Confidence Based Bidirectional Global Context Aware Training Framework for Neural Machine Translation

Chulun Zhou[1†], **Fandong Meng**[2], **Jie Zhou**[2],
Min Zhang[3], **Hongji Wang**[1], **Jinsong Su**[1,4*]

[1]School of Informatics, Xiamen University, Xiamen
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
[3]Harbin Institute of Technology, Shenzhen [4]Pengcheng Lab, Shenzhen
clzhou@stu.xmu.edu.cn {fandongmeng,withtomzhou}@tencent.com
zhangminmt@hotmail.com {whj,jssu}@xmu.edu.cn

Abstract

Most dominant neural machine translation (NMT) models are restricted to make predictions only according to the local context of preceding words in a left-to-right manner. Although many previous studies try to incorporate global information into NMT models, there still exist limitations on how to effectively exploit bidirectional global context. In this paper, we propose a **C**onfidence **B**ased **B**idirectional **G**lobal **C**ontext **A**ware (CB-BGCA) training framework for NMT, where the NMT model is jointly trained with an auxiliary conditional masked language model (CMLM). The training consists of two stages: (1) multi-task joint training; (2) confidence based knowledge distillation. At the first stage, by sharing encoder parameters, the NMT model is additionally supervised by the signal from the CMLM decoder that contains bidirectional global contexts. Moreover, at the second stage, using the CMLM as teacher, we further pertinently incorporate bidirectional global context to the NMT model on its unconfidently-predicted target words via knowledge distillation. Experimental results show that our proposed CB-BGCA training framework significantly improves the NMT model by +1.02, +1.30 and +0.57 BLEU scores on three large-scale translation datasets, namely WMT'14 English-to-German, WMT'19 Chinese-to-English and WMT'14 English-to-French, respectively.

1 Introduction

In recent years, Neural Machine Translation (NMT) has achieved great progress and attracted more attention. Most dominant NMT models mainly adopt an encoder-decoder framework (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017; Meng and Zhang, 2019; Song et al., 2019;

Miao et al., 2021) with the teacher-forcing strategy (Goodfellow et al., 2016) for training. Despite its success, the unidirectional property of teacher-forcing strategy restricts NMT models to only focus on the local context, i.e., the preceding words of the to-be-predicted target word at each decoder step. Apparently, this strategy tends to be limited because word dependencies are always bidirectional involving both preceding and succeeding words on the target side.

To address this issue, many previous researches attempt to exploit global information on the target side (Liu et al., 2016; Zhang et al., 2016; Serdyuk et al., 2018; Zhang et al., 2018; Su et al., 2018; Zhang et al., 2019a,b; Su et al., 2019; Zhou et al., 2019; Zhang et al., 2020). Typically, they introduce the modelling of target-side global context in the reverse direction by pairing the conventional left-to-right (L2R) NMT model with a right-to-left (R2L) auxiliary model. However, in these methods, the modelling of reverse global context is separate from the local context of preceding words. Thus, they cannot sufficiently encourage the NMT model to exploit bidirectional global context (Devlin et al., 2019). Meanwhile, some of them adopt bidirectional decoding, which often relies on multi-pass decoding or specially customized decoding algorithms (Liu et al., 2016; Zhang et al., 2018; Zhou et al., 2019; Zhang et al., 2020).

Another series of studies (Conneau and Lample, 2019; Edunov et al., 2019; Weng et al., 2020; Baziotis et al., 2020; Yang et al., 2020; Chen et al., 2020) resort to leveraging target-side bidirectional global context contained in large-scale pre-trained language models (PLM), such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). These PLMs are normally not bilingual-aware for translation and trained independently of the NMT model. As a special case, Chen et al. (2020) design a conditional masked language modelling objective to make BERT aware of source input during the fine-

---

[†]This work is done when Chulun Zhou was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
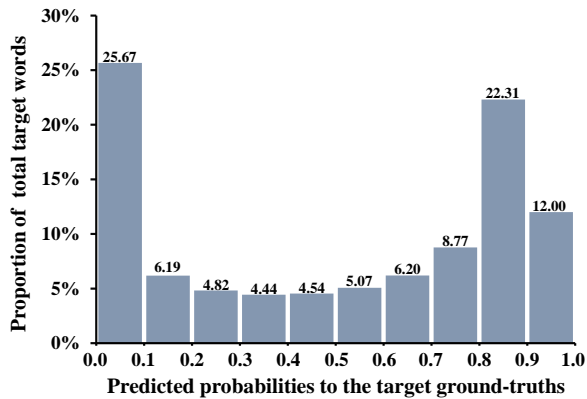[*]Corresponding author

Figure 1: The distribution of the NMT-predicted probabilities to the corresponding ground-truth words on the training set of WMT'14 English-to-German translation task, which is output by a fully-trained Transformer model using teacher-forcing strategy. For instance, the model predicts 25.67% target ground-truth words with probabilities between 0.0~0.1 given totally correct preceding words at each time step.

tuning stage. Nevertheless, in these approaches, the pre-trainings of PLMs are independent of NMT models, limiting the potential of model performance.

As for how to effectively incorporate global information into NMT models, another notable deficiency of previous work is that they do not pertinently enhance the NMT model according to its word-level prediction confidence. Ideally, under the teacher-forcing strategy, a well-trained NMT model should assign high probabilities to the target ground-truth words based on correct previous words, which, however, is not the case. Figure 1 depicts the predicted word-level probabilistic distribution of a fully-trained Transformer model. We find that, even based on totally correct preceding words, there is a considerable portion of target ground-truth words that the model predicts with relatively low probabilities. The reasonable cause of this phenomenon is that the NMT model cannot confidently predict these target words according to only the local context of preceding words (Watanabe and Sumita, 2002; Hoang et al., 2017). Hence, we should especially refine the NMT model on these unconfidently-predicted target words.

In this paper, we propose a **C**onfidence **B**ased **B**idirectional **G**lobal **C**ontext **A**ware (CBBGCA) training framework for NMT. Under our framework, the NMT model is jointly trained with a conditional masked language model (CMLM) which is essentially bilingual-aware and contains bidirectional global context on the target side. Specifically,

the CBBGCA training consists of two stages. At the first stage, we jointly train the NMT model and CMLM in a multi-task learning manner by sharing the encoders of the two models. This preliminarily enhances the NMT model because the encoder is additionally supervised by the signal from the CMLM decoder that contains bidirectional global context. At the second stage, we employ the CMLM to pertinently refine the training of the NMT model on those unconfidently-predicted target words via confidence based knowledge distillation. By doing so, our model can be further encouraged to effectively leverage the bilingual-aware bidirectional global context contained in the CMLM.

To sum up, the major contributions of our paper are as follows:

- We introduce multi-task learning to benefit the NMT model by sharing its encoder with an auxiliary CMLM, which preliminarily enhances the NMT model to capture bidirectional global context.

- We further propose confidence based knowledge distillation using the CMLM as teacher to especially refine the NMT model on unconfidently-predicted target words, more effectively exploiting the bidirectional global contextual information.

- Extensive experiments on large-scale WMT'14 English-to-German, WMT'19 Chinese-to-English and WMT'14 English-to-French translation tasks show that our CBBGCA training framework respectively improves the state-of-the-art Transformer model by +1.02, +1.30 and +0.57 BLEU points, which demonstrate the effectiveness and generalizability of our approach.

2 CBBGCA Training Framework

In this section, we will introduce our proposed CBBGCA training framework that employs a CMLM to enhance the NMT model according to its prediction confidence. In the following subsections, we first describe the basic architectures of our NMT model and CMLM. Then, we introduce the training procedures of our CBBGCA framework, involving two stages.

2.1 The NMT model and CMLM

Both the NMT model and CMLM are based on Transformer (Vaswani et al., 2017), which is essen-

tially an attentional encoder-decoder framework.[1]

2.1.1 Encoder

The encoders of the NMT model and the CMLM are identical, which are mainly used to learn the semantic representations of the source sentence.

Generally, the encoder consists of L_e identical layers, each of which contains two sub-layers: a self-attention (SelfAtt) sub-layer and a position-wise feed-forward network (FFN) sub-layer. The SelfAtt sub-layer takes the hidden states of the previous layer as inputs and conducts multi-head scaled dot-product attention. Let $\mathbf{h}^{(l)}$ denote the hidden states of the l-th encoder layer, the SelfAtt sub-layer can be formulated as

$$\mathbf{c}^{(l)} = \text{AN}(\text{SelfAtt}(\mathbf{h}^{(l-1)}, \mathbf{h}^{(l-1)}, \mathbf{h}^{(l-1)})), \quad (1)$$

where $\text{AN}(\cdot)$ denotes the *AddNorm*, i.e., layer normalization with residual connection. Afterwards, the FFN sub-layer is applied,

$$\mathbf{h}^{(l)} = \text{AN}(\text{FFN}(\mathbf{c}^{(l)})). \quad (2)$$

Note that $\mathbf{h}^{(0)}$ is initialized as the embedding sequence of the source sentence and the hidden states of the L_e-th layer $\mathbf{h}^{(L_e)}$ are used as the final word-level representations of the source sentence.

2.1.2 Decoders

The decoders of the NMT model and the CMLM are similar except their self-attention mechanisms and prediction manners.

The NMT Decoder. It is comprised of L_d identical layers with each having three sub-layers: a masked self-attention (MaskSelfAtt) sub-layer, a cross-attention (CrossAtt) sub-layer and an FFN sub-layer. Particularly, to preserve the autoregressive property at each time step, the MaskSelfAtt sub-layer performs self-attention with an attention mask that prevents the decoder from seeing succeeding words. To generate the hidden states $\mathbf{s}^{(l)}$ of the l-th decoder layer, the MaskSelfAtt sub-layer can be formulated as

$$\mathbf{a}^{(l)} = \text{AN}(\text{MaskSelfAtt}(\mathbf{s}^{(l-1)}, \mathbf{s}^{(l-1)}, \mathbf{s}^{(l-1)})). \quad (3)$$

Then, the CrossAtt sub-layer conducts cross-attention using $\mathbf{a}^{(l)}$ and the source representations $\mathbf{h}^{(L_e)}$,

$$\mathbf{z}^{(l)} = \text{AN}(\text{CrossAtt}(\mathbf{a}^{(l)}, \mathbf{h}^{(L_e)}, \mathbf{h}^{(L_e)})). \quad (4)$$

---

[1]Please note that our framework can also be adapted to other NMT models.

Next, the FFN sub-layer maps $\mathbf{z}^{(l)}$ into $\mathbf{s}^{(l)}$:

$$\mathbf{s}^{(l)} = \text{AN}(\text{FFN}(\mathbf{z}^{(l)})). \quad (5)$$

Finally, with the source sentence $\mathbf{x}$, the target translation $\mathbf{y}_{<t}$ and the learned top-layer hidden states $\mathbf{s}$, the decoder models the probability distribution over the target vocabulary at the t-th time step as follows:

$$p(y_t|\mathbf{y}_{<t}, \mathbf{x}) = \text{softmax}(\mathbf{W}\mathbf{s}_t), \quad (6)$$

where $\mathbf{W}$ represents the learnable parameter matrix for the linear transformation.

The CMLM Decoder. Typically, it predicts a set of masked target words $\mathbf{y}_m$ given the source sentence $\mathbf{x}$ and the set of observable target words $\mathbf{y}_o$. The CMLM decoder also contains L_d identical layers, each of which also includes a SelfAtt sub-layer, a CrossAtt sub-layer, and an FFN sublayer. Unlike the MaskSelfAtt sub-layer of the NMT decoder, the attention mask is removed in the SelfAtt sub-layer of the CMLM decoder.

Finally, with the learned top-layer hidden states s' of the CMLM decoder, the predicted probability distribution for every masked target word $y_t \in \mathbf{y}_m$ can be formalized as

$$p(y_t|\mathbf{y}_o, \mathbf{x}) = \text{softmax}(\mathbf{W}'\mathbf{s}'_t), \quad (7)$$

where $\mathbf{W}'$ is the learnable parameter matrix of the linear transformation. Note that since the CMLM decoder takes $\mathbf{y}_o$ rather than $\mathbf{y}_{<t}$ as input, which includes both preceding and succeeding words with respect to every masked target word, it should contain bidirectional global contextual information.

2.2 Two-stage Training

The training of CBBGCA framework involves two stages. At the first stage, we jointly train the NMT model and CMLM by multi-task learning. At the second stage, according to the word-level prediction confidence, we employ the CMLM to refine the training of the NMT model through knowledge distillation.

2.2.1 Stage 1: Multi-task Joint Training

In the first training stage, given a batch of training instances, we jointly train the NMT model and CMLM by simultaneously optimizing their respective objectives:

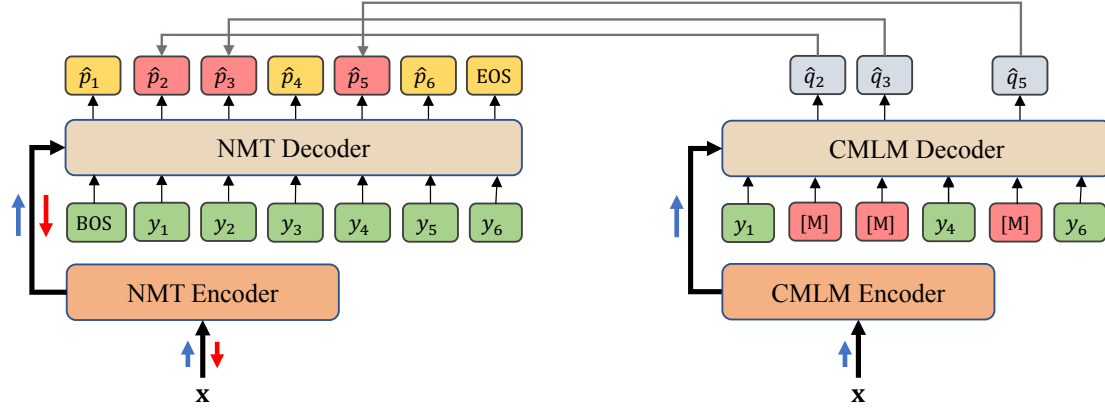$$L_1(\theta_e, \theta_{nd}, \theta_{cd}) = \lambda L_{\text{nmt}} + (1-\lambda)L_{\text{cmlm}}, \quad (8)$$

Figure 2: The second training stage. Supposing the NMT-predicted $\hat{p}_2^*$, $\hat{p}_3^*$ and $\hat{p}_5^*$ are lower than ϵ, the set $\mathbf{y}_m$ of masked target words becomes $\{y_2, y_3, y_5\}$. Accordingly, the input $\mathbf{y}_o$ to the CMLM decoder is $y_1, [M], [M], y_4, [M], y_6$. The blue and red arrow also represent the forward an backward propagation pass, respectively. Note that we separate the previously-shared encoder and fix the parameters of the CMLM at this stage.

where λ is a balancing hyper-parameter and θ_e, θ_{nd} and θ_{cd} denote the parameters of the shared encoder, the NMT decoder and the CMLM decoder, respectively. Very importantly, during this procedure of joint training, we share the same encoder for the NMT model and CMLM. In this manner, the NMT model benefits from the multi-task learning because the encoder is additionally supervised by the signal from the CMLM decoder which contains bidirectional global context.

Specifically, using the teacher-forcing strategy, the NMT model is optimized through the following objective:

$$L_{\mathrm{nmt}}(\theta_e, \theta_{nd}) = -\sum_{t=1}^{|\mathbf{y}|} \log p(y_t|\mathbf{y}_{<t}, \mathbf{x}). \quad (9)$$

Besides, we adopt the strategy used in (Ghazvininejad et al., 2019) to optimize the CMLM. Concretely, we randomly select n words, where $n \sim uniform(1, |\mathbf{y}|)$, and replace each word with a special token [M], splitting $\mathbf{y}$ into $\mathbf{y}_o$ and $\mathbf{y}_m$. Formally, we optimize the CMLM by minimizing the following objective for every word in $\mathbf{y}_m$:

$$L_{\mathrm{cmlm}}(\theta_e, \theta_{cd}) = -\sum_{y_t \in \mathbf{y}_m} \log p(y_t|\mathbf{y}_o, \mathbf{x}). \quad (10)$$

2.2.2 Stage 2: Confidence Based Knowledge Distillation (CBKD)

At the second stage, once we obtain the two fully-trained models, we use the CMLM to further refine the training of the NMT model through knowledge distillation (KD). The reason why we introduce such a KD-based model training is that the conventional NMT model predicts a considerable portion of target ground-truth words with relatively low probabilities, as shown in Figure 1. This phenomenon indicates that the NMT model cannot confidently predict these target words based on only local context of preceding words. Therefore, we aim to pertinently distill the knowledge of CMLM into the NMT model on these unconfidently-predicted target words because the CMLM contains bilingual-aware bidirectional global context.

Figure 2 depicts the training procedure of this stage with an illustrative example. Given the source sentence $\mathbf{x}$ and the preceding ground-truth words $\mathbf{y}_{<t}$ at each time step t, we first let the NMT model make predictions for every target word using Equation 6, producing word-level probability distributions $\hat{p}_1, \hat{p}_2, ..., \hat{p}_{|\mathbf{y}|}$. Then, we determine the word set $\mathbf{y}_m$ where the predicted probabilities $\hat{p}_t^*$ to the corresponding ground-truth words are lower than a threshold value ϵ,

$$\mathbf{y}_m = \{y_t | \hat{p}_t^* \leq \epsilon, 1 \leq t \leq |\mathbf{y}|\}. \quad (11)$$

Next, we obtain the set $\mathbf{y}_o$ of partially observable target words by replacing those selected ground-truth words with a special token [M]. Subsequently, we feed $\mathbf{y}_o$ to the CMLM and obtain its predicted probability distribution $\hat{q}_t$ for every word in $\mathbf{y}_m$ using Equation 7.

To pertinently refine the NMT model on the set $\mathbf{y}_m$ of its unconfidently-predicted target words, we use the CMLM with fixed parameters as teacher and transfer its knowledge to the NMT model. Along with the supervision from the corresponding ground-truth words, we optimize the NMT model with a balancing factor α through the following

objective:

$$L_{\mathrm{kd}}(\theta_{ne},\theta_{nd}) = \sum_{y_t \in \mathbf{y}_m} \{\alpha \mathrm{KL}(\hat{q}_t \| \hat{p}_t) - (1-\alpha)\log \hat{p}_t^*\}, \tag{12}$$

where θ_{ne}, θ_{nd} and $\mathrm{KL}(\cdot)$ represent the NMT encoder parameters, the NMT decoder parameters and the Kullback–Leibler divergence (Sohn et al., 2015), respectively. Here, we follow (Clark et al., 2019) to linearly decrease the factor α from 1 to 0 throughout training. This guides the NMT model to absorb more knowledge from the CMLM at the early period of the stage 2 and gradually re-focus on the ground-truth words to learn better. Finally, the total training objective of this stage is as follows:

$$L_2(\theta_{ne},\theta_{nd}) = L_{\mathrm{kd}}(\theta_{ne},\theta_{nd}) - \sum_{y_t \in \mathbf{y}_o \backslash [\mathrm{M}]} \log \hat{p}_t^*, \tag{13}$$

where $\mathbf{y}_o \backslash [\mathrm{M}]$ represents $\mathbf{y}_o$ excluding all special tokens [M]. By doing so, we can fully strengthen the ability of the NMT model to leverage the bilingual-aware bidirectional global context contained in the CMLM. Note that the CMLM is not involved at the inference time.

3 Experiments

3.1 Datasets

We carry out experiments on three large-scale translation tasks, WMT'14 English-to-German (En→De), WMT'19 Chinese-to-English (Zh→En) and WMT'14 English-to-French (En→Fr). The data are preprocessed using Byte-Pair-Encoding[2] (Sennrich et al., 2016) (BPE). More dataset statistics and the detailed preprocessing procedures are described in Appendix A.

3.2 Implementation Details

We follow the settings used in (Vaswani et al., 2017) to build the NMT model under *Transformer-base* configuration. Concretely, the *Transformer-base* architecture is comprised of 6 encoder and decoder layers, each with 512 as hidden size, the FFN sublayers of 2,048 dimension and 8 heads in multihead attentions. For more details about the training and inference, please refer to Appendix B.

3.3 Hyper-parameters

Apart from all the hyper-parameters we empirically set based on previous experience, the balancing
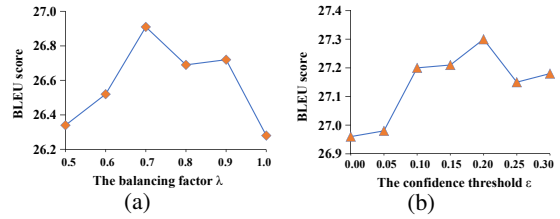
[2]https://github.com/rsennrich/subword-nmt

Figure 3: The performance of the NMT model on the WMT'14 En→De validation set with different hyperparameters. (a) The BLEU scores after the first stage with different λ, the balancing factor in Equation 8. (b) The BLEU scores after the second stage with different ϵ, the confidence threshold in Equation 12.

factor λ in Equation 8 and the confidence threshold ϵ for determining $\mathbf{y}_m$ in Equation 13 are the hyper-parameters we need to manually tune on the validation set.

To balance the training of the NMT model and CMLM, we select the minimum λ that can bring steady improvements to the NMT model within 200,000 steps. As shown in Figure 3(a), we gradually vary λ from 0.5 to 1.0 with an increment of 0.1 and evaluate the performance on the validation set. We find that the NMT model achieves its peak when $\lambda = 0.7$. Hence, λ is set to 0.7 for the joint training of the two models at the first stage.

Given the selected λ, at the second training stage, we also investigate the impact of ϵ on the validation set. We adjust its value from 0.0 to 0.3 with an interval of 0.05. As shown in Figure 3(b), the NMT model performs the best when the ϵ is 0.2. Therefore, we set $\epsilon = 0.2$ as the confidence threshold for the second training stage.

3.4 Main Results

In our experiments, **CBBGCA** is the system under our proposed training framework, as described in Section 2.2. **Multi-300k** denotes the baseline system that jointly optimizes the NMT model and CMLM by sharing their encoders throughout the whole 300k training steps, which is used to make comparison with conducting CBKD at the second training stage. For evaluation, in addition to the widely used BLEU (Papineni et al., 2002), we also adopt the Comet (Rei et al., 2020) which is recently a more welcomed metric.

Results on WMT'14 En→De. Table 1 lists several existing competitive NMT systems and ours. First, we can see that "Multi-300k" surpasses "Transformer" by +0.58 BLEU and +0.0101 Comet scores. Moreover, the BLEU and Comet scores

System	BLEU	Comet
Existing Systems		
Transformer (Vaswani et al., 2017)*	27.30	-
Rerank-NMT (Liu et al., 2016)*	27.81	-
ABD-NMT (Zhang et al., 2018)*	28.22	-
FKD-NMT (Zhang et al., 2019a)**	27.84	-
SB-NMT (Zhou et al., 2019)*	**29.21**	-
DBERT-NMT (Chen et al., 2020)*	27.53	-
Our Systems		
Transformer	27.30	0.2602
Multi-300k	27.88†	0.2703†
CBBGCA	**28.32‡◇**	**0.2828‡◇**

Table 1: BLEU (%) and Comet scores on WMT'14 En→De. '*': results are taken from corresponding papers. '**': results are reproduced by running original code. '†' and '‡': significantly better than our Transformer with t-test $p<0.05$ and $p<0.01$, respectively. '◇': significantly better than FKD-NMT with t-test $p<0.01$.

System	Zh→En		En→Fr	
	BLEU	**Comet**	**BLEU**	**Comet**
Transformer	25.54	0.3187	40.97	0.5241
Multi-300k	25.89‡	0.3263‡	41.26‡	0.5271‡
CBBGCA	**26.84‡**	**0.3544‡**	**41.54‡**	**0.5343‡**

Table 2: BLEU (%) and Comet scores on WMT'19 Zh→En and WMT'14 En→Fr. '‡': significantly ($p<0.01$) better than the Transformer.

System	BLEU	Comet
Transformer	27.30	0.2602
Multi-300k	27.88	0.2703
w/. Dynamic	27.97	0.2734
CBBGCA	28.32	0.2828
w/o. ShareEnc	27.76	0.2683
w/o. CBKD	27.92	0.2712

Table 3: BLEU (%) and Comet scores of ablation results on the test set of WMT'14 En→De.

of "CBBGCA" are respectively +1.02 and +0.0226 higher than "Transformer", verifying that CBKD at the second training stage brings further improvement.

Next, our proposed framework outperforms most recent competitive models. Specifically, "CB-BGCA" yields a better result than "ABD-NMT" (Zhang et al., 2018) and "FKD-NMT" (Zhang et al., 2019a) that only provide the NMT model with unidirectional future contexts. This proves the power of incorporating target-side bidirectional global context into the NMT model. Note that "ABD-NMT" needs two-pass decoding, which first obtains reverse state sequence by a backward decoder and then uses the forward decoder to generate final translations. The only exception is "SB-NMT" (Zhou et al., 2019) that designs elaborately customized bidirectional decoding algorithms, which is actually not fairly comparable to ours because of its decoding manners and the involvement of synthetic training data.[3] These results demonstrate the effectiveness of our proposed training framework.

Results on WMT'19 Zh→En and WMT'14 En→Fr. From Table 2, the "Multi-300k" preliminarily gains +0.35 and +0.29 BLEU scores over "Transformer" on Zh→En and En→Fr, respectively. Moreover, our "CBBGCA" further strongly outperforms "Multi-300k" and achieves a total of +1.30 and +0.57 BLEU score improve-

ments over "Transformer" on the two datasets, respectively. In term of Comet, comparing different models, we can see similar results. Note that the sizes of WMT'19 Zh→En (20M) and WMT'14 En→Fr (36M) datasets are much larger than that of WMT'14 En→De (4.5M) dataset, demonstrating the effectiveness of our proposed framework on various language pairs.

4 Analysis

4.1 Ablation Study

To fully investigate each part of our proposed training framework, we conduct ablation studies on WMT'14 En→De translation task. Table 3 reports the ablation results on the test set.

We first validate the necessity of our two-stage strategy by only training the model using the multi-task joint training. "*w/.* Dynamic" means the weights in Equation 8 are dynamically adjusted. Specifically, we linearly increase λ from 0.5 to 1.0 throughout the whole training process. We can see that its performance is just slightly higher than the fixed-weight "Multi-300k" and still significantly inferior to the two-stage "CBBGCA".

For the feasible options in two-stage strategy, "*w/o.* ShareEnc" represents not sharing encoders at the first training stage and its performance decreases by 0.56 BLEU score. This shows that the NMT encoder is enhanced by the joint training

---

[3] The result of "SB-NMT" is the reported performance with checkpoint averaging technique from (Zhou et al., 2019).

with the CMLM. As for "*w/o.* CBKD"[4], which means not performing KD on any target words at the second training stage (i.e., $\alpha=0$), its performance also decreases by 0.40 BLEU score. This demonstrates the effect of pertinently incorporating bidirectional global context into the NMT model on its unconfidently-predicted target words.

4.2 Effects of Different KD Strategies

In our training framework, for each sentence pair, we adopt KD to transfer the knowledge of the CMLM into the NMT model only on the word set $\mathbf{y}_m$. The set is determined by masking k target words whose NMT-predicted probabilities to the corresponding ground-truths are lower than a threshold ϵ. Obviously, there are alternative strategies for the above process. Therefore, we further investigate the following variants:

- **Random**: Regardless of confidence, we randomly select k words of a target sentence to be masked for the CMLM.
- **NMT-High**: As a contrast, we mask the target words whose NMT-predicted probabilities to the ground-truths are higher than $1 - \epsilon$.
- **NMT-Wrong**: We mask the target words where the predictions of the NMT model do not coincide with the corresponding ground-truths.
- **All-at-Once**: In this variant, to validate the necessity of selectively distilling knowledge on a portion rather than all of target words, we generate CMLM-predicted probability distributions for all target words. As an extreme case, we mask all target words at once with only source sentences as input to the CMLM.
- **Part-to-All**: Instead of masking all target words at once, we generate the CMLM-predicted probability distributions in a part-to-all way. Concretely, we first generate the NMT-predicted probability distributions for all words. Then, all target words are divided into several non-overlapping subsets, each corresponding to a certain probability interval. For each time, we mask a subset of target words whose probabilities to the ground-truths are located within the corresponding interval.

System	BLEU	Comet
Transformer	27.30	0.2602
Multi-300k	27.88	0.2703
CBBGCA	**28.32**	**0.2828**
KD Strategy		
Random	28.09	0.2774
NMT-High	27.85	0.2711
NMT-Wrong	27.98	0.2734
All-at-Once	27.64	0.2687
Part-to-All	28.06	0.2769

Table 4: BLEU (%) and Comet scores on the test set of WMT'14 En→De with different KD strategies.

Particularly, because the hyper-parameter ϵ is 0.2 in "CBBGCA", we have a total of 5 iterations and the intervals are set to $[0.0, 0.2]$, $[0.2, 0.4]$, $[0.4, 0.6]$, $[0.6, 0.8]$, $[0.8, 1.0]$.

Table 4 lists the results with different KD strategies. We can observe that all these variants are inferior to our "CBBGCA" method. Particularly, the results of "Random" and "NMT-High" indicate that conducting knowledge distillation on either randomly selected or confidently-predicted target words is less effective than on those unconfidently-predicted ones. Next, the result of "NMT-Wrong" is lower than "CBBGCA". It may be due to the fact that the NMT model assigns low probabilities to some correctly-predicted target words. Thus, the NMT model fails to absorb the beneficial knowledge from CMLM on these words. Lastly, "All-at-Once" and "Part-to-All" represent two approaches to generate CMLM-predicted probability distributions for all target words. It is reasonable for "All-at-Once" to obtain a a worse performance since the CMLM cannot predict well without any observable word on the target side. For "Part-to-All", we can see it improves the NMT model over "Multi-300k" but is still worse and takes more computational cost than "CBBGCA". This also echoes the finding in "NMT-High" that applying KD on confidently-predicted target words is not optimal. All these results demonstrate that it is crucial for the NMT model to pertinently exploit bidirectional global contexts on its unconfidently-predicted target words.

4.3 Change of Model Confidence

We also investigate the change of model confidence with respect to target ground-truth words on the

---

[4]It differs from "Multi-300k" in that the CMLM is not optimized at the second stage (200k∼300k steps) in "*w/o.* CBKD". In "Multi-300k", between 200k∼300k steps, the CMLM and NMT model continue to be jointly optimized by sharing their encoders.

	[0,0.1)	[0.1,0.2)	[0.2,0.3)	[0.3,0.4)	[0.4,0.5)	[0.5,1]
Transformer	25.67	6.19	4.82	4.44	4.54	54.34
CBBGCA	25.36	5.81	4.49	4.18	4.40	55.76
Δ	-0.31	-0.38	-0.33	-0.26	-0.14	+1.42

Table 5: The percentage of tokens within each probability interval on the WMT'14 En→De training set.

training set. Table 5 lists the percentage of tokens within each interval, in terms of NMT-predicted probability. Because the probability higher than 0.5 must be the maximum across the vocabulary, we group 0.5∼1.0 as a whole high-confidence interval while the others are low-confidence intervals. From the table, we can observe that the number of tokens in low-confidence intervals drops. For instance, the number of tokens locating in $[0.0, 0.2]$ becomes 0.69% fewer, which is a notable change considering that the WMT'14 En→De training set contains roughly 4.5 million sentences with a total of approximately 140 million tokens. This indicates that the NMT model becomes more confident about the target ground-truth words.

4.4 Integration with Large-scale PLM

There are also some researches (Edunov et al., 2019; Conneau and Lample, 2019; Yang et al., 2020; Weng et al., 2020; Chen et al., 2020) that focus on incorporating large-scale PLMs into the NMT model. Different from these approaches that require pre-training on massive external data, in this work, the integration of CMLM into the training procedure is to directly provide the NMT model with target-side bidirectional global context without external data. To show that our proposed training framework is compatible and orthogonal to existing approaches involving large-scale PLMs, we conduct experiments with an external Roberta model (Liu et al., 2019) in Appendix C.

4.5 Comparison over Stronger Systems

To further validate our proposed training framework, we conduct experiments over stronger systems. Particularly, we first compare our proposed training framework with (Baziotis et al., 2020) that used a bidirectional LM as prior to regularize the NMT model, which is similar to our method to some extent. Following their setting, we train a target-side language model using a 6-layer Transformer decoder. Then, it is used as the teacher model to impose soft constraints on the output of the NMT model. The upper rows of Table 6

System	BLEU	Δ
Transformer	25.54	ref.
Transformer + LM prior	25.85	+0.31
CBBGCA	**26.84**	**+1.30**
Transformer *w/.* BT data	25.85	ref.
Multi-300k *w/.* BT data	26.30	+0.45
CBBGCA *w/.* BT data	**27.12**	**+1.27**

Table 6: BLEU (%) scores on the test set of WMT'19 Zh→En. The upper rows list the results of models trained using only the original training set of WMT'19 Zh→En. The lower rows represents the performances of models trained with additional back-translated data.

give the comparison results. We can see that our "CBBGCA" still outperforms "Transformer + LM prior".

In addition, back-translated data are often used to boost NMT models. We also involve additional back-translated data during training. Specifically, we use an in-house English-to-Chinese *Transformer-base* model to translate the English sentences in the WMT'19 Zh→En training set. Then, we add these back-translated data to the WMT'19 Zh→En training set. The lower rows of Table 6 lists the performance of our models under this setting. Similar to previous results, it shows that both of "Multi-300k" and "CBBGCA" consistently improve the NMT model on the BT-augmented WMT'19 Zh→En, demonstrating the the effectiveness of our proposed training framework.

4.6 Case Study

In Appendix D, we give an illustrative example on the WMT'19 Zh→En test set to show the improvements of our model.

5 Related Work

5.1 Exploiting Global Context

This line of research aims at modelling target-side global context in the reverse direction with an auxiliary model. Liu et al. (2016) first adopt L2R and R2L NMT models to independently generate translations through beam search, then re-rank the candidate list via their agreement. Zhang et al. (2018) employ a backward decoder to capture reverse target-side context, which is then exploited by the forward decoder. Serdyuk et al. (2018) pro-

pose twin networks, where the forward network is encouraged to generate hidden states similar to those of the backward network. Zhang et al. (2019a) present a future-aware knowledge distillation framework enabling the unidirectional decoder to explore the future context for word predictions. Zhou et al. (2019) propose a synchronous bidirectional NMT model with revised beam search algorithm that involves interactive L2R and R2L decodings. Zhang et al. (2019b) also combine the L2R and R2L NMT models by considering their agreement, helping the model to generate sentences with better prefixes and suffixes. Although these methods indeed gain some improvements, the modelling of reverse global context is independent of the local context of preceding words. Meanwhile, they usually rely on elaborately designed mechanisms for burdensome multi-pass decoding.

5.2 KD from Pre-trained Language Model

Another line of research is to exploit global contextual information contained in large-scale pre-trained language models (PLM) via knowledge distillation (KD). For example, Edunov et al. (2019) and Conneau and Lample (2019) feed the top-layer representations of ELMo or BERT to NMT encoders. Yang et al. (2020) explore three techniques to apply BERT on NMT models, namely asymptotic distillation, dynamic switch for knowledge fusion, and rate-scheduled updating. Weng et al. (2020) propose a training framework consisting of a dynamic fusion mechanism and a continuous KD paradigm to leverage the knowledge of various PLMs. Baziotis et al. (2020) incorporate a language model prior for low-resource NMT. Chen et al. (2020) fine-tune the BERT on the parallel corpus to make it aware of source input, and then utilize it to improve the NMT model via KD over all target words. Compared to this, CBBGCA jointly optimizes the NMT model and auxiliary model, leading to better performance. Moreover, our method just selectively conducts KD on a portion of target words, giving higher distillation efficiency, which is similar to (Wang et al., 2021). Even though these PLM-based approaches have gained remarkable improvements, they unavoidably have some inherent limitations: (1) the monolingual PLMs lacks crucial bilingual information for translation; (2) the independence between PLM pre-trainings and NMT model training. In contrast, our proposed model is able to overcome these limitations.

6 Conclusion

In this paper, we propose a CBBGCA training framework for the NMT model to effectively exploit target-side bidirectional global context with an auxiliary CMLM. The training consists of two stages. At the first stage, we introduce multi-task learning to benefit the NMT model by sharing its encoder with an auxiliary CMLM. Then, at the second stage, through confidence based knowledge distillation, we use the CMLM as teacher to especially refine the NMT model on unconfidently-predicted target words. Experimental results show that our framework can significantly improve the NMT model. Compared with previous work, neither external nor synthetic data are needed and only the NMT model is involved during inference.

Acknowledgements

References

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

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *EMNLP 2020*.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *ACL 2020*.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc V Le. 2019. Bam! born-again multi-task networks for natural language understanding. In *ACL 2019*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NIPS 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. In *NAACL-HLT 2019*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP 2019*.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA.

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2017. Towards decoding as continuous optimisation in neural machine translation. In *EMNLP 2017*, pages 146–156.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP 2004*.

Lemao Liu, Masao Utiyama, Andrew M. Finch, and Eiichiro Sumita. 2016. Agreement on target-bidirectional neural machine translation. In *NAACL-HLT 2016*.

Yinhan Liu, 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. *CoRR*, abs/1907.11692.

Fandong Meng and Jinchao Zhang. 2019. DTMT: A novel deep transition architecture for neural machine translation. In *AAAI 2019*.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. In *ACL 2021*.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT 2018*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL 2016*.

Dmitriy Serdyuk, Nan Rosemary Ke, Alessandro Sordoni, Adam Trischler, Chris Pal, and Yoshua Bengio. 2018. Twin networks: Matching the future for sequence generation. In *ICLR 2018*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *NIPS 2015*.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Trans. Assoc. Comput. Linguistics*, 7.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. In *AAAI 2018*.

Jinsong Su, Xiangwen Zhang, Qian Lin, Yue Qin, Junfeng Yao, and Yang Liu. 2019. Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding. *Artificial Intelligence*, 277.

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

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

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *ACL 2021*.

Taro Watanabe and Eiichiro Sumita. 2002. Bidirectional decoding for statistical machine translation. In *COLING 2002*.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2020. Acquiring knowledge from pre-trained model to neural machine translation. In *AAAI 2020*.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of BERT in neural machine translation. In *AAAI 2020*.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *EMNLP 2016*.

Biao Zhang, Deyi Xiong, Jinsong Su, and Jiebo Luo. 2019a. Future-aware knowledge distillation for neural machine translation. *TASLP*, 27(12).

Jiajun Zhang, Long Zhou, Yang Zhao, and Chengqing Zong. 2020. Synchronous bidirectional inference for neural sequence generation. *Artificial Intelligence*, 281:103234.

Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Rongrong Ji, and Hongji Wang. 2018. Asynchronous bidirectional decoding for neural machine translation. In *AAAI 2018*.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. 2019b. Regularizing neural machine translation by target-bidirectional agreement. In *AAAI 2019*.

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous bidirectional neural machine translation. *Trans. Assoc. Comput. Linguistics*, 7:91–105.

Appendix

A   Datasets

WMT'14 En→De. For English-to-German translation, the training set from the WMT 2014 contains about 4.5 million parallel sentence pairs. We use the *newstest2013* and *newstest2014* as the validation and test sets, respectively. We apply BPE to preprocess the data, obtaining a shared vocabulary of approximately 32,000 tokens. For evaluation, we adopt the case-sensitive BLEU scores by using the multi-bleu.perl script[5] and the Comet scores as our metrics.

WMT'19 Zh→En. For Chinese-to-English translation, we use the training data from the WMT 2019, which consists of about 20 million sentence pairs. The *newstest2018* and *newstest2019* are used as the validation and test sets, respectively. We also apply BPE to preprocess the data with 30,000 merge operations for both source and target languages. Then, we construct their corresponding vocabularies having roughly 47,000 Chinese tokens and 32,000 English tokens. To evaluate the models, we use cased BLEU scores calculated with the Moses mteval-v13a.pl script[6] as well as the Comet scores.

WMT'14 En→Fr. For English-to-French translation, the training set contains about 36 million parallel sentence pairs. We use the *newstest2013* and *newstest2014* as the validation and test sets, respectively. The preprocessing and evaluation are the same as those for WMT'14 En→De.

B   Implementation Details

During training, we set dropout to 0.1 and use label-smoothing technique of value 0.1. For parameter updating, we employ the Adam optimizer (Kingma and Ba, 2015) with $\beta 1 = 0.9$, $\beta 2 = 0.998$ and $\epsilon = 10^{-9}$. As for learning rate scheduling, we adopt the same strategy as (Vaswani et al., 2017) and set warm-up steps to 4,000. For all the three translation tasks, we train our models on Tesla V-100 GPU where we batch sentence pairs of similar lengths together containing roughly 32000 tokens. At the first stage, we train all models by sharing their encoders for 200,000 steps. At the second stage, we

System	BLEU	Δ
w/o Roberta		
Transformer	27.30	ref.
Multi-300k	27.88	+0.58
CBBGCA	**28.32**	**+1.02**
w/ Roberta		
Transformer	27.64	+0.34
Multi-300k	28.13	+0.83
CBBGCA	**28.68**	**+1.38**

Table 7: BLEU scores (%) on the test set of WMT'14 En→De. The upper rows (w/o Roberta) represent the performance without using Roberta while the lower rows (w/ Roberta) are the results with the incorporation of the Roberta model using dynamic switch mechanism.

separate their encoders and fix the CMLM parameters. Then, the NMT model is solely optimized with the fixed CMLM by additional 100,000 steps, resulting in a total of 300,000 training steps.

At the inference time, we use beam search with the beam size 5. The results are reported with the statistical significance test (Koehn, 2004) for both translation tasks.

C   Integration with Large-scale PLM

In this experiment, we adopt the dynamic switching mechanism used in (Yang et al., 2020) to incorporate externally pre-trained information into the NMT model. Specifically, we use Roberta (Liu et al., 2019) as the large-scale PLM that provides our NMT model with external knowledge. Formally, a context gate is employed to control the amount of information flowing from the pre-trained model and our NMT model, which is computed as

$$g = \sigma(\mathbf{U}r + \mathbf{V}h), \qquad (14)$$

where $\sigma(\cdot)$ is the sigmoid function, r and h respectively represent the hidden states of each token obtained from the Roberta and the NMT encoder, $\mathbf{U}$ and $\mathbf{V}$ are the trainable parameter matrix. Then, the hidden states of the Roberta model and the NMT encoder are combined as

$$h' = g \odot r + (1 - g) \odot h, \qquad (15)$$

where $\odot$ is the operation of element-wise multiplication.

Table 7 lists the performance of our model with the integration of Roberta using dynamic switch

Source	第 三 届 世界 互联 网 大会 首次 发布 了 15 项 领先 科技 成果 。
	dì sān jiè shì jiè hù lián wǎng dà huì shǒu cì fā bù le 15 xiàng lǐng xiān kē jì chéng guǒ 。
Reference	The third World Internet Conference has published 15 leading scientific and technological achievements for the first time .
Transformer	For the first time , 15 leading technology achievements were released at the Third World Internet Congress .
Multi-300k	For the first time , 15 leading scientific and technological achievements were launched at the Third World Internet Congress .
CBBGCA	For the first time , 15 leading scientific and technological achievements were launched at the Third World Internet Conference .

Figure 4: An example on the WMT'19 Zh→En test set. The green and blue words are the source Chinese words and their corresponding English ground-truth translations. Those in red are wrongly translated words.

mechanism. We can see that the incorporation of external Roberta brings some improvements to the NMT model under different settings (w/ Roberta vs. w/o Roberta). Besides, it is notable that "CBBGCA" still achieves consistent improvements over "Transformer" and "Multi-300k" regardless of whether using Roberta or not, which demonstrates the compatibility and orthogonality of our proposed training framework to those approaches (Edunov et al., 2019; Conneau and Lample, 2019; Yang et al., 2020; Weng et al., 2020; Chen et al., 2020) of integrating external large-scale PLMs.

D Case Study

We conduct case study to illustrate the improvements of our model. Figure 4 gives an example on the WMT'19 Zh→En test set with the outputs of different models.

We can see that "Transformer" inappropriately translates the Chinese words "*dà huì*" and "*kē jì*" into "*Congress*" and "*technology*", respectively. For "Multi-300k", it correctly translates "*kē jì*" into "*scientific and technological*". Although it is somewhat acceptable for "*kē jì*" to be translated as "*technology*", "*scientific and technological*" is a stricter translation and and more proper modifier considering that the succeeding word "*chéng guǒ*" ("*achievements*" is a noun. The reason is that the multi-task joint training makes the NMT model aware of bidirectional global context rather than only the local context of preceding words.

Moreover, "CBBGCA" successfully translates both "*dà huì*" and "*kē jì*". This indicates that the CBKD at the second training stage further gives our model more confidence to predict the ground truth "*Conference*" rather than its near-synonym

"*Congress*". The above analyses show that our proposed training framework actually enhances the NMT model to capture bidirectional global context and significantly improves translation quality.