x},\mathbf{y}\}\in\mathcal{B}} \log P(\mathbf{y}|\mathbf{x}). \quad (6)$$

And the final training objective of our proposed approach is:

$$\mathcal{L}_{\mathrm{F}} = \mathcal{L}_{\mathrm{T}} + \alpha \cdot \mathcal{L}_{\mathrm{M}} + \beta \cdot \mathcal{L}_{\mathrm{A}} + \gamma \cdot \mathcal{L}_{\mathrm{C}}, \quad (7)$$

where α, β and γ are used to balance the preference among the external losses, which are empirically set to 0.3 individually. Note that due to the different inputs, $\mathcal{L}_{\mathrm{M}}$ should be computed separately.

The training strategy as follows: at the tth NMT training epoch, we are going to sample part of the sentences from the training set, the sampling ratio is computed by:

$$ratio = \max(d^{(t-1)} * 20\%, 5\%), \quad (8)$$

where d is the decay rate, we set as 0.9 here. To avoid decreasing training efficiency, the sampled data will be translated by $f_{\theta_t}(\cdot)$ at the tth epoch and used at the $t+1$th epoch. And the first epoch will not use this method as a warm-up.

The overview of the training strategy is shown in Figure 4 The NMT will begin to translate sampled sentences at the end of the tth epoch, which is synchronous with the training process. Then, when both of the training process and translation process are finished, the multi-task learning paradigm will be employed to continue train the NMT model.

3 Experiment

3.1 Implementation Detail

We conduct experiments on the WMT data-sets[2], including WMT17 Chinese to English CWMT part (Zh→En), WMT 14 English to German (En→De) and English to Romanian (En→Ro). On the Zh→En, our training set has about 7.5M sentence pairs. We use `newsdev2017` as dev set which has 2002 sentence pairs, and `newstest2017` as test set which has 2001 sentence pairs. On the En→De, our training set has about 4.5M sentence pairs. We use `newstest2013` as dev set which has 3000 sentence pairs, and `newstest2014` as test set, which has 3003 sentence pairs. On the En→Ro, our training set has about 0.6M sentence pairs. We use `newstest2015` as dev set which has 2000 sentence pairs, and `newstest2016` as test set which has 2000 sentence pairs. We apply the byte pair encoding (BPE) (Sennrich et al., 2015) to all language pairs and limit the vocabulary to 32K. All out-of-vocabulary words were mapped to the *UNK* token. The same training sets were used to train a word alignment model using fast_align[3]. Then, the bilingual phrase table is extracted by Koehn et al. (2003, 2007a). We limit the length of phrase is 2-4, and finally 6.7M, 3.4M and 0.2M phrases are extracted from Zh→En, En→De and En→Ro.

Following Transformer-Base and Transformer-Big settings, we set the dimension of the input and output of all layers as 512/768, and that of the feed-forward layer to 2048/3072. We employ 8/12 parallel attention heads. The number of layers for the encoder and decoder are 6. Sentence pairs are batched together by approximate sentence length. Each batch has approximately 25000 source and 25000 target tokens. We use label smoothing with value 0.1 and dropout with a rate of 0.1. We use the Adam (Kingma and Ba, 2014) with the learning rate of 1e-3, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and it was varied under the warm-up with 4000 steps. Other settings of Transformer follow Vaswani et al. (2017).

We use beam search for heuristic decoding, and the size is set to 4. We use the *sacreBLEU*[4] to calcu-

[2]http://www.statmt.org/wmt17/translation-task.html
[3]https://github.com/clab/fast_align
[4]https://github.com/mjpost/sacreBLEU

#	Model	#Param.	Zh→En	En→De	En→Ro
1	Transformer-Base	62M	24.41	27.37	32.23
2	Transformer-Big	207M	24.72	28.47	–
3	Transformer-Base* (Vaswani et al., 2017)	65M	–	27.3	–
4	Transformer-Base* (Hassan et al., 2018)	–	24.13	–	–
5	Transformer-Base* (Gu et al., 2017)	–	–	27.02	31.91
6	Transformer-Big* (Vaswani et al., 2017)	213M	–	27.3	–
7	(Feng et al., 2020)	–	–	27.55	–
8	AOL* (Kong et al., 2019)	–	–	28.01	–
9	AOL*(Big) (Kong et al., 2019)	–	–	28.99	–
10	Dynamic Past&Future* (Zheng et al., 2019)	54M	–	28.10	32.96
11	Reorder Embedding* (Chen et al., 2019)	107M	–	28.22	–
12	Deliberation Network*(Big) (Wang et al., 2019)	372M	–	29.11	–
13	Self-supervised Learning	62M	24.39	27.50	31.98
14	MRT (Shen et al., 2016)	62M	–	27.71	–
15	Knowledge Distillation	62M	24.55	27.93	–
16	FENMT	65M	25.47‡	28.25†	33.43‡
17	FENMT (Big)	211M	26.16‡	29.36‡	–

Table 1: The comparison of our FENMT, Transformer baselines and related work on the WMT17 Chinese to English (Zh→En), WMT14 English to German (En→De), and WMT16 English to Romania (En→Ro) tasks (* indicates the results came from their paper, †/‡ indicate significantly better than the baseline ($p < 0.05/0.01$)).

late *case-sensitive* BLEU (Papineni et al., 2002) as the automatic metric. We implement the proposed approach with the implementation of Transformer derived from the *tensor2tensor*[5].

3.2 Automatic Evaluation

Translation quality. The results are summarized in Table 1. We implement the Transformer-Base and Transformer-Big as our baselines. Several Transformer systems with the same settings (Vaswani et al., 2017; Hassan et al., 2018; Gu et al., 2017) are reported as a comparison (line 1-6). Then, several related researches about improve faithfulness of NMT (Kong et al., 2019; Zheng et al., 2019; Chen et al., 2019; Feng et al., 2020) or exploiting translations for improving NMT (Xia et al., 2017; Wang et al., 2019) also be reported (line 7-12). We implement three comparable approaches on our Transformer baseline, including: 1). self-supervised learning: we use the translations of training data as a self-supervision signal to fine tune the NMT model; 2). minimum risk training (MRT): we implement the MRT following Shen et al. (2016); 3). Knowledge Distillation: we adopt the KL divergence to distill knowledge from Transformer-Big to Transformer-Base (line 13-15).

The results on the ZH→EN task are shown in the third column of Table 1. The improvement of our model (FENMT) could be up to 1.03 based on the Transformer-Base baseline (line 16 vs. line 1), and 1.44 base on the Transformer-Big baseline (line 17 vs. line 2). Then, the results on the En→De task are shown in the fourth column. On this task, the proposed model with base and big settings could attain 28.25 BLEU (+0.88) and 29.36 BLEU (+0.89), which outperforms all previous studies. We also experiment our method on low resource language pair of the En→Ro. Results are shown in the last column. The improvement is 1.20 BLEU on the base setting, which is a material improvement in low resource scenario.

Experimental results on three machine translation tasks show that the proposed approach can improve translation quality which is not limited by the language or size of training data. Moreover, our method is more effect on Zh→En than De→En, which may appeal the unfaithful problem is more serious on the language pair which have a larger difference in morphology.

Model size and efficiency. The number of parameters is shown in Table 1, our work only adds 3M/4M parameters on the Base/Big settings. The training efficiency of our FENMT based on the base

[5]https://github.com/tensorflow/tensor2tensor

Model	Degree	Addition	Omission	Grammar	Style	Others	Total
Baseline	Minor	0	0	50	15	3	68
	Major	6	54	3	0	0	63
	Critical	0	5	0	0	1	6
FENMT	Minor	0	0	41	12	8	61(-10.3%)
	Major	1	43	4	0	0	48(-23.8%)
	Critical	0	3	0	0	0	3(-50.0%)

Table 2: Human evaluation on 100 sentences sampled from Zh→En test set. We divide mistranslations into several types: **Addition** includes repetitive and useless translation, **Omission** means a consecutive part is not be translated correctly (miss or wrong), **Grammar** includes word order, word form, function word, etc. **Critical**, **Major** and **Minor** mean the degree of errors. We invite a professional translator to label errors in the sampled sentences.

Quality	Baseline	FENMT
Incomprehensible (1)	0	0
Bad (2)	7	3(-57.1%)
Understandable (3)	43	29(-32.6%)
Good (4)	42	54(+28.6%)
Excellent (5)	8	14(+75.0%)
Overall score	3.51	3.79

Table 3: Human evaluation on 100 sentences sampled from Zh→En test set. We divide translation quality into 5 levels and give score 1 to 5 (larger is better). We ask a professional translator to score them. The overall score is the weighted average of above categories.

Model	BLEU	Δ
Baseline	27.37	–
FENMT-Base	28.25	+0.88
w/o $\mathcal{L}_A$ w/o $\mathcal{L}_C$	28.00	+0.63
w/o $\mathcal{L}_M$ w/o $\mathcal{L}_C$	27.70	+0.33
w/o $\mathcal{L}_M$ w/o $\mathcal{L}_A$	27.89	+0.42
w/o $\mathcal{L}_C$	28.14	+0.77
w/o $\mathcal{L}_A$	28.10	+0.73
w/o $\mathcal{L}_M$	27.86	+0.49

Table 4: Ablation study on the En→De task.

setting is 0.86x compared with Transformer-Base, and based on the big setting is 0.94x compared with Transformer-Big.[6] Our approach only influence the training process of NMT, so the inference efficiency will not be affected.

3.3 Human Evaluation

The automatic metric, i.e., BLEU, sometimes can't accurately evaluate translation quality. For example, the sentence missing content words has decrease more on faithfulness than missing function words, but the BLEU scores may be equal. So, we make detailed human evaluations to see the variations of translation quality in the real environment.

Number of mistranslations. We divide mistranslations into several types and each type has three degrees. We sample 100 sentences from the Zh→En test set, and invite a professional translator to label errors contained in these translations.

The results are reported in Table 2. Our method can reduce the number of mistranslations at the most of categories. Typically, our approach significantly reduce the number of the **Omission**, which means a continue part from the input doesn't be translated correctly. At the **Addition** category, our approach also achieves remarkable improvement even it's not a main error type in current NMT. Omission and Addition are two serious error types greatly hurting the faithfulness of translations. The reduction of these errors will improve the faithfulness of translations obviously.

Translation quality ranking. Besides evaluating the error types in the sampled sentences, we also evaluate the overall quality for each sentence. Here, the translation quality is divided into 5 levels and give score 1 to 5 (larger is better) and a professional translator is invited to score them.

The results are shown in Table 3, the overall score of the proposed method is better than baseline ($3.79 > 3.51$). Specifically, the good (4) and excellent(5) translations from our approach are more than baseline (+75.0% and + 28.6%) by revising the errors from the bad (2) and understandable (3) translations (-57.1% and -32.6%). This results show that the reduction of mistranslations really improve the overall quality for human readers.

[6] All comparisons here were on a single GPU (Tesla P100).

Model	Number of Phrases	Accuracy
Reference	8082	–
Baseline	5676	70.2%
FENMT	6453	79.8%

Table 5: The accuracy of phrase translation on the En→De task.

Model	Sampling Rate	BLEU
FENMT	5%	27.79
	20%	28.27
	100%	28.39
	ours (20%→5%)	28.25

Table 6: The effectiveness of different sampling rates on the En→De task. *ours*: computed by Eq. 8.

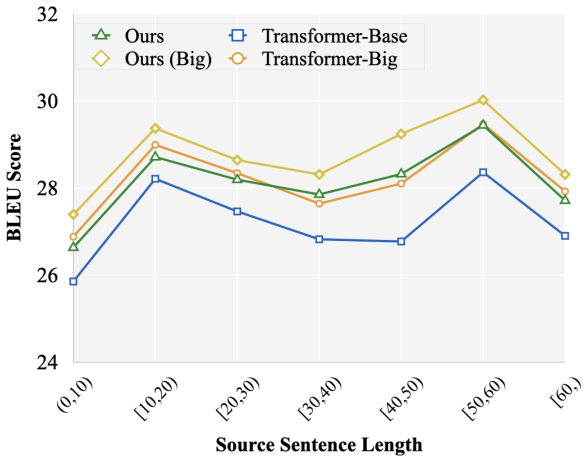

Figure 5: Performance of translations with different lengths of source sentences on the En→De task. "Ours" means the proposed FENMT.

3.4 Analysis

Ablation study. To further show the function of each task in our approach, we make ablation study in this section. Specifically, we investigate how the *masked language model objective*, *word alignment objective*, and *max-margin objective* affect the translation performance.

The results are shown in Table 4. Firstly, we analysis the effect of each task. The model achieves 0.63, 0.33 and 0.42 gains when only using masked language model ($\mathcal{L}_{\text{M}}$), word alignment ($\mathcal{L}_{\text{A}}$) and max-margin ($\mathcal{L}_{\text{C}}$) individually. Then, the results of combining two of three tasks are shown in the second part. The masked language model combines word alignment or max-margin can get improvements of 0.77 and 0.73, which are close to the best performance. While the combination of word alignment and max-margin is not work well (+0.49).

The above experimental results show that each task could get a decent improvement. But compared with improving the ability of the decoder, the high quality contextual representation learned from the masked language model is more important.

Accuracy of phrase translation. We compute the accuracy of phrase translation on the En→De task to evaluate the proposed multi-task objective in a fine-grained aspect. The result are shown in Table 5. The total number of phrases in the references is 8082. Our approach successfully translate the 6453 (79.8%) and the baseline correctly translate the 5676 (70.2%). The accuracy of our approach largely improves 9.6% compared with the baseline.

Analysis of different sampling rate. The results of the FENMT with different sampling rate are shown in Table 6. When the sampling rate is 5%, the performance decreases 0.46 compared with the rate computed by Eq. 8. When the sampling rate is larger than 20%, the performance does not change significantly. But the dynamical sampling rate will reduce the number of sentences needed to be translated, which can avoid dropping training efficiency.

Analysis of sentence length. We group the En→De test set by the length of source sentences, and then re-evaluate the BLEU score of each group. The test set is divided into 7 subsets. Figure 5 shows the results. We find that our model outperforms the baseline in all categories in both base and big setting. The proposed model performs better on long sentences (e.g., [30,60]). Because long sentences are usually complex and difficult to translate which causes the number of mistranslations in them is more than short sentences. Our approach can avoid these mistranslations compared with baselines.

Case study. We show two cases from the Zh→En task to see the difference between baseline and our approach, which are shown in Table 7.

Our approach could learn how to translate the difficult fragments in the input which are easier to be mistranslated. For example, the idiom "*turn the table*" in case 1 is translated to *loss* by the baseline, which only observe the word "败" in the input. In case 2, the baseline makes a serious mistake at the beginning of the sentence. The translation of "私

Input	不论他是从前面，还是后面靠近你，它都会教你如何反败为胜。
Refer.	whether you're approached from in front or behind, it will show you how to turn the tables on your mugger.
Baseline	whether he comes from the front, or from the *front*, it will teach you how to *lose*.
FENMT	whether you're approached from in **front or back**, it will show you how to **turn the tables**.
Input	私募股权基金等领域风险事件的爆发, 资产托管机制是一个重要原因。
Refer.	asset custody mechanism is a major reason to explain the outbreak of risk events in private equity funds and other sectors .
Baseline	*trust* is an important reason for the outbreak of risk events in private equity fund .
FENMT	**asset custody mechanism** is an important reason for the outbreak of risk events in private equity funds and **other sectors** .

Table 7: Translation cases from Transformer and FENMT on the Zh→En task. Words with **Bold** and blue fonts are correct translations revised by our model. Words with *Italic* and red fonts are mistranslations from baseline. Words with underline are the corresponding parts in the reference.

募股权基金" is omitted. Our FENMT avoids this kind of mistakes by specialized modeling mistranslated parts in the NMT model.

4 Related Work

Enhancing faithfulness for NMT. Faithfulness and Fluency are two fundamental factors of translation quality. NMT has been able to generate fluent sentences. While translating faithful sentences is an urgent problem to be solved. In the RNN-based NMT, Tu et al. (2016) and Mi et al. (2016) proposed a coverage mechanism to improve the accuracy of translation outputs. Following this intuition, Zheng et al. (2018) divided source representation into past and future parts to fine-grained control translation process. These studies focus on using source representation effectively. On the other hand, improving the ability of the decoder is another way. Tu et al. (2017) proposed to introduce a reconstruction loss to make translation can reconstruct the input sentence. Weng et al. (2017) proposed a bag-of-words loss to constrain decoding process. These methods are similar to multi-task learning, but the motivation of them are different.

Recent studies found that Transformer also suffer this problem even its translation quality is far better than RNN model. Kong et al. (2019) proposed a coverage difference ratio metric as a reward to train the Transformer model. Weng et al. (2020) proposed to model global representation in the source side to improve the source representation. Zheng et al. (2019) proposed a capsule based module to control the source representation dynamically in the decoding process. Zhang et al. (2019), Feng et al. (2020) and Garg et al. (2019) proposed to introduce word alignment information in Transformer to improve translation accuracy. However, they only focus on one side causing this problem while don't have an overall solution. Our study is the first work to pay attention to using mistranslations guides NMT model to avoid making these mistakes again.

Multi-task learning in NMT. Multi-task learning has been widely used in NMT. Dong et al. (2015) proposed to share an encoder between different translation tasks to exploit multi lingual knowledge. Luong et al. (2015a) proposed to jointly learn the translation task for different languages, the parsing task and the image captioning task, with a shared encoder or decoder. Zhang and Zong (2016) and Domhan and Hieber (2017) proposed to use multi-task learning for incorporating source/target side monolingual data in NMT. Zhou et al. (2019) introduced noisy data with multi-task learning to improve the robustness of NMT. Different from these attempts, our approach wants to improve the faithfulness of current NMT model, while learning extra knowledge from other tasks.

5 Conclusion

In this paper, we address the problem that current NMT can't generate faithful translations which will observably decrease translation quality. We propose a FENMT to learn the faithful translation from mistranslated parts. We implement the proposed method based on the Transformer model and evaluate it on three translation tasks. Both the automatic and human evaluations show that our approach can

effectively improve the faithfulness of translations. Our work can employ on different text generation tasks, e.g., text summarization and dialogue, to enhance the key phrases (or terms) generation. In the future, we will continue investigate the learning method for effectively utilizing self-generated samples and expand to other text generation tasks.

Acknowledgements

We would like to thank the reviewers for their insightful comments. Thanks to Shanbo Cheng, Shaohui Kuang and Changfeng Zhu for their constructive comments. This work is supported by National Key R&D Program of China (2018YFB1403202).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*.

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Neural machine translation with reordering embeddings. In *ACL*.

David Chiang. 2007. Hierarchical phrase-based translation. *CL*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *EMNLP*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL*.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. In *AAAI*.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *EMNLP-IJCNLP*.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv*.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007a. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007b. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL*.

Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard H. Hovy, and Tong Zhang. 2019. Neural machine translation with adequacy-oriented learning. In *AAAI*.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015a. Multi-task sequence to sequence learning. *CoRR*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015b. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *ACL*.

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

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *AAAI*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *ACL*.

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

Yiren Wang, Yingce Xia, Fei Tian, Fei Gao, Tao Qin, Cheng Xiang Zhai, and Tie-Yan Liu. 2019. Neural machine translation with soft prototype. In *NIPS*.

Xiangpeng Wei, Heng Yu, Yue Hu, Yue Zhang, Rongxiang Weng, and Weihua Luo. 2020. Multiscale collaborative deep models for neural machine translation. In *ACL*.

Rongxiang Weng, Shujian Huang, Zaixiang Zheng, Xinyu Dai, and Jiajun Chen. 2017. Neural machine translation with word predictions. In *EMNLP*.

Rongxiang Weng, H. Wei, Shujian Huang, Heng Yu, Lidong Bing, Weihua Luo, and Jiajun Chen. 2020. Gret: Global representation enhanced transformer. In *AAAI*.

Lijun Wu, Yiren Wang, Yingce Xia, Fei Tian, Fei Gao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Depth growing for neural machine translation. In *ACL*.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *NIPS*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, FeiFei Zhai, Jingfang Xu, and Yang Liu. 2019. Neural machine translation with explicit phrase alignment. *arXiv preprint arXiv:1911.11520*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*.

Zaixiang Zheng, Shujian Huang, Zhaopeng Tu, Xin-Yu Dai, and Jiajun Chen. 2019. Dynamic past and future for neural machine translation. In *EMNLP*.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lili Mou, Xinyu Dai, Jiajun Chen, and Zhaopeng Tu. 2018. Modeling past and future for neural machine translation. *TACL*.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation*.