Cross-model Back-translated Distillation for Unsupervised Machine Translation

Xuan-Phi Nguyen [1 2] **Shafiq Joty** [1 3] **Thanh-Tung Nguyen** [1 2] **Wu Kui** [2] **Ai Ti Aw** [2]

Abstract

Recent unsupervised machine translation (UMT) systems usually employ three main principles: initialization, language modeling and iterative back-translation, though they may apply them differently. Crucially, iterative back-translation and denoising auto-encoding for language modeling provide data diversity to train the UMT systems. However, the gains from these diversification processes have seemed to plateau. We introduce a novel component to the standard UMT framework called Cross-model Back-translated Distillation (CBD), that is aimed to induce another level of data diversification that existing principles lack. CBD is applicable to all previous UMT approaches. In our experiments, CBD achieves the state of the art in the WMT'14 English-French, WMT'16 English-German and English-Romanian bilingual unsupervised translation tasks, with BLEU scores of 38.2, 30.1, and 36.3, respectively. It also yields 1.5 – 3.3 BLEU improvements in IWSLT English-French and English-German tasks. Through extensive experimental analyses, we show that CBD is effective because it embraces data diversity while other similar variants do not.

1. Introduction

Machine translation (MT) is a core task in natural language processing that involves both language understanding and generation. Recent neural approaches (Vaswani et al., 2017; Wu et al., 2019) have advanced the state of the art with near human-level performance (Hassan et al., 2018). However, they continue to rely heavily on large parallel data. As a result, the search for unsupervised alternatives using only monolingual data has been active. While Ravi & Knight (2011) and Klementiev et al. (2012) proposed various unsupervised techniques for statistical MT (SMT), Lample et al. (2018a;c) established a general framework for modern unsupervised MT (UMT) that works for both SMT and neural MT (NMT) models. The framework has three main principles: model initialization, language modeling and iterative back-translation. Model initialization bootstraps the model with a knowledge prior like word-level cross-lingual transfer (Lample et al., 2018b). Language modeling, which takes the form of denoising auto-encoding (DAE) in NMT (Lample et al., 2018c), trains the model to generate plausible sentences in a language. Meanwhile, iterative back-translation (IBT) facilitates cross-lingual translation training by generating noisy source sentences for original target sentences. The recent approaches differ on how they apply each of these three principles. For instance, Lample et al. (2018a) use an unsupervised word translation model (Lample et al., 2018b) for model initialization, while Conneau & Lample (2019) use a pre-trained cross-lingual masked language model (XLM).

In this paper, we focus on a different aspect of the UMT framework, namely, its *data diversification* process. In this context, we refer data diversification as only sentence level variations, and not contextual topics or genres. If we look from this view, the DAE and IBT steps of the UMT framework also perform some form of data diversification to train the model. Specifically, the noise model in the DAE process generates *new*, but noised, versions of the input data, which are used to train the model with a reconstruction objective. Likewise, the IBT step involves the same UMT model to create synthetic parallel pairs (with the source being synthetic), which are then used to train the model. Since the NMT model is updated with DAE and IBT simultaneously, the model generates *fresh* translations in each back-translation step. Overall, thanks to DAE and IBT, the model gets better at translating by iteratively training on the newly created and diversified data whose quality also improves over time. This argument also applies to statistical UMT, except for the lack of the DAE (Lample et al., 2018c). However, we conjecture that these diversification methods may have reached their limit as the performance does not improve further the longer we train the UMT models.

[1]Nanyang Technological University [2]Institute for Infocomm Research (I[2]R), A*STAR [3]Salesforce Research Asia. Correspondence to: Xuan-Phi Nguyen <nguyenxu002@e.ntu.edu.sg>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

In this work, we introduce a fourth principle to the standard UMT framework: Cross-model Back-translated Distillation[1] or CBD (§3), with the aim to induce another level of diversification that the existing UMT principles lack. CBD initially trains two bidirectional UMT agents (models) using existing approaches. Then, one of the two agents translates the monolingual data from one language s to another t in the first level. In the second level, the generated data are back-translated from t to s by the *other agent*. In the final step, the synthetic parallel data created by the first and second levels are used to distill a supervised MT model. Crucially, the second level agent should be a different one from the first level (hence the name, 'cross-model'). CBD is applicable to any existing UMT method and is more efficient than ensembling approaches (Freitag et al., 2017) (§5.3).

In the experiments (§4), CBD establishes the state of the art (SOTA) in the bilingual unsupervised translation tasks of WMT'14 English-French, WMT'16 English-German and WMT'16 English-Romanian, with 38.2, 30.1 and 36.3 BLEU, respectively. Without large scale pretrained models and data, our method shows consistent improvements of 1.0-2.0 BLEU compared to the baselines in these tasks. It also boosts the performance on IWSLT'14 English-German and IWSLT'13 English-French tasks significantly. In our analysis, we explain with experiments why other similar variants (§5.1) and other alternatives from the literature (§5.4) do not work well and cross-model back-translation is crucial for our method. We further demonstrate that CBD enhances the baselines by achieving greater diversity as measured by back-translation BLEU (§5.2).

2. Background

Ravi & Knight (2011) were among the first to propose a UMT system by framing the problem as a *decipherment* task that considers non-English text as a cipher for English. Nonetheless, the method is limited and may not be applicable to the current well-established NMT systems (Luong et al., 2015; Vaswani et al., 2017; Wu et al., 2019). Lample et al. (2018a) set the foundation for modern UMT. They propose to maintain two encoder-decoder networks simultaneously for both source and target languages, and train them via denoising auto-encoding, iterative back-translation and adversarial training. In their follow-up work, Lample et al. (2018c) formulate a common UMT framework for both Phrase-based SMT (PBSMT) and NMT with three basic principles that can be customized. Specifically, the three main principles of UMT are:

- **Initialization**: A non-randomized cross- or multi-lingual initialization that represents a knowledge prior to bootstrap the UMT model. For instance, Lample et al. (2018a) and Artetxe et al. (2019) use an unsupervised word-translation model MUSE (Lample et al., 2018b) as initialization to promote word-to-word cross-lingual transfer. Lample et al. (2018c) use a shared jointly trained sub-word (Sennrich et al., 2016b) dictionary. On the other hand, Conneau & Lample (2019) use a pretrained cross-lingual masked language model (XLM) to initialize the unsupervised NMT model.
- **Language modeling**: Training a language model on monolingual data helps the UMT model to generate fluent texts. The neural UMT approaches (Lample et al., 2018a;c; Conneau & Lample, 2019) use denoising auto-encoder training to achieve language modeling effects in the neural model. Meanwhile, the PBSMT variant proposed by Lample et al. (2018c) uses the KenLM smoothed n-gram language models (Heafield, 2011).
- **Iterative back-translation**: Back-translation (Sennrich et al., 2016a) brings about the bridge between source and target languages by using a backward model that translates data from target to source. The (source and target) monolingual data is translated back and forth iteratively to progress the UMT model in both directions.

During training, the initialization step is conducted once, while the denoising and back-translation steps are often executed in an alternating manner.[2] It is worth noting that depending on different implementations, the parameters for backward and forward components may be separate (Lample et al., 2018a) or shared (Lample et al., 2018c; Conneau & Lample, 2019). A parameter-shared cross-lingual NMT model has the capability to translate from either source or target, while a UMT system with parameter-separate models has to maintain two models. Either way, we deem a standard UMT system to be bidirectional, i.e., it is capable of translating from either source or target language.

Our proposed cross-model back-translated distillation (CBD) works outside this well-established framework. It employs two UMT agents to create extra diversified data apart from what existing methods already offer, rendering it a useful add-on to the general UMT framework. Furthermore, different implementations of UMT as discussed above can be plugged into the CBD system to achieve a performance boost, even for future methods that may potentially not employ the three principles.

3. Cross-model Back-translated Distillation

In this section, we explain our CBD method in more details. Specifically, let $\mathbb{X}_s$ and $\mathbb{X}_t$ denote the two sets of monolingual data for languages s and t, respectively.

[1] Code: https://github.com/nxphi47/multiagent_crosstranslate.

[2] The KenLM language model in PBSMT (Lample et al., 2018c) was kept fixed during the training process.

We first train two UMT agents independently with two different parameter sets θ_1 and θ_2 using existing methods (Lample et al., 2018a;c; Conneau & Lample, 2019).[3] Since a UMT agent with parameter set $\theta_i \in \{\theta_1, \theta_2\}$ is deemed *bidirectional* in our setup, we denote $y_t \sim P(\cdot|x_s, \theta_i)$ to be a translation sample from language s to t of input sentence x_s using model θ_i. Assuming $\Theta = \{\theta_1, \theta_2\}$, we then define $x_s \sim \mathbb{X}_s$, $y_t \sim P(\cdot|x_s, \theta_\alpha)$ and $z_s \sim P(\cdot|y_t, \theta_\beta)$ to be a sample x_s from $\mathbb{X}_s$, a translation of x_s to language t using model θ_α, and a translation of y_t back to language s using θ_β, respectively, with θ_α being either θ_1 or θ_2 and $\theta_\beta = \Theta \setminus \theta_\alpha$. Note that in this formulation, the model θ_α that produces y_t is different from the one θ_β that produces z_s. Similarly, we define $x_t \sim \mathbb{X}_t$, $y_s \sim P(\cdot|x_t, \theta_\alpha)$ and $z_t \sim P(\cdot|y_s, \theta_\beta)$ in the same manner for $\mathbb{X}_t$. Figure 1 further illustrates this process.

With these generated samples, we train a *supervised* MT model parameterized by θ to maximize the joint probabilities of the aforementioned six random variables, *i.e.,* x_s, y_t, z_s, x_t, y_s and z_t. Equivalently, we minimize the following derived objective function:

$$\begin{aligned}\mathcal{J}(\theta) = \frac{1}{2}\Big[& -\log P_\theta(y_t|z_s) - \log P_\theta(y_t|x_s) \\ & -\log P_\theta(z_s|y_t) - \log P_\theta(x_s|y_t) - \log P_\theta(y_s|z_t) \\ & -\log P_\theta(y_s|x_t) - \log P_\theta(z_t|y_s) - \log P_\theta(y_s|x_t)\Big]\end{aligned} \quad (1)$$

Mathematical derivations and detailed explanations of objective $\mathcal{J}(\theta)$ are further given in the Appendix. Considering the sampling process of x_s, y_s, z_s, x_t, y_t and z_t, the model θ is trained by minimizing the following CBD loss function:

$$\mathcal{L}_\theta(\theta_\alpha, \theta_\beta) = \mathop{\mathbb{E}}_{\substack{z_s \sim P(\cdot|y_t,\theta_\beta), y_t \sim P(\cdot|x_s,\theta_\alpha), x_s \sim \mathbb{X}_s \\ z_t \sim P(\cdot|y_s,\theta_\beta), y_s \sim P(\cdot|x_t,\theta_\alpha), x_t \sim \mathbb{X}_t}} [\mathcal{J}(\theta)] \quad (2)$$

where $\theta_\alpha, \theta_\beta \in \Theta$ are the given UMT models; θ_α is used to generate y_t and y_s from x_s and x_t respectively, while θ_β is used to back-translate y_t and y_s to z_s and z_t respectively. Algorithm 1 describes the overall CBD training process, where the ordered pair $(\theta_\alpha, \theta_\beta)$ is alternated between (θ_1, θ_2) and (θ_2, θ_1)

To describe the CBD strategy more conceptually, in each iteration step of Algorithm 1, each agent $\theta_\alpha \in \{\theta_1, \theta_2\}$ generates translations from the monolingual data $\mathbb{X}_s$ and $\mathbb{X}_t$ of both languages s and t to acquire the *first level* of synthetic parallel data (x_s, y_t) and (x_t, y_s). In the *second level*, the other agent $\theta_\beta = \{\theta_1, \theta_2\} \setminus \theta_\alpha$ is used to generate the translation z_s of the translation y_t of x_s (and similarly for z_t from the translation y_s of x_t). This process is basically *back-translation*, but with the backward model coming from a different regime than that of the forward model. The fact that the first level agent must be different from the second level agent is crucial to achieve the desirable level of diversity in data generation. After this, we update the model θ using all the synthetic pairs (x, y) and (y, z) using the objective function defined in Equation (1).

[3]For neural approaches, changing the random seeds would do the trick, while PBSMT methods would need to randomize the initial embeddings and/or subsample the training data.

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow[s\to t]{\theta_\alpha} y_t \xrightarrow[t\to s]{\theta_\beta} z_s \xrightarrow[(y_t,z_s),(z_s,y_t)]{(x_s,y_t),(y_t,x_s)} \theta$$

$$\mathbb{X}_t \dashrightarrow x_t \xrightarrow[t\to s]{\theta_\alpha} y_s \xrightarrow[s\to t]{\theta_\beta} z_t \xrightarrow[(y_s,z_t),(z_t,y_s)]{(x_t,y_s),(y_s,x_t)} \theta$$

Figure 1: The sampling process of $x_s, y_t, z_s, x_t, y_s, z_t$. The variable ordered set $(\theta_\alpha, \theta_\beta)$ is replaced with (θ_1, θ_2) and (θ_2, θ_1) iteratively in Algorithm 1. All synthetic parallel pairs are used to train θ in a supervised way.

Algorithm 1 Cross-model Back-translated Distillation: Given monolingual data $\mathbb{X}_s$ and $\mathbb{X}_t$ of languages s and t, return a UMT model with parameters θ.

```
Train the 1st UMT agent with parameters θ_1
Train the 2nd UMT agent with parameters θ_2
Initialize model θ (randomly or with pretrained model)
while until convergence do
    θ ← θ − η∇_θ L_θ(θ_α = θ_1, θ_β = θ_2)
    θ ← θ − η∇_θ L_θ(θ_α = θ_2, θ_β = θ_1)
end while
return θ
```

In this way, firstly, the model θ gets trained on the translated products $\{(x_s \leftrightarrow y_t), (x_t \leftrightarrow y_s)\}$ of the UMT teachers, making it as capable as the teachers. Secondly, the model θ is also trained on the second-level data $\{(y_t \leftrightarrow z_s), (y_s \leftrightarrow z_t)\}$ which is slightly different from the first-level data. Thus, this mechanism provides extra data diversification to the system θ in addition to what the UMT teachers already offer, resulting in our final model outperforming the UMT baselines (§4). However, one may argue that since θ_1 and θ_2 are trained in a similar fashion, z will be the same as x, resulting in a duplicate pair. In our experiments, on the contrary, the back-translated dataset contains only around 14% duplicates across different language pairs, as shown in our analysis on data diversity in §5.2.

In the Appendix, we provide a more generalized version of CBD with n (≥ 2) UMT agents, where we also analyze its effectiveness on the IWSLT translation tasks.

4. Experiments

We present our experiments on the large scale WMT (§4.1) and base WMT (§4.2) tasks, followed by IWSLT (§4.3).

4.1. Large Scale WMT Experiments

Setup. We use the codebase from Conneau & Lample (2019) and follow exactly their model setup. Specifically, we use all of the monolingual data from 2007-2017 WMT News Crawl datasets, which yield 190M, 78M, 309M and 3M sentences for language English (En), French (Fr), German (De) and Romanian (Ro), respectively. We filter out sentences whose lengths are over 175 tokens. For each language pair, we build a jointly bilingual dictionary of 60K sub-word units using Byte-Pair Encoding (Sennrich et al., 2016b). To save computation resources, we reuse the pretrained XLM (Conneau & Lample, 2019) and MASS[4] (Song et al., 2019) UMT finetuned models as our two initial models θ_1 and θ_2, respectively. We initialize the CBD supervised MT model θ with the pretrained XLM model provided by Conneau & Lample (2019) for En-Fr and De-En pairs and the pretrained MASS model from Song et al. (2019) for En-Ro pairs, both of which are Transformers with 6 layers and 1024 model dimensions. We train the model with a 2K tokens per batch on a 8-GPU system. Like all previous work, we evaluate the models using the tokenized Moses *multi-bleu.perl* script (Koehn et al., 2007).

Results. Table 1 shows the performance of CBD in comparison with recent UMT methods. Our method establishes the SOTA in the WMT unsupervised tasks with 38.2, 35.5, 30.1, 36.3, 36.3 and 33.8 BLEU for the large scale WMT En-Fr, Fr-En, En-De, De-En, En-Ro and Ro-En tasks, respectively. This translates to up to 1.8 BLEU improvements over the previous SOTA (Song et al., 2019). More interestingly, given that the hard work in training the teacher and initial models θ_1, θ_2 and θ has been done by Conneau & Lample (2019) and Song et al. (2019), our CBD requires a fraction of additional resources to outperform the baselines. This is illustrated in Figure 2, where CBD only needs around 20K updates to converge while the baseline XLM requires up to 200K updates to converge.

4.2. Base WMT Experiments

Setup. Since the results in §4.1 may have been influenced by large scale data and pretrained models, we then seek to evaluate the effectiveness of our CBD method in scenarios where none of the above conveniences are provided. Specifically, we use the News Crawl 2007-2008 datasets for English (En), French (Fr) and German (De), and News Crawl 2015 dataset for Romanian (Ro), and limit the total number of sentences per language to 5M. This is, in fact, the default data setup in the code provided by Lample et al. (2018c); Conneau & Lample (2019). For the NMT models, we follow Lample et al. (2018c) to train the UMT agents with a parameter-shared Transformer (Vaswani et al., 2017) that has 6 layers and 512 dimensions and a batch size of 32 sentences. We use joint Byte-Pair Encoding (BPE) (Sennrich et al., 2016b) and train fastText (Bojanowski et al., 2017) on the BPE tokens to initialize the token embeddings. For the PBSMT (Koehn et al., 2003) models, following Lample et al. (2018c), we use MUSE (Lample et al., 2018b) to generate the initial phrase table and run 4 iterations of back-translation. We subsample 500K sentences from the 5M monolingual sentences at each iteration to train the PBSMT models.[5] For XLM (Conneau & Lample, 2019), we follow the same setup as described in §4.1, except that we pretrain and finetune the XLM model from scratch on the base 5M dataset. We choose the best model based on validation loss and use a beam size of 5. We use a 4-GPU system to train the models. To ensure randomness in the PBSMT agents, we use different seeds for MUSE training and randomly subsample different sets of data during training. To achieve the same for neural agents (NMT and XLM), we simply use different seeds to initialize the models and sample batches of training data.

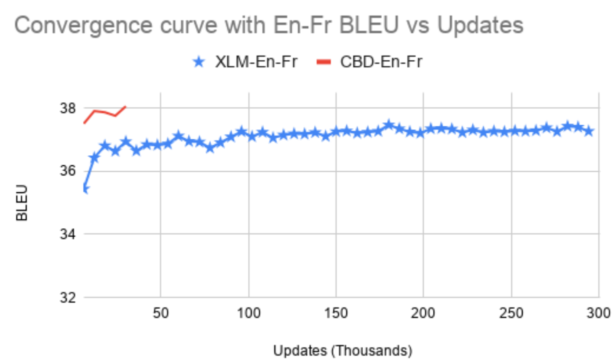

Figure 2: Convergence speed of CBD in comparison with baseline XLM, represented by BLEU score on the WMT'14 En-Fr testset after a number of training updates. Analyses of other languages are given in the Appendix.

Results. Table 2 shows the experimental results of different UMT approaches with and without CBD. First of all, with the datasets that are 30-50 times smaller than the ones used in §4.1, the baselines perform around 2 to 3 BLEU worse than the scores reported in Lample et al. (2018c); Conneau & Lample (2019) (see Table 1). As shown, the

[4]MASS outperforms XLM in our Romanian-related experiments.

[5]For PBSMT, Lample et al. (2018c) subsampled 5M out of 193M sentences of monolingual data.

Table 1: BLEU scores on the *large scale* WMT'14 English-French (En-Fr), WMT'16 English-German (En-De) and WMT'16 English-Romanian (En-Ro) unsupervised translation tasks.

Method / Data	**En-Fr**	**Fr-En**	**En-De**	**De-En**	**En-Ro**	**Ro-En**
NMT (Lample et al., 2018c)	25.1	24.2	17.2	21.0	21.1	19.4
PBSMT (Lample et al., 2018c)	27.8	27.2	17.7	22.6	21.3	23.0
Multi-agent dual learning (Wang et al., 2019)	—	—	19.3	23.8	—	—
XLM (Conneau & Lample, 2019)	33.4	33.3	26.4	34.3	33.3	31.8
MASS (Song et al., 2019)	37.5	34.9	28.3	35.2	35.2	33.1
CBD	**38.2**	**35.5**	**30.1**	**36.3**	**36.3**	**33.8**

Table 2: BLEU scores on the *base* WMT'14 English-French (En-Fr), WMT'16 English-German (En-De) and WMT'16 English-Romanian (En-Ro) unsupervised translation tasks.

Method	**En-Fr**	**Fr-En**	**En-De**	**De-En**	**En-Ro**	**Ro-En**
Data	5M	5M	5M	5M	3M	3M
NMT	24.7	24.5	14.5	18.2	16.7	16.3
+ CBD	26.6	25.7	16.6	20.5	18.1	17.8
PBSMT	17.1	16.4	10.9	13.6	10.5	11.7
+ CBD	21.6	20.6	15.0	17.7	11.3	14.5
XLM	33.0	31.5	23.9	29.3	30.6	27.9
+ CBD	35.4	33.0	26.1	31.5	32.2	29.2

CBD-enhanced model with the pretrained XLM achieves 35.4 and 33.0 BLEU on the WMT'14 En-Fr and Fr-En tasks respectively, which are 2.4 and 1.5 BLEU improvements over the baseline. It also surpasses Conneau & Lample (2019) by 2.0 BLEU in En-Fr task, despite the fact that their model was trained with 274M combined bilingual sentences (compared to our setup of 10M sentences). CBD also boosts the scores for XLM in En-De, De-En, En-Ro, Ro-En by around 2.0 BLEU. For the NMT systems, CBD also outperforms the baselines by 1 to 2 BLEU. More interestingly, PBSMT models are known to be deterministic, but CBD is still able to improve data diversity and provide performance boost by up to 4.0 BLEU points.

4.3. IWSLT Experiments

We also demonstrate the effectiveness of CBD on relatively small datasets for IWSLT En-Fr and En-De translation tasks. The IWSLT'13 En-Fr dataset contains 200K sentences for each language. We use the IWSLT15.TED.tst2012 set for validation and the IWSLT15.TED.tst2013 set for testing. The IWSLT'14 En-De dataset contains 160K sentences for each language. We split it into 95% for training and 5% for validation, and we use IWSLT14.TED.{dev2010, dev2012, tst2010,tst1011, tst2012} for testing. For these experiments, we use the neural UMT method (Lample et al., 2018c) with a Transformer of 5 layers and 512 model dimensions, and trained using only 1 GPU.

From the results in Table 3, we can see that CBD improves the performance in all the four tasks by 2-3 BLEU compared to the NMT baseline of (Lample et al., 2018c).

Table 3: BLEU scores on the unsupervised IWSLT'13 English-French (En-Fr) and IWSLT'14 English-German (En-De) tasks.

Method	**En-Fr**	**Fr-En**	**En-De**	**De-En**
NMT	29.6	30.7	15.8	19.1
+ CBD	31.8	31.8	18.4	21.7

5. Understanding CBD

5.1. Cross-model Back-translation is Key

As mentioned, crucial to our strategy's success is the cross-model back-translation, where the agent operating at the first level must be different from the one in the second level. To verify this, we compare CBD with similar variants that do not employ the cross-model element in the WMT tasks. We refer to these variants commonly as *back-translation distillation* (BD). The first variant *BD(1,1)* has only 1 UMT agent that translates the monolingual data only once and uses these synthetic parallel pairs to distill the model θ. The second variant *BD(1,2)* employs 2 UMT agents, similar to CBD, to produce 2 sets of synthetic parallel data from the monolingual data and uses both of them for distillation. Finally, the third variant *BD(2,2)* uses 2 UMT agents to sample translations from the monolingual data in forward and backward directions using the same respective agents. In other words, BD(2,2) follows similar procedures in Algorithm 1, except that it optimizes the following loss with θ_α being alternated between θ_1 and θ_2:

$$\overline{\mathcal{L}_\theta}(\theta_\alpha) = \mathop{\mathbb{E}}_{\substack{z_s \sim P(\cdot|y_t,\theta_\alpha), y_t \sim P(\cdot|x_s,\theta_\alpha), x_s \sim \mathbb{X}_s \\ z_t \sim P(\cdot|y_s,\theta_\alpha), y_s \sim P(\cdot|x_t,\theta_\alpha), x_t \sim \mathbb{X}_t}} [\mathcal{J}(\theta)] \quad (3)$$

From the comparison in Table 4, we see that none of the BD variants noticeably improves the performance across the language pairs, while CBD provides consistent gains of 1.0-2.0 BLEU. In particular, the BD(1,1) variant fails to improve as the distilled model is trained on the same synthetic data that the UMT agent is already trained on. The variant BD(1,2) is in fact similar in sprit to (Nguyen et al., 2020), which improves supervised and semi-supervised MT. However, it fails to do so in the unsupervised setup, due to the lack of supervised agents. The variant BD(2,2) also fails because the 2nd level synthetic data is already optimized during iterative back-translation training of the UMT agents, leaving the distilled model with no extra information to exploit. Meanwhile, cross-model back-translation enables CBD to translate the second-level data by an agent other than the first-level agent. In this strategy, the second agent produces targets that the first agent is not aware of, while the second agent receives as input the sources that are foreign to it. This process creates corrupted but new information, which the supervised MT model can leverage to improve the overall MT performance through more data diversity.

Table 4: BLEU comparison of CBD vs. no cross-model variants in the *base* WMT'14 English-French (En-Fr), WMT'16 English-German (En-De) and English-Romanian (En-Ro) tasks.

Method	En-Fr	Fr-En	En-De	De-En	En-Ro	Ro-En
NMT	24.7	24.5	14.5	18.2	16.7	16.3
BD(1/1)	24.5	24.5	14.0	17.5	16.1	15.9
BD(1/2)	24.6	24.6	14.1	17.8	16.4	16.2
BD(2/2)	24.8	24.7	14.4	18.1	16.9	16.4
CBD	**26.6**	**25.7**	**16.6**	**20.5**	**18.1**	**17.8**

5.2. CBD Produces Diverse Data

Having argued that cross-model back-translation creates extra information for the supervised MT model to leverage on, we hypothesize that such extra information can be measurable by the diversity of the generated data. To measure this, we compute the *reconstruction BLEU* and compare the scores for BD(2,2) and CBD in the WMT En-Fr, En-De and En-Ro tasks. The scores are obtained by using the first agent to translate the available monolingual data in language s to t and then the second agent to translate those translations back to language s. After that, a BLEU score is measured by comparing the reconstructed text with the original text. In BD(2,2), the first and second agents are identical, while they are distinct for CBD. From the results in Table 5, we observe that the reconstruction BLEU scores of CBD are more than 10 points lower than those of BD, indicating that the newly generated data by CBD are more diverse and different from the original data.

Table 5: Reconstruction BLEU scores of BD and CBD in different languages for the *base* WMT unsupervised translation tasks. Lower BLEU means more diverse.

Method	En-Fr	Fr-En	En-De	De-En	En-Ro	Ro-En
BD	76.0	72.4	75.3	63.7	73.2	71.5
CBD	63.1	59.7	60.3	50.5	61.1	56.9

In Table 6, we further report the ratio of duplicate source-target pairs to the amount of synthetic parallel data created by CBD. We sample 30M synthetic parallel data using the CBD strategy and examine the amount of duplicate pairs for the WMT En-Fr, En-De and En-Ro tasks. We can notice that across the language pairs, only around 14% of the parallel data are duplicates. Given that only about 5M (3.5M for En-Ro) sentences are *real* sentences and the remaining 25M sentences are synthetic, this amount of duplicates is surprisingly low. This fact also explains why CBD is able to exploit extra information better than any standard UMT to improve the performance.

Table 6: Comparison between the amount of real data, generated data by CBD and the duplicates per language pair for the *base* WMT'14 En-Fr, WMT'16 En-De and En-Ro unsupervised MT tasks.

Method	En-Fr	En-De	En-Ro
Real data	5M	5M	3.5M
Generated data	30M	30M	29M
Duplicate pairs	4.4M (14.5%)	3.8M(12.7%)	3.9M (13.4%)

5.3. Comparison with Ensembles of Models and Ensemble Knowledge Distillation

Since CBD utilizes outputs from two UMT agents for supervised distillation, it is interesting to see how it performs compared to an ensemble of UMT models and ensemble knowledge distillation (Freitag et al., 2017). To perform ensembling, we average the probabilities of the two UMT agents at each decoding step. For ensemble distillation, we generate synthetic parallel data from an ensemble of UMT agents, which is then used to train the supervised model.

From the results on the WMT translation tasks in Table 7, we observe that ensembles of models improve the performance only by 0.5-1.0 BLEU, while CBD provides larger gains (1.0-2.0 BLEU) across all the tasks. These results demonstrate that CBD is capable of leveraging the potentials of multiple UMT agents better than how an ensemble of agents does. This is in contrast to data diversification (Nguyen et al., 2020), which is shown to mimic and perform similarly to model ensembling. More importantly, during

Table 7: BLEU comparison of CBD vs. an ensemble of UMT agents and ensemble knowledge distillation (Freitag et al., 2017) on *base* WMT'14 En-Fr, WMT'16 En-De and En-Ro translation tasks.

Method	En-Fr	Fr-En	En-De	De-En	En-Ro	Ro-En
NMT Baseline	24.7	24.5	14.5	18.2	16.7	16.3
Ensemble of 2 agents	25.2	24.8	15.3	19.1	17.7	17.1
Ensemble distillation	17.3	20.0	3.5	3.7	1.2	1.1
CBD	**26.6**	**25.7**	**16.6**	**20.5**	**18.1**	**17.8**

Table 8: Comparison with other alternatives on the *base* WMT En-Fr, Fr-En, En-De and De-En, with XLM as the base model.

WMT	En-Fr	Fr-En	En-De	De-En
XLM	33.0	31.5	23.9	29.3
Sampling (temp=0.3)	33.5	32.2	24.3	30.2
Top-k sampling	33.18	32.26	24.0	29.9
Top-p sampling		Diverge		
Target noising	32.8	30.7	24.0	29.6
Multi-agent dual learning	33.5	31.7	24.6	29.9
CBD	35.4	33.0	26.1	31.5

inference, an ensemble of models requires more memory and computations (twice in this case) to store and execute multiple models. In contrast, CBD can throw away the UMT teacher agents after training and needs only one single model for inference. Meanwhile, ensemble knowledge distillation (Freitag et al., 2017), which performs well with supervised agents, performs poorly in unsupervised MT tasks. The reason could be that the UMT agents may not be suitable for the method originally intended for supervised learning. Further inspection in the Appendix suggests that many samples in the ensemble translations contain incomprehensible repetitions.

5.4. Comparison with Other Potential Alternatives

In this section, we compare CBD with other alternatives in the text generation literature that also attempt to increase diversity. While many of these methods are generic, we adopt them in the UMT framework and compare their performance with our CBD technique in the WMT En-Fr, Fr-en, En-De, and De-en tasks, taking the XLM (Conneau & Lample, 2019) as the base model.

One major group of alternatives is *sampling* based methods, where the model samples translations following multinomial distributions during iterative back-translation. Specifically, we compare the CBD with (*i*) sampling with temperature 0.3 (Edunov et al., 2018; Fan et al., 2018), (*ii*) top-k sampling (Radford et al., 2019), and (*iii*) nucleus or top-p sampling (Holtzman et al., 2020). Plus, we compare CBD with *target noising*, where we add random noises to the translations of the UMT model during iterative back-translation. Finally, multi-agent dual learning (Wang et al., 2019) is also considered as another alternative, where multiple unsupervised agents are used to train the end supervised model.

The results are reported in Table 8. We can see that while the sampling based methods indeed increase the diversity significantly, they do not improve the performance as much as CBD does. The reason could be that the extra data generated by (stochastic) sampling are noisy and their quality is not as good as deterministic predictions from the two UMT agents via cross-model back-translation. On the other hand, target noising does not provide a consistent improvement while multi-agent dual learning achieves less impressive gains compared to CBD.

5.5. Translationese Effect

It can be seen that our cross-model back-translation method is indeed a modified version of back-translation (Sennrich et al., 2016a). Therefore, it is necessary to test if this method suffers from the *translationese effect* (Edunov et al., 2020). As pointed out in their work, back-translation only shows performance gains with translationese source sentences but does not improve when the sentences are natural text.[6] Nguyen et al. (2020) show that the translationese effect only exhibits in a semi-supervised setup, where there are both parallel and monolingual data. However, while they show that their supervised back-translation technique is not impacted by the translationese effect, they left out the question whether unsupervised counterparts are affected.

Therefore, we test our unsupervised CBD method against the translationese effect by conducting the same experiment. More precisely, we compare the BLEU scores of our method versus the XLM baseline (Conneau & Lample, 2019) in the WMT'14 English-German test sets in the three setups devised by Edunov et al. (2020):

- Natural source → translationese target ($X \to Y^*$).
- Translationese source → natural target ($X^* \to Y$)
- Translationese of translationese of source to translationese of target ($X^{**} \to Y^*$).

Table 9 shows that our method outperforms the baseline significantly in the natural source → translationese target scenario ($X \to Y^*$), while it may not improve the

[6]Translationese is human translation of a natural text by a professional translator. Translationese tends to be simpler, more grammatically correct, but lacks contextual sentiments and fidelity.

Table 9: BLEU scores of CBD and the baseline (Conneau & Lample, 2019) on the translationese effect (Edunov et al., 2020), in the *base* WMT'14 English-German setup.

WMT'14 En-De	$X \rightarrow Y^*$	$X^* \rightarrow Y$	$X^{**} \rightarrow Y^*$
XLM Baseline	18.63	18.01	25.59
CBD	20.40	18.31	27.72

translationese source scenario ($X^* \rightarrow Y$) considerably. The results demonstrate that our method behaves differently than what the translationese effect indicates. More importantly, the translations of the natural source sentences are improved, which indicates the practical usefulness of our method. Furthermore, in line with the findings in Nguyen et al. (2020), the experiment shows that the translationese effect may only exhibit in a semi-supervised setup, but not in supervised or unsupervised setups.

6. Related Work

The first step towards utilizing the vast monolingual data to boost MT quality is through semi-supervised training. Back-translation (Sennrich et al., 2016a; Edunov et al., 2018) is an effective approach to exploit target-side monolingual data. Dual learning (He et al., 2016; Wang et al., 2019), meanwhile, trains backward and forward models concurrently and intertwines them together. Recently, Zheng et al. (2020) proposed a variational method to couple the translation and language models through a shared latent space. There have also been attempts in solving low-resource translation problems with limited parallel data (Gu et al., 2018; Irvine & Callison-Burch, 2014; Guzmán et al., 2019). Mohiuddin et al. (2021) propose a contextualized LM based data augmentation for neural machine translation and show its advantages over traditional back-translation gaining improved performance in low-resource scenarios. In the realm of SMT, cross-lingual dictionaries have been used to reduce parallel data reliance (Irvine & Callison-Burch, 2016; Klementiev et al., 2012).

In recent years, unsupervised word-translation via cross-lingual word embedding has seen a huge success (Lample et al., 2018b; Artetxe et al., 2017; 2018a). This opened the door for UMT methods that employ the three principles described in §2. Lample et al. (2018a) and Artetxe et al. (2018c) were among the first of this kind, who use denoising autodecoder for language modeling and iterative back-translation. Lample et al. (2018a) use MUSE word translation (Lample et al., 2018b) as the initialization to bootstrap the model, while Artetxe et al. (2018c) use the VecMap cross-lingual word embeddings (Artetxe et al., 2017). Lample et al. (2018c) later suggested the use of BPE (Sennrich et al., 2016b) and fastText (Bojanowski et al., 2017) to initialize the model and the parameters sharing. Pretrained language models (Devlin et al., 2019) are then used to initialize the entire network (Conneau & Lample, 2019). Song et al. (2019) proposed to pretrain an encoder-decoder model while Artetxe et al. (2018b) suggested a combination of PBSMT and NMT with subword information.[7] Plus, pretraining BART (Lewis et al., 2020) on multi-lingual corpora improves the initialization process (Liu et al., 2020).

Our proposed CBD works outside the three-principle UMT framework and is considered as an add-on to any underlying UMT system. There exist some relevant approaches to CBD. First, it is similar to Nguyen et al. (2020), which generates a diverse set of data from multiple *supervised* MT agents. Despite being effective in supervised and semi-supervised settings, a direct implementation of it in UMT underperforms due to lack of supervised signals (§5.1). In order to successfully exploit unsupervised agents, CBD requires cross-model back-translation which is the key to its effectiveness.

Second, CBD can be viewed as an augmentation technique (Fadaee et al., 2017; Wang et al., 2018). Although the denoising autoencoding built in a typical UMT system also performs augmentation, the noising process is rather naive, while CBD augments data by well-trained agents. Sampling based methods are also considered data diversification strategies, where the model samples translation tokens not by greedy selection (taking $\arg\max$ of probabilities), but by a predefined multinomial distribution. Simple sampling with temperature is often used in many text generation tasks (Edunov et al., 2018; Fan et al., 2018). More advanced top-k sampling (Radford et al., 2019) is used in GPT-2, where a subset of the vocabulary is selected and re-scaled to compute probabilities. Meanwhile, top-p sampling (Holtzman et al., 2020) is used to tackle text degeneration. Our CBD method draws a clear distinction from these methods in that the presumed extra synthetic data is generated not by a random stochastic process, but by well-trained models through the cross-translation procedure.

Third, CBD is related to ensembling (Perrone & Cooper, 1992) and ensemble knowledge distillation (Kim & Rush, 2016; Freitag et al., 2017). Ensembling (Perrone & Cooper, 1992) refers to a type of inference strategies, where multiple differently trained models are used to predict the output probabilities given an input, which are then averaged out to acquire the final output. Ensemble knowledge distillation (Kim & Rush, 2016; Freitag et al., 2017), meanwhile, use multiple models to perform ensemble inference to generate one-way synthetic targets from the original source data, which are then used to distill the final model. The major

[7] Since Artetxe et al. (2018b) did not provide the code, we were unable to apply CBD to their work.

difference between our method and the aforementioned ensembling methods is that they seek to produce the most accurate translations with less variance, while ours seeks to produce as much diverse data as possible. Along with the fact that these distillation schemes are currently applied to supervised settings only, the results in Table 7 indicate that they may not be suitable for unsupervised MT. Similar to our method, multi-agent dual learning (Wang et al., 2019) also uses multiple models in both forward and backward directions, but the data is generated in an ensembling style and its objective to minimize the reconstruction losses instead of to generate diverse synthetic data.

7. Conclusion

We have proposed cross-model back-translated distillation (CBD) - a method that works outside the three existing principles for unsupervised MT and is applicable to any UMT methods. CBD establishes the state of the art in the unsupervised WMT'14 English-French, WMT'16 English-German and English-Romanian translation tasks. It also outperforms the baselines in the IWSLT'14 German-English and IWSLT'13 English-French tasks by up to 3.0 BLEU. Our analysis shows that CBD embraces data diversity and extracts more model-specific intrinsic information than what an ensemble of models would do.

Acknowledgements

We deeply appreciate the efforts of our anonymous reviewers and meta-reviewer in examining and giving us feedback on our paper. We also thank Prathyusha Jwalapuram for proofreading the paper. Xuan-Phi Nguyen is supported by the A*STAR Computing and Information Science (ACIS) scholarship, provided by the Agency for Science, Technology and Research Singapore (A*STAR). Shafiq Joty would like to thank the funding support from NRF (NRF2016IDM-TRANS001-062), Singapore.

References

Artetxe, M., Labaka, G., and Agirre, E. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, 2017.

Artetxe, M., Labaka, G., and Agirre, E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, 2018a. URL https://www.aclweb.org/anthology/P18-1073/.

Artetxe, M., Labaka, G., and Agirre, E. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3632–3642, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1399. URL https://www.aclweb.org/anthology/D18-1399.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. Unsupervised neural machine translation. In *International Conference on Learning Representations*, 2018c. URL https://openreview.net/forum?id=Sy2ogebAW.

Artetxe, M., Labaka, G., and Agirre, E. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 194–203, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1019. URL https://www.aclweb.org/anthology/P19-1019.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Conneau, A. and Lample, G. Cross-lingual language model pretraining. In Wallach, H., Larochelle, H., Beygelzimer, A., dÁlché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 7059–7069. Curran Associates, Inc., 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Edunov, S., Ott, M., Auli, M., and Grangier, D. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://www.aclweb.org/anthology/D18-1045.

Edunov, S., Ott, M., Ranzato, M., and Auli, M. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2836–2846, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.

acl-main.253. URL https://www.aclweb.org/anthology/2020.acl-main.253.

Fadaee, M., Bisazza, A., and Monz, C. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 567–573, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2090. URL https://www.aclweb.org/anthology/P17-2090.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://www.aclweb.org/anthology/P18-1082.

Freitag, M., Al-Onaizan, Y., and Sankaran, B. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*, 2017.

Gu, J., Hassan, H., Devlin, J., and Li, V. O. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*, 2018.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1632. URL https://www.aclweb.org/anthology/D19-1632.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567, 2018. URL http://arxiv.org/abs/1803.05567.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., and Ma, W.-Y. Dual learning for machine translation. In *Advances in neural information processing systems*, pp. 820–828, 2016.

Heafield, K. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pp. 187–197. Association for Computational Linguistics, 2011.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Irvine, A. and Callison-Burch, C. Hallucinating phrase translations for low resource mt. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 160–170, 2014.

Irvine, A. and Callison-Burch, C. End-to-end statistical machine translation with zero or small parallel texts. *Natural Language Engineering*, 22(4):517–548, 2016.

Kim, Y. and Rush, A. M. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://www.aclweb.org/anthology/D16-1139.

Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 130–140. Association for Computational Linguistics, 2012.

Koehn, P., Och, F. J., and Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pp. 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073462. URL https://doi.org/10.3115/1073445.1073462.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pp. 177–180, 2007.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=rkYTTf-AZ.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. In *International Conference on Learning Representations*, 2018b. URL https://openreview.net/forum?id=H196sainb.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, Brussels, Belgium, October-November 2018c. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549. URL https://www.aclweb.org/anthology/D18-1549.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020.

Luong, T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1412–1421. ACL, 2015.

Mohiuddin, T., Bari, M. S., and Joty, S. Augvic: Exploiting bitext vicinity for low-resource nmt. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, 2021. Association for Computational Linguistics.

Nguyen, X.-P., Joty, S., Kui, W., and Aw, A. T. Data diversification: A simple strategy for neural machine translation. In *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc., 2020.

Perrone, M. P. and Cooper, L. N. When networks disagree: Ensemble methods for hybrid neural networks. Technical report, BROWN UNIV PROVIDENCE RI INST FOR BRAIN AND NEURAL SYSTEMS, 1992.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Ravi, S. and Knight, K. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 12–21, 2011.

Sennrich, R., Haddow, B., and Birch, A. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725. Association for Computational Linguistics, 2016b. doi: 10.18653/v1/P16-1162. URL http://www.aclweb.org/anthology/P16-1162.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Wang, X., Pham, H., Dai, Z., and Neubig, G. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 856–861, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1100. URL https://www.aclweb.org/anthology/D18-1100.

Wang, Y., Xia, Y., He, T., Tian, F., Qin, T., Zhai, C., and Liu, T.-Y. Multi-agent dual learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGhN2A5tm.

Wu, F., Fan, A., Baevski, A., Dauphin, Y., and Auli, M. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkVhlh09tX.

Zheng, Z., Zhou, H., Huang, S., Li, L., Dai, X.-Y., and Chen, J. Mirror-generative neural machine translation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxQRTNYPH.