# Bridging the Gap between Training and Inference: Multi-Candidate Optimization for Diverse Neural Machine Translation

Huan Lin<sup>1</sup> Baosong Yang<sup>2</sup> Liang Yao<sup>2</sup> Dayiheng Liu<sup>2</sup> Haibo Zhang<sup>2</sup>

Jun Xie<sup>2</sup> Min Zhang<sup>3</sup> Jinsong Su<sup>1,4\*</sup>

<sup>1</sup>School of Informatics, Xiamen University

<sup>2</sup>Alibaba Group <sup>3</sup>Soochow University <sup>4</sup>Pengcheng Lab, Shenzhen

huanlin@stu.xmu.edu.cn

{yangbaosong.ybs,yaoliang.yl,liudayiheng.ldyh,zhanhui.zhb}@alibaba-inc.com gingjing.xj@alibaba-inc.com minzhang@suda.edu.cn jssu@xmu.edu.cn

#### Abstract

Diverse NMT aims at generating multiple diverse yet faithful translations given a source sentence. In this paper, we investigate a common shortcoming in existing diverse NMT studies: the model is usually trained with single reference, while expected to generate multiple candidate translations in inference. The discrepancy between training and inference enlarges the confidence variance and quality gap among candidate translations and thus hinders model performance. To deal with this defect, we propose a multi-candidate optimization framework for diverse NMT. Specifically, we define assessments to score the diversity and the quality of candidate translations during training, and optimize the diverse NMT model with two strategies based on reinforcement learning, namely hard constrained training and soft constrained training. We conduct experiments on NIST Chinese-English and WMT14 English-German translation tasks. The results illustrate that our framework is transparent to basic diverse NMT models, and universally makes better trade-off between diversity and quality. Our source code is available at https://github.com/ DeepLearnXMU/MultiCanOptim.

# 1 Introduction

Recently, neural machine translation (NMT) has achieved impressive progress in improving translation quality (Sutskever et al., 2014; Luong et al., 2015; Vaswani et al., 2017). Despite the remarkable success, NMT models still suffer from lacking translation diversity, which is essential due to the following reasons. First, similar to natural language, variability and expressiveness are the core features of translations. Second, only focusing on increasing translation accuracy during training will bias the NMT model to common phrases, exacerbating data sparsity (Khayrallah et al., 2020). In conclusion, improving translation diversity is a promising direction in NMT community.

To achieve diverse NMT, several studies have explored various training or decoding strategies, including: 1) constraining decoding with diversity regularization (Li et al., 2016; Vijavakumar et al., 2018), 2) sampling from the mixture of models (Shen et al., 2019; Wu et al., 2020), and 3) conditioning decoding with diverse signals (Shu et al., 2019; Sun et al., 2020). However, all these approaches train models on single-reference corpus, while expecting them to generate multiple candidate translations during inference. We argue that such discrepancy between training and inference prevents the models from learning one-to-many relations efficiently. Firstly, since the predictions of NMT models are encouraged to fit the one-hot distribution of single-reference corpus, the model confidence of generating Top1 candidate translations will be much larger than that of the rest candidates, limiting translation diversity. Secondly, only one reference is used to get optimization signal at the training time, resulting in significant quality drops of Top2-TopK translations. One direct way addressing these issues is to train the models using multi-reference training data. Nevertheless, its construction is quite expensive and thus impractical.

To overcome the above issues, in this paper, we propose a novel multi-candidate optimization framework for diverse NMT. The basic idea is to guide an NMT model to learn diverse translation from its candidate translations based on reinforcement learning (RL). During training, the model generates multiple candidate translations, of which rewards are quantified according to their diversity and quality. Since directly optimizing model parameters with the above two rewards involves backpropagating through discrete decoding decision, we explore two specific methods to train the diverse NMT model: 1) *Hard constrained training*. We transform the rewards to discrete scalars, prevent-

<sup>\*</sup> Jinsong Su is the corresponding author. This work was done when Huan Lin was interning at DAMO Academy, Alibaba Group.

ing the model from learning those candidate translations with low rewards. 2) Soft constrained training. We introduce minimum risk training (MRT) to minimize the risks of obtaining diversity and quality rewards. Compared with previous works, our proposed framework reduces the confidence variance among candidate translations and improves the quality of Top2-TopK translations during inference, achieving better performance in terms of both diversity and quality. Overall, the major contributions of our work are three-fold:

- We point out and empirically verify that the discrepancy between training and inference in diverse NMT negatively impacts the translation diversity and quality.
- We propose a novel multi-candidate optimization framework based on RL, enabling an NMT model to learn one-to-many relations from its candidate translations. Our framework is transparent to model architecture, thereby can be employed individually or complemented to existing diverse NMT models.
- Extensive experimental results on NIST Chinese-English and WMT14 English-German datasets show that our framework can efficiently smooth the confidence distribution and raise the quality of Top2-Top*K* candidate translations, surpassing several commonly-used diverse NMT models.

# 2 Related Work

**Diverse NMT.** Improving translation diversity has been a hot topic in NMT community in recent years, such as lattice-based NMT (Su et al., 2017; Tan et al., 2018) and personalized NMT (Michel and Neubig, 2018; Lin et al., 2021). Existing works for diverse NMT can be categorized into three major categories. The first category produces diverse translations by applying diversity regularization to decoding algorithm (Li et al., 2016; Vijayakumar et al., 2018). The second category improves translation diversity by sampling from a mixture of models. In this aspect, Shen et al. (2019) adopt conditional mixture models to control the generation of translations. Wu et al. (2020) derive a large number of models with Bayesian modeling, which are sampled to generate diverse translations. Unlike the former two categories, the third one attempts to condition the decoding procedure with diverse signals. Typically, Shu et al. (2019) use syntactic codes to condition translation process.

Further, Lachaux et al. (2020) replace the syntactic codes with latent domain variables derived from target sentences, which is more computationally efficient. Sun et al. (2020) sample the encoder-decoder attention heads of Transformer to affect source word selection. Despite their successes, an obstacle of these approaches lies in the discrepancy between training and inference, that is, learning diverse translations from a single-reference corpus. This enlarges the confidence and quality gaps among candidate translations, limiting the potential of diverse NMT models.

Multi-Candidate Optimization in Natural Language Generation. Since single-reference corpus is insufficient to model one-to-many relations in natural language generation (NLG), researchers have introduced multi-candidate optimization to NLG tasks such as image captioning and paraphrasing. Most of representative works among them generate pseudo training references and focus on improving diversity (Zheng et al., 2018; Hou et al., 2018; Gao et al., 2020). Conversely, in NMT community, previous studies on multicandidate optimization mainly aim at improving low-resource translation quality rather than diversity, which is similar to other data augmentation methods in NMT, such as back-translation and forward-translation (Sennrich et al., 2016; Edunov et al., 2018a; Cheng et al., 2020; Wan et al., 2020). For example, Khayrallah et al. (2020) improve the translation quality of low-resource language pairs by sampling paraphrases of the reference sentence. Different from these tasks, diverse NMT is more challenging since it requires the generation results to be accurate as well as diverse. For better balancing quality and diversity, we propose a novel multi-candidate optimization framework with RL.

**Reinforcement Learning in NMT.** Reinforcement learning (RL) has become an appealing path for advancement in NMT, as it firstly allows to optimize non-differentiable objectives, and secondly reduces exposure bias in auto-regressive sequence generators. To this end, various methods have been proposed. In Ranzato et al. (2016), Wu et al. (2017) and Edunov et al. (2018b), the authors employ the REINFORCE algorithm to optimize models with metric-based reward (i.e. senetnce-level BLEU). Different from them, He et al. (2016) propose to train two reverse NMT models through a duallearning mechanism. And Bahdanau et al. (2017)



Figure 1: Our multi-candidate optimization framework: the diverse NMT model generates K candidate translations to receive individual diversity and quality rewards, which are then used for model optimization.

use actor-critic method that predicts the reward by a critic network. In this work, we follow Shen et al. (2016), Wieting et al. (2019) and Wang and Sennrich (2020) that adopt minimum risk training (MRT) to minimize the reward during training. To the best of our knowledge, our work is the first attempt employing RL to model one-to-many relations for diverse NMT.

# 3 Multi-Candidate Optimization Framework

As a significant extension of conventional NMT, given a source sentence, a diverse NMT model aims at producing a set of different candidate translations. Similar to conventional NMT, the most commonly-used training strategy of diverse NMT is to minimize the training objective based on maximum likelihood estimation (MLE):

$$\mathcal{L}_{mle}(\theta) = -\sum_{n=1}^{N} \log P_{\theta}(y^{(n)}|x^{(n)}), \quad (1)$$

where  $(x^{(n)}, y^{(n)})$  is the *n*-th instance in the training corpus of size *N*, and  $P_{\theta}(y^{(n)}|x^{(n)})$  denotes the translation model with parameters  $\theta$ . It can be said that the one-to-many relations are the basis of diverse NMT. However, as mentioned above, the model is unable to effectively learn such relations from a single-reference training corpus. Accordingly, the discrepancy between training and inference has become a bottleneck limiting the performance of diverse NMT models.

To deal with the above issue, we propose a multicandidate optimization framework based on reinforcement learning (RL). As shown in Figure 1, a diverse NMT model generates K candidate translations using its original method as additional references during training. Particularly, given a source sentence  $x^{(n)}$ , the model picks an action each time it generates a candidate translation  $y_k^{(n)}$ . Diversity and quality rewards of  $y_k^{(n)}$  are observed once it is completed, which are then used to optimize model parameters. Please note that our framework is model-irrelevant and thus can be compatible with any diverse NMT model. Next, we will introduce the reward computation and training procedure in following subsections.

#### 3.1 Reward Computation

Conventional RL in NMT usually takes sentencelevel BLEU (Papineni et al., 2002) as reward. However, in diverse NMT, ideal translations should be semantically equal to their source sentences, as well as diverse from each other. To this end, we exploit two highly generic evaluation metrics to encourage the diversity and quality of candidate translations:

**Diversity Reward.** This reward measures the difference between each candidate translation  $\hat{y}_k^{(n)}$  and other translations, including the original reference  $y^{(n)}$  and the rest candidate translations  $\{\hat{y}_{k'}^{(n)}\}_{k'=1,k'\neq k}^{K}$ . We can model the difference with arbitrary method such as Jaccard distance (Jaccard, 1901), edit distance (Levenshtein et al., 1966) or BLEU (Papineni et al., 2002). Here, we serve BLEU as the similarity assessment since it is less sensitive to sentence length. Formally, the diversity reward of  $\hat{y}_k^{(n)}$  is defined as follows:

$$\mathrm{DR}(\hat{y}_{k}^{(n)}) = 1 - \mathrm{BLEU}_{\mathrm{s}}\left(\hat{y}_{k}^{(n)}, \{y^{(n)}\} \cup \{\hat{y}_{k'}^{(n)}\}_{k'=1, k' \neq k}^{K}\right),$$
(2)

where  $BLEU_s(*)$  indicates sentence-level  $BLEU^1$ .

Quality Reward. One common approach to evaluate the quality of each candidate translation is to compare it with the corresponding reference. However, such a method biases the model to candidate translations syntactically similar to original references, therefore harms translation diversity. In order to tackle this problem, it is better to use semantic evaluation metrics such as reconstruction-BLEU (He et al., 2016), COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), UniTE (Wan et al., 2022) and so on. Here, we choose reconstruction-BLEU, which uses a reverse NMT model trained on initial single-reference corpus to translate each candidate  $\hat{y}_k^{(n)}$  back to a source sentence  $\hat{x}_k^{(n)}$ , and

<sup>&</sup>lt;sup>1</sup>We calculate sentence-level BLEU with SacreBLEU in https://github.com/mjpost/sacrebleu

then evaluates the BLEU score between  $\hat{x}_k^{(n)}$  and  $x^{(n)}$ :

$$QR(\hat{y}_k^{(n)}) = BLEU_s\left(\hat{x}_k^{(n)}, x^{(n)}\right).$$
(3)

Compared with COMET and BLEURT that pretrained using out-of-domain data, the reconstruction model in our metric is trained on the same corpus as NMT model, thereby reducing the impact of domain inconsistency on evaluation accuracy.

#### 3.2 Model Training

The calculations of diversity and quality rewards involve undifferentiated discrete operations, leading to a challenge in the back-propagation at training time. To address this issue, we explore two approaches, separately termed as *hard constrained training* and *soft constrained training*.

Hard Constrained Training (HCT). An intuitive idea is utilizing translations only with high diversity and quality rewards. Along with this strategy, we pair  $x^{(n)}$  with its candidate translations  $\{\hat{y}_k\}_{k=1}^K$  and original reference  $y^{(n)}$  to form a new multi-reference training instance, then optimize the model with MLE objective:

$$\mathcal{L}_{hct}(\theta) = \sum_{n=1}^{N} \log P_{\theta}(y^{(n)} | x^{(n)})$$
(4)  
+ 
$$\sum_{n=1}^{N} \sum_{k=1}^{K} \alpha_k \cdot \log P_{\theta}(\hat{y}_k^{(n)} | x^{(n)}),$$

where

$$\alpha_k = \begin{cases} 0 & \mathrm{DR}(\hat{y}_k^{(n)}) < \delta_d \text{ or } \mathrm{QR}(\hat{y}_k^{(n)}) < \delta_q \\ 1 & \text{else.} \end{cases}$$

Here,  $\alpha_k$  is used to re-weight the training objective of candidate translations.  $\delta_d$  and  $\delta_q$  indicate corpuslevel diversity and quality rewards of the initial model (i.e. a pre-trained diverse NMT model) on development set, respectively. Hard constrained training is easy to implement. However, the candidate translations are still far from utilization. First, all candidate translations are treated equally although they possess different diversity and quality. Second, some candidate translations are totally discarded although they may provide guidance for the model training.

**Soft Constrained Training (SCT).** To fully utilize all candidate translations, we employ MRT to directly optimize diversity and quality rewards.

We choose MRT since it does not require extra parameters compared with other RL techniques (He et al., 2016; Bahdanau et al., 2017). Similar to hard constrained training, a multi-reference training instance consists of each original source sentence and all its candidate translations. Specifically, we define the losses of diversity and quality rewards as  $1 - DR(\hat{y}_k^{(n)})$  and  $1 - QR(\hat{y}_k^{(n)})$ , respectively. Please refer to Equations 2 and 3 for definitions of DR(\*) and QR(\*). We then apply these two losses to softly weight the posterior distribution  $P_{\theta}(\hat{y}_k^{(n)}|x^{(n)})$ . The goal is to minimize two risks:

$$\mathcal{R}_{d}(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} P_{\theta}(\hat{y}_{k}^{(n)} | x^{(n)}) \cdot (1 - \mathrm{DR}(\hat{y}_{k}^{(n)})),$$
(5)

$$\mathcal{R}_{q}(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} P_{\theta}(\hat{y}_{k}^{(n)} | x^{(n)}) \cdot (1 - \text{QR}(\hat{y}_{k}^{(n)})).$$
(6)

Completely different to MLE (Equation 1) that aims at reducing the discrepancy between a candidate translation and its single reference, MRT encourages the model to maximize rewards via generating more diverse and accurate translations. Following Wieting et al. (2019), we first pre-train the diverse NMT model with  $\mathcal{L}_{mle}(\theta)$ , and then finetune it with the combination of  $\mathcal{L}_{mle}(\theta)$ ,  $\mathcal{R}_d(\theta)$ and  $\mathcal{R}_q(\theta)$ :<sup>2</sup>

$$\mathcal{L}_{sct}(\theta) = \mathcal{L}_{mle}(\theta) + \mathcal{R}_d(\theta) + \mathcal{R}_q(\theta).$$
(7)

Consequently, soft constrained training possesses two advantages comparing with its hard counterpart: 1) It provides more guidance for model training by exploiting all candidate translations; 2) It directly incorporates the diversity and quality rewards into training objective, thereby distinguishing different effects of candidate translations.

### 4 **Experiments**

In this section, we carry out several groups of experiments to investigate the effectiveness of our proposed framework on Chinese-English and English-German translation tasks.

<sup>&</sup>lt;sup>2</sup>We have also tried weighted sum of three terms. Results shows no significant difference with the non-weighted version.

#### 4.1 Setup

In order to make comparison with existing diverse NMT models (Li et al., 2016; Vijayakumar et al., 2018; Shen et al., 2019; Sun et al., 2020), we build multi-reference corpus and examine our framework on translation tasks commonly used in previous diverse NMT studies:

- NIST Chinese-to-English. This training set contains about 1.34M news sentence pairs.<sup>3</sup> We use MT03 as development set and MT04, MT05, MT06 as test sets, and report the average scores on test sets as final results.
- WMT14 English-German. This training data consists of 4.5M sentence pairs<sup>4</sup>. We use the newstest 2013 as the development set, and the newstest 2014 as the test set.

For the above two datasets, We adopt Moses tokenizer (Koehn et al., 2007) to deal with English and German sentences, and segment the Chinese sentences with the Stanford Segmentor<sup>5</sup>. Following common practices, we employ byte pair encoding (Sennrich et al., 2015) with 32K merge operations to segment words into subword units. We use a joint dictionary for English-German translation task while assigning individual vocabularies for Chinese-English translation task. In addition, we remove the examples in datasets where the length of source or target sentence exceeds 100 words.

We develop all diverse NMT models on Transformer-base (Vaswani et al., 2017)<sup>6</sup>. At the pre-training stage, we set the batch size as 32,768 tokens for NIST and 12,500 tokens for WMT14. Other configurations are identical to common settings in previous studies (Vaswani et al., 2017; Shu et al., 2019; Sun et al., 2020). During fine-tuning, we keep other settings consistent with the pre-training stage, but reduce the learning rate by a factor of 10. Using early-stopping strategy, we evaluate the model every 500 steps and stop training if the translation diversity or quality on development set does not raise for 10 consecutive evaluations. Considering computational efficiency, we set *K* as 3 in all our experiments as default.

#### 4.2 Evaluation

We use the following three metrics to assess the quality and diversity of candidate translation sets.

- **BLEU.** Following previous studies (Shen et al., 2019; Wu et al., 2020; Sun et al., 2020), we use average BLEU of *K* candidate translation sets to evaluate translation quality.
- **COMET.** It is based on pre-trained language model and has shown higher correlations with human judgements in a variety of metrics tasks (Mathur et al., 2020b). We adopt it since n-gram-based metrics may fail to robustly match paraphrases and capture distant dependencies, resulting in a diverse translation with high faithfulness and fluency but a low BLEU score (Smith et al., 2016; Mathur et al., 2020a). Similar to BLEU, we report the average COMET score of *K* candidate translation sets as default. Particularly, we normalize the results of COMET with sigmoid function.
- **divBLEU.** We define divBLEU to measure the differences among K candidate translations on a test set of size S:

$$1 - \text{BLEU}_{c}(\{\hat{y}_{k}^{(s)}\}_{s=1}^{S}, \{\hat{y}_{k'}^{(s)}\})$$

where  $1 \le k \le K, 1 \le k' \le K, k \ne k'$ . The second term denotes pairwise-BLEU (Shen et al., 2019; Wu et al., 2020; Sun et al., 2020) that compares each candidate translation set with each other.

All BLEU metrics used in this paper are casesensitive. Concretely, corpus-level BLEU is calculated with Moses script<sup>7</sup>. To raise the reliability, we run all models three times with different random seeds and report the average results.

# 4.3 Baselines

We apply our framework to the following models:

- **Transformer** (Vaswani et al., 2017) refers to the baseline. We pick its Top*K* hypotheses in beam search as the diverse translations.
- **Tree2Code** (Shu et al., 2019) generates diverse candidates with various syntactic codes.
- Head Sampling (Sun et al., 2020) generates different words by sampling attention heads.

We also display the reported results of several dominant diverse NMT models on the same datasets: **Diverse Decoding** (Li et al., 2016) employing diversity regularization terms to encourage translation

<sup>&</sup>lt;sup>3</sup>The training set is a combination of LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

<sup>&</sup>lt;sup>4</sup>The preprocessed data can be found and downloaded from http://nlp.stanford.edu/projects/nmt/.

<sup>&</sup>lt;sup>5</sup>https://nlp.stanford.edu/

<sup>&</sup>lt;sup>6</sup>Our codes are implemented upon https://github. com/facebookresearch/XLM/.

<sup>&</sup>lt;sup>7</sup>https://github.com/moses-smt/ mosesdecoder/blob/master/scripts/ generic/multi-bleu.perl

| Madal                                    | Chinese-English |       |                           | <b>English-German</b> |       |                           |  |  |
|--|-----------------|-------|---------------------------|-----------------------|-------|---------------------------|--|--|
| Model                                    | BLEU            | COMET | divBLEU                   | BLEU                  | COMET | divBLEU                   |  |  |
| Existing Diverse NMT Systems             |                 |       |                           |                       |       |                           |  |  |
| Diverse Decoding (Li et al., 2016)       | 43.18           | -     | 19.76                     | 25.27                 | _     | 21.43                     |  |  |
| Diverse Beam (Vijayakumar et al., 2018)  | 39.58           | -     | 41.93                     | 23.27                 | -     | 33.87                     |  |  |
| HardMOE (Shen et al., 2019)              | 38.54           | _     | 39.30                     | 23.22                 | -     | 31.97                     |  |  |
| Multinominal Sampling (Sun et al., 2020) | 20.62           | _     | 89.28                     | 11.99                 | -     | 87.16                     |  |  |
| Head Sampling (Sun et al., 2020)         | 42.66           | -     | 33.82                     | 25.62                 | -     | 21.34                     |  |  |
| Our Implementations                      |                 |       |                           |                       |       |                           |  |  |
| Transformer (Vaswani et al., 2017)       | 44.67           | 57.83 | 13.89                     | 26.29                 | 55.03 | 19.06                     |  |  |
| +HCT                                     | 44.23           | 57.72 | 14.01                     | 26.07                 | 55.01 | 19.76                     |  |  |
| +SCT                                     | 43.78           | 57.41 | <b>14.98</b> ‡            | 26.01                 | 54.85 | <b>20.33</b> <sup>†</sup> |  |  |
| Tree2Code (Shu et al., 2019)             | 42.99           | 56.79 | 34.80                     | 25.43                 | 53.52 | 26.11                     |  |  |
| +HCT                                     | 42.52           | 56.88 | 37.57 <sup>‡</sup>        | 25.15                 | 53.74 | 28.71 <sup>‡</sup>        |  |  |
| +SCT                                     | 42.26           | 57.06 | <b>38.78</b> <sup>‡</sup> | 25.40                 | 54.19 | <b>29.98</b> ‡            |  |  |
| Head Sampling (Sun et al., 2020)         | 42.52           | 56.40 | 34.02                     | 25.16                 | 53.28 | 21.24                     |  |  |
| +HCT                                     | 42.46           | 56.58 | 36.38 <sup>‡</sup>        | 25.03                 | 53.66 | 23.55 <sup>†</sup>        |  |  |
| +SCT                                     | 42.02           | 57.03 | <b>37.56</b> ‡            | 24.91                 | 54.02 | <b>24.07</b> <sup>‡</sup> |  |  |

Table 1: Main results on NIST Chinese-English (average scores of MT04, MT05 and MT06) and WMT14 English-German tasks. "HCT" and "SCT" individually indicate hard constrained training and soft constrained training. "BLEU" and "COMET" denote translation quality assessed by the n-gram-based and the model-based metrics, respectively. "divBLEU" indicates the diversity among candidates. We also calculate p-value with bootstrap sampling (Koehn, 2004) to estimate statistical significance.  $\ddagger/\dagger$ : significantly better than corresponding basic models (p < 0.01/0.05). All results are derived from 3 independent runs.

diversity during beam search; **Diverse Beam** (Vijayakumar et al., 2018) that improves the method of Li et al. (2016) by grouping the outputs; **Hard-MOE** (Shen et al., 2019) utilizing a mixture model, where different translations are obtained by controlling hidden states; **Multinominal Sampling** (Sun et al., 2020) that randomly selects words at each timestep to form diverse translations.

#### 4.4 Main Results

Table 1 shows the main results. Obviously, all basic models equipped with multi-candidate optimization achieve higher diversity while preserving semantic quality of translations, including conventional NMT model (Transformer) and diverse NMT models (Head Sampling and Tree2Code), showing universal effectiveness of the proposed framework. We further draw several conclusions:

1) The higher diversity among translations, the lower BLEU score they obtain, which is consistent with prior findings (Shen et al., 2019; Wu et al., 2020; Sun et al., 2020). The main reason is that the n-gram-based metric (BLEU) fails to accurately evaluate the quality of translations that syntactically differ from their references. The model-based metric (COMET) shows that our framework yields comparable translation quality compared with corresponding basic models. More discussions about the correlation between these two automatic metrics and human evaluation are given in Section 4.6.

2) Soft constrained training exhibits better performance than hard constrained training on three basic models. The underlying reason is that soft constrained training can fully utilize candidate translations to optimize models.

3) The improvement of Transformer is smaller than that of Head Sampling and Tree2Code. We attribute this to the relatively less diversity of references generated by conventional NMT model, which limits the effects of our framework.

# 4.5 Ablation Study

To investigate the effectiveness of different components in our framework, we further compare hard constrained training and soft constrained training with their several variants upon our best basic model Tree2Code on Chinese-English translation task, as concluded in Table 2:

1) Directly fine-tuning models on the whole multi-reference training set (Tree2Code+HCT without DR and QR) benefits translation quality while significantly harms its diversity, suggesting the importance of two rewards.

2) Using only the diversity reward (HCT $_d$  and

| Model     | DR           | QR           | BLEU<br>(Top1) | BLEU  | COMET | div-<br>BLEU |
|-----------|--------------|--------------|----------------|-------|-------|--------------|
| Tree2Code | X            | X            | 44.97          | 42.99 | 56.79 | 35.09        |
| +HCT      | X            | X            | 45.14          | 43.10 | 57.05 | 33.80        |
|           | $\checkmark$ | X            | 44.50          | 42.16 | 56.40 | 37.23        |
|           | X            | $\checkmark$ | 44.72          | 43.38 | 56.98 | 34.88        |
|           | $\checkmark$ | $\checkmark$ | 44.31          | 42.52 | 56.88 | 37.72        |
| +SCT      | $\checkmark$ | X            | 44.57          | 41.87 | 56.63 | 38.95        |
|           | X            | $\checkmark$ | 44.86          | 43.67 | 57.34 | 36.97        |
|           | ED           | $\sim$       | 44.20          | 42.37 | 57.21 | 38.01        |
|           | $\checkmark$ | CM           | 43.98          | 41.67 | 56.16 | 38.64        |
|           | $\checkmark$ | $\checkmark$ | 44.17          | 42.22 | 56.76 | 38.93        |

Table 2: Ablation study examined on the Chinese-English translation task. "HCT" and "SCT" individually represent hard constrained training and soft constrained training. "DR" and "QR" denote diversity and quality rewards, respectively. "ED": using edit distance as diversity reward; "CM": using COMET as quality reward.

 $SCT_d$ ) significantly increases divBLEU while decreases BLEU and COMET. We speculate that candidate translations with high diversity but low quality lead to this phenomenon.

3) On the contrary, when we only consider the quality reward (HCT<sub>q</sub> and SCT<sub>q</sub>), the results show high COMET but limited improvements on divBLEU. This is because candidate translations are semantically closer to references under current setting, which may harm the diversity of multi-reference pseudo corpus.

4) Jointly considering both diversity and quality (HCT and SCT) makes a better trade-off between translation diversity and quality, suggesting that both rewards are essential for diverse NMT.

5) When replacing BLEU with edit distance to define diversity reward, we observe the diversity drop of translations (SCT (ED) v.s. SCT). Mean-while, changing reconstruction-BLEU to COMET also harms the translation quality (SCT (CM) v.s. SCT). All these confirm the advantages of our proposed two rewards.

6) We additionally report the BLEU score of Top1 candidate translation set (BLEU (Top1)). Interestingly, BLEU fluctuates more than BLEU (Top1) among different variants, which gives us a hint that the superiority of our frameworks lies in the translation on Top2-TopK variants. We will further explore this problem in Section 4.6.

#### 4.6 Analyses

Furthermore, we propose several hypotheses and experimental analyses for deeper insights to diverse NMT task, therefore explain why and how



Figure 2: Manual Evaluation Results. (a) Basic model equipped with our framework (Tree2Code+SCT) achieves comparable manual evaluation score with conventional NMT model (Transformer). (b) COMET has higher correlation with manual evaluation considering Kendall's Tau coefficient (Kendall, 1938).

our framework benefits the model performance. Specifically, we choose to analyze Transformer, Tree2Code and Tree2Code+SCT on Chinese-English translation task.

**Hypothesis 1** *Model-based metric COMET is more suitable for quality evaluation of diverse translations than n-gram-based metric BLEU.* 

**Analysis** Intuitively, the improvement of translation diversity may cause more mismatched n-grams between hypotheses and references, leading to a drop in n-gram-based metrics, i.e. BLEU. In order to make the evaluation more convincing, we conduct human evaluation on the translation results. Specifically, we randomly sample 300 source sentences from MT04-06 sets, and then use three models to generate diverse translations as humanevaluated cases. Next, three linguistic experts are asked to score (0-5) these translations according to the fluency and the accuracy. Each sentence is evaluated by two experts independently, and will be further reviewed by another expert if the disagreement of the former two experts exceeds 3. From Figure 2 (a), we can observe that Tree2Code+SCT gets higher manual evaluation scores than its basic model (Tree2Code), and yields comparable translation quality to the Transformer baseline.<sup>8</sup>

Furthermore, we employ the Kendall's Tau coefficient  $\tau$  (Kendall, 1938) to quantify the correlation between automatic evaluation and manual evaluation, which is calculated over all the humanevaluated cases and defined as

$$\tau = \frac{2}{m(m-1)}(|C| - |D|).$$
(8)

<sup>&</sup>lt;sup>8</sup>More details of manual evaluations are in Appendix A.



Figure 3: The confidence variances among TopK hypothesis translations at different training steps.

Here, m is the number of human-evaluated cases, |C| is the number of times a metric assigns a higher score to the "better" hypothesis and |D| is the number of times a metric assigns a higher score to the "worse" hypothesis. As illustrated in Figure 2 (b), we find that the  $\tau$  is 0.45 for BLEU and 0.58 for COMET, indicating the latter one is more suitable for evaluating translation quality of diverse NMT models than the former one.

**Hypothesis 2** *Multi-candidate optimization improve translation diversity by reducing the confidence variance among candidate translations.* 

Analysis We serve the predicted translation probability as the confidence of each candidate translation (Nguyen and O'Connor, 2015; Wang et al., 2019), and draw the confidence variance of Top Ktranslations during training in Figure 3. When training on a single-reference corpus (Tree2Code), the confidence variance of TopK translations shows an upward trend as the pre-training step grows. Then, it will keep growing if we fine-tune the diverse NMT model with original training strategy, while starting to decline if using our training strategy. This proves that single-reference training encourages the model to fit the one-hot translation. On the contrary, multi-candidate optimization can reduce the confidence variance, and thus offer NMT model more opportunities to generate diverse translations.

# **Hypothesis 3** *Multi-candidate optimization improves the quality of Top2-TopK translations.*

Analysis We measure the quality of TopK candidate translations using COMET and manual evaluation, respectively. As shown in Figure 4, there exist large quality gaps between Top1 and the rest translations. However, after introducing our framework, the COMET and manual evaluation scores of Top2-TopK translations are improved. This shows that



Figure 4: The COMET scores of Top*K* hypothesis generated by Tree2Code and Tree2Code+SCT.



Figure 5: The COMET and divBLEU scores under different settings of  $K_{train}/K_{infer}$ .  $K_{train}$  represents the reference number during training, while  $K_{infer}$  indicates the candidate translation number during inference.

multi-candidate optimization can provide effective guidance for Top2-TopK candidate translations, thus improving overall quality. <sup>9</sup>

**Hypothesis 4** *More references during training leads to better overall performance.* 

**Analysis** We explore different combinations of reference number for training and candidates translation number for inference on Tree2Code+SCT. as illustrated in Figure 5, each row ( $K_{train}$ ) and column ( $K_{infer}$ ) represents the number of generated references for training and the candidate translation number during inference, respectively. We have several interesting observations:

As for the quality, the COMET scores in upper left triangle ( $K_{train} \ge K_{infer}$ ) are higher than those in lower right triangle ( $K_{train} \le K_{infer}$ ). This suggests that references for training should be more than candidates generated during inference for sufficient guidance. As for the diversity, it is obvious that the divBLEU scores in the upper right triangle are also higher than those in the lower left triangle. That is, divBLEU raises as  $K_{infer}$ 

<sup>&</sup>lt;sup>9</sup>From another point of view, the existing optimization on mini-batch is a local fit to single-reference training data, while multi-candidate optimization narrows such quality gap by affecting the distribution of training data.

|               | 依 巴拉 告诉 今日 新闻 电视台 说,「这                           |  |  |
|---------------|--|--|--|
| Src           | 是一个恐怖夜晚 。  |  |  |
| Ref           | Ibarra told today 's news television station : " |  |  |
|               | This is a horrible night . "                     |  |  |
| Transformer   | Ibarra told today 's news television station, "  |  |  |
|               | This is a terrorist evening . "                  |  |  |
|               | Ibarra told today 's news television station , " |  |  |
|               | This is a terrible evening . "                   |  |  |
|               | Ibarra told today 's news television station     |  |  |
|               | that " This is a terrorist evening . "           |  |  |
| Tree2Code     | Ibarra told today 's news television station     |  |  |
|               | that " This is a terrible night . "              |  |  |
|               | According to a barra, today 's news televisior   |  |  |
|               | station said : " This is a terrible night . "    |  |  |
|               | This is a terrible night, according to a news    |  |  |
|               | television station today.                        |  |  |
| Tree2Code+SCT | This is a terrible night according to Ibarra     |  |  |
|               | told today 's news TV station .                  |  |  |
|               | Ibarra told today 's news television station, "  |  |  |
|               | This is a terrible night . "                     |  |  |
|               | Speaking to news TV today, Ibarra said, "        |  |  |
|               | This is a terrible night . "                     |  |  |

Figure 6: An example of NIST Chinese-English diverse translation.

and  $K_{train}$  grow. However, the improvements of diversity gradually become marginal.

# 5 Case Study

From Figure 6, we can see that there are only some simple substitutions (highlighted in blue) in Transformer's results. Tree2Code generates more diverse translations, while containing more mis-translation and under-translation problems (highlighted in red). After applying our framework, Tree2Code+SCT generates better translations in terms of both diversity and quality.

# 6 Conclusion

In this paper, we first point out that the widely used single-reference training is not the preferred solution for diverse NMT. It causes discrepancy between training and inference, and prevents the model from learning one-to-many mapping relationships. Consequently, we propose a novel multicandidate optimization framework which is modelirrelevant and can be compatible with any diverse NMT model. Empirical results suggest that: 1) Multi-candidate optimization is an universally effective manner on boosting the performance of diverse NMT; 2) Model-based metrics can better reflect the translation quality than its n-gram-based counterpart under diverse NMT context; 3) Multicandidate optimization offers NMT abilities to reduce the confidence variance and improve the translation quality of candidate translations.

# Acknowledgements

The project was supported by National Natural Science Foundation of China (No. 62036004, No. 61672440), Natural Science Foundation of Fujian Province of China (No. 2020J06001), Youth Innovation Fund of Xiamen (No. 3502Z20206059), National Key Research and Development Program of China (No. 2018YFB1403202), and Alibaba Group through Alibaba Innovative Research Program. We also thank the reviewers for their insightful comments.

# References

- Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *ICLR 2017*.
- Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. In *ACL* 2020, pages 5961–5970.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018a. Understanding back-translation at scale. In *EMNLP 2018*, pages 489–500.
- Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018b. Classical structured prediction losses for sequence to sequence learning. In *NAACL-HLT 2018*, pages 355–364.
- Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *ACL 2020*, pages 639–649.
- Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *NeurIPS 2016*, pages 820–828.
- Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In COLING 2018, pages 1234–1245.
- Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.
- Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.
- Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *EMNLP 2020*, pages 82–89.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP* 2004, pages 388–395.

- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In ACL 2007, pages 177–180.
- Marie-Anne Lachaux, Armand Joulin, and Guillaume Lample. 2020. Target conditioning for one-to-many generation. In *Findings, EMNLP 2020*, pages 2853– 2862.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.
- Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.
- Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. In *ACL 2021*, pages 4008–4018.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP 2015*, pages 1412–1421.
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *ACL* 2020, pages 4984–4997.
- Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 Metrics Shared Task. In WMT 2020, pages 688–725.
- Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *ACL 2018*, pages 312–318.
- Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *EMNLP 2015*, pages 1587–1598.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.
- Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR 2016*.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP 2020*, pages 2685–2702.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *ACL* 2020, pages 7881–7892.

- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *ACL 2015*, pages 1715–1725.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL 2016*, pages 86–96.
- 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* 2016.
- Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *ICML 2019*, pages 5719–5728.
- Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In ACL 2019, pages 1823–1827.
- Aaron Smith, Christian Hardmeier, and Joerg Tiedemann. 2016. Climbing Mont BLEU: The Strange World of Reachable High-BLEU Translations. In *EAMT 2016*, pages 269–281.
- Jinsong Su, Zhixing Tan, Deyi Xiong, Rongrong Ji, Xiaodong Shi, and Yang Liu. 2017. Lattice-based recurrent neural network encoders for neural machine translation. In *AAAI 2017*, volume 31.
- Zewei Sun, Shujian Huang, Hao-Ran Wei, Xinyu Dai, and Jiajun Chen. 2020. Generating diverse translation by manipulating multi-head attention. In *AAAI* 2020, pages 8976–8983.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS 2014*, pages 3104–3112.
- Zhixing Tan, Jinsong Su, Boli Wang, Yidong Chen, and Xiaodong Shi. 2018. Lattice-to-sequence attentional neural machine translation models. *Neurocomputing*, 284:138–147.
- 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 *NeurIPS 2017*, pages 5998–6008.
- Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In AAAI 2018, pages 7371–7379.
- Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F. Won, and Lidia S. Chao. 2022. UniTE: Unified Translation Evaluation. In *ACL* 2022.
- Yu Wan, Baosong Yang, Derek F Wong, Lidia S Chao, Haihua Du, and Ben CH Ao. 2020. Unsupervised neural dialect translation with commonality and diversity modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9130–9137.

- Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *ACL 2020*, pages 3544–3552.
- Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *EMNLP-IJCNLP 2019*, pages 791–802.
- John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: training neural machine translation with semantic similarity. In ACL 2019, pages 4344–4355.
- Lijun Wu, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2017. Sequence prediction with unlabeled data by reward function learning. In *IJCAI 2017*, pages 3098–3104.
- Xuanfu Wu, Yang Feng, and Chenze Shao. 2020. Generating diverse translation from model distribution with dropout. In *EMNLP 2020*, pages 1088–1097.
- Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. In *EMNLP* 2018, pages 3188–3197.

#### **A** Manual Evaluation Details

#### A.1 Score Definition

We define the quality in manual evaluation as follows: **1 - Totally incomprehensible.** The content is confused and most of the target is left untranslated or unintelligible. **2 - Bad.** Only a small part of target sentence can be understood, specific details are unintelligible, target is very poor in terms of readability or fluency. **3 - Neither good nor bad.** Translation has notable fluency and readability issues, but it is understandable overall. **4 - Good.** It is grammatically correct, but could be better in terms of style and readability. **5 - Very good.** It equals quality of human translation. Only a few minor issues (like capitalization), that don't affect the readability of the target, are allowed.

#### **A.2** Results of Top*K* Hypotheses



Figure 7: Manual scores of TopK hypotheses.

As illustrated in Figure 7, our framework (Tree2Code+SCT) leads to higher manual scores than basic model (Tree2Code) in terms of Top1-TopK hypotheses, which is consistent with the overall results in Figure 2 (a).