Universal Conditional Masked Language Pre-training for Neural Machine Translation

Pengfei Li[1] **Liangyou Li**[1] **Meng Zhang**[1] **Minghao Wu**[2] **Qun Liu**[1]
[1]Huawei Noah's Ark Lab [2]Monash University
{lipengfei111, liliangyou, zhangmeng92, qun.liu}@huawei.com
minghao.wu@monash.edu

Abstract

Pre-trained sequence-to-sequence models have significantly improved Neural Machine Translation (NMT). Different from prior works where pre-trained models usually adopt an unidirectional decoder, this paper demonstrates that pre-training a sequence-to-sequence model but with a bidirectional decoder can produce notable performance gains for both Autoregressive and Non-autoregressive NMT. Specifically, we propose CeMAT, a conditional masked language model pre-trained on large-scale bilingual and monolingual corpora in many languages.[1] We also introduce two simple but effective methods to enhance the CeMAT, *aligned code-switching & masking* and *dynamic dual-masking*. We conduct extensive experiments and show that our CeMAT can achieve significant performance improvement for all scenarios from low- to extremely high-resource languages, i.e., up to +14.4 BLEU on low-resource and +7.9 BLEU on average for Autoregressive NMT. For Non-autoregressive NMT, we demonstrate it can also produce consistent performance gains, i.e., up to +5.3 BLEU. To the best of our knowledge, this is the first work to pre-train a unified model for fine-tuning on both NMT tasks.

1 Introduction

Pre-trained language models have been widely adopted in NLP tasks (Devlin et al., 2019; Radford and Narasimhan, 2018). For example, XLM (Conneau and Lample, 2019) demonstrated that cross-lingual pre-training is effective in improving neural machine translation (NMT), especially on low-resource languages. These methods all directly pre-train a bidirectional encoder or an unidirectional decoder. The encoder and decoder in NMT models are then independently initialized with them and fine-tuned (Guo et al., 2020; Zhu et al., 2020). Recently, pre-training standard sequence-to-sequence (Seq2Seq) models has shown significant improvements and become a popular paradigm for NMT tasks (Song et al., 2019; Liu et al., 2020; Lin et al., 2020).

Approach	Enc.	Dec.	Mono.	Para.
mBERT (Devlin et al., 2019)	•		•	
XLM (Conneau and Lample, 2019)	•		•	•
MASS (Song et al., 2019)	•	$\rightarrow$	•	
mBART (Liu et al., 2020)	•	$\rightarrow$	•	
mRASP (Lin et al., 2020)	•	$\rightarrow$		•
CeMAT (Ours)	•	$\Longleftrightarrow$	•	•

Table 1: Comparison and summary of existing pre-trained models for machine translation. Enc: encoder; Dec: decoder; Mono: monolingual; Para: bilingual. "•" denotes the corresponding model is pre-trained or the corresponding data is used. "$\rightarrow$" denotes the decoder of model is unidirectional, "$\Longleftrightarrow$" denotes the decoder is bidirectional.

However, some experimental results from XLM (Conneau and Lample, 2019) have shown that the decoder module initialized by the pre-trained bidirectional masked language model (MLM) (Devlin et al., 2019), rather than the unidirectional causal language model (CLM, Radford and Narasimhan, 2018), would achieve better results on Autoregressive NMT (AT). Especially, compared to random initialization, initialized by GPT (Radford and Narasimhan, 2018) might result in performance degradation sometimes. We conjecture that when fine-tuning on generation tasks (e.g., NMT), the representation capability of the pre-trained models may be more needed than the generation capability. Therefore, during pre-training, we should focus on training the representation capability not only for the encoder, but also for the decoder more explicitly.

Inspired by that, we present CeMAT, a **multilingual Conditional masked language prE-training model for MAchine Translation**, which

[1]Code, data, and pre-trained models are available at https://github.com/huawei-noah/Pretrained-Language-Model/CeMAT

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6379 - 6391
May 22-27, 2022 ©2022 Association for Computational Linguistics

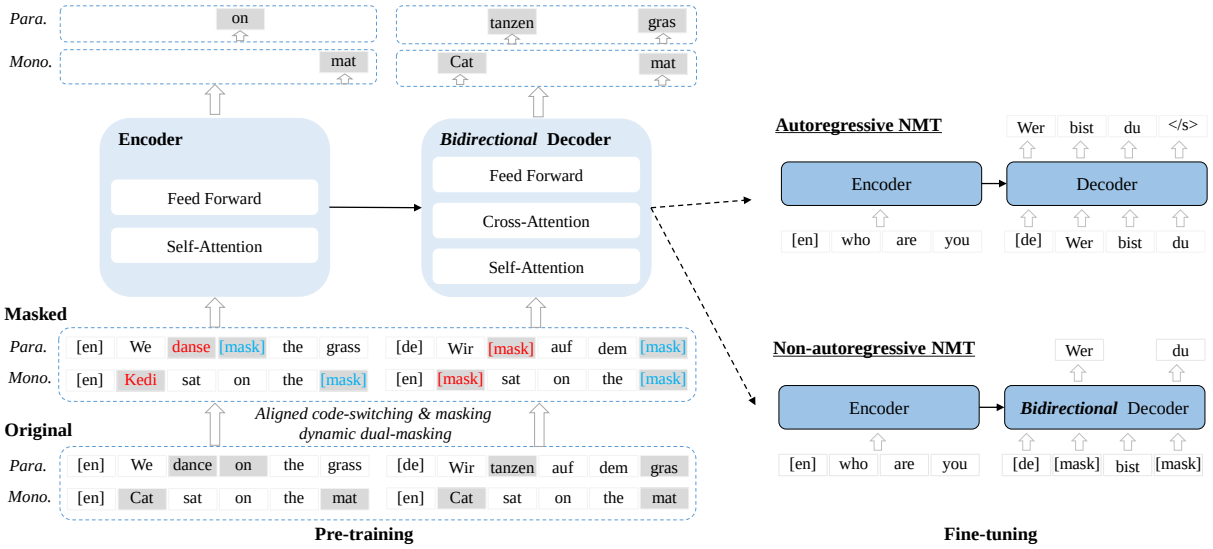

Figure 1: The framework for CeMAT, which consists of an encoder and a ***bidirectional decoder***. "*Mono*" denotes monolingual, "*Para*" denotes bilingual. During the pre-training (left), the original monolingual and bilingual inputs in many languages are augmented (the words are replaced with new words with same semantics or "[mask]", please see Figure 2 for more details) and fed into the model. Finally, we predict all the "[mask]" words on the source side and target side respectively. For fine-tuning (right), CeMAT provides unified initial parameter sets for AT and NAT.

consists of a bidirectional encoder, a *bidirectional* decoder, and a cross-attention module for bridging them. Specifically, the model is jointly trained by MLM on the encoder and Conditional MLM (CMLM) on the decoder with large-scale monolingual and bilingual texts in many languages. Table 1 compares our model with prior works. Benefiting from the structure, CeMAT can provide unified initialization parameters not only for AT task, but also for Non-autoregressive NMT (NAT) directly. NAT has been attracting more and more attention because of its feature of parallel decoding, which helps to greatly reduce the translation latency.

To better train the representation capability of the model, the masking operations are applied in two steps. First, some source words that have been aligned with target words are randomly selected and then substituted by new words of similar meanings in other languages, and their corresponding target words are masked. We call this method *aligned code-switching & masking*. Then, the remaining words in both source and target languages will be masked by *dynamic dual-masking*.

Extensive experiments on downstream AT and NAT tasks show significant gains over prior works. Specifically, under low-resource conditions ($<$ 1M bitext pairs), our system gains up to +14.4 BLEU points over baselines. Even for extremely high-resource settings ($>$ 25M), CeMAT still achieves significant improvements. In addition, experiments on the WMT16 Romanian→English task demonstrate that our system can be further improved (+2.1 BLEU) by the Back-Translation (BT; Sennrich et al., 2016a).

The main contributions of our work can be summarized as follows:

- We propose a multilingual pre-trained model CeMAT, which consists of a bidirectional encoder, a *bidirectional* decoder. The model is pre-trained on both monolingual and bilingual corpora and then used for initializing downstream AT and NAT tasks. To the best of our knowledge, this is the first work to pre-train a unified model suitable for both AT and NAT.

- We introduce a two-step masking strategy to enhance the model training under the setting of bidirectional decoders. Based on a multilingual translation dictionary and word alignment between source and target sentences, *aligned code-switching & masking* is firstly applied. Then, *dynamic dual-masking* is used.

- We carry out extensive experiments on AT and NAT tasks with data of varied sizes. Consistent improvements over strong competitors demonstrate the effectiveness of CeMAT.

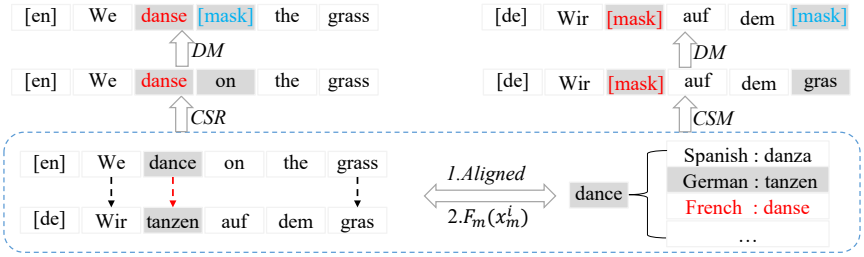

Figure 2: The details of our two-step masking. We first obtain the aligned pair set $\Lambda = \{$("dance","tanzen"),...$\}$ (marked with $\dashrightarrow$) from the original inputs by looking up the cross-lingual dictionary (denote as *1.Aligned*), and then randomly select a subset (marked as "dance"$\dashrightarrow$"tanzen" with red color) from it, in the lower left of the figure. For each element in the subset, we select a new word by $F_m(x_m^i)$, and perform *CSR* to replace the source fragment ("danse" marked as red color) and *CSM* for target ("[mask]" marked as red color) respectively. Finally, we do the DM process to mask the contents of the source and target respectively ("[mask]" marked as light-blue color).

2 Pre-training Approach

Our CeMAT is jointly trained by MLM and CMLM on the source side and the target side, respectively. The overall framework is illustrated in Figure 1. In this section, we first introduce the multilingual CMLM task (Section 2.1). Then, we describe the two-step masking, including the *aligned code-switching & masking* (Section 2.2) and the *dynamic dual-masking* (Section 2.3). Finally, we present training objectives of CeMAT (Section 2.4).

Formally, our training data consists of M language-pairs $D = \{D_1, D_2, ..., D_M\}$. $D_k(m, n)$ is a collection of sentence pairs in language L_m and L_n, respectively. In the description below, we denote a sentence pair as $(X_m, Y_n) \in D_k(m, n)$, where X_m is the source text in the language L_m, and Y_n is the corresponding target text in the language L_n. For monolingual corpora, we create pseudo bilingual text by copying the sentence, namely, $X_m = Y_n$.

2.1 Conditional Masked Language Model

CMLM predicts masked tokens y_n^{mask}, given a source sentence X_m and the remaining target sentence $Y_n \backslash y_n^{mask}$. The probability of each $y_n^j \in y_n^{mask}$ is independently calculated:

$$P(y_n^j | X_m, Y_n \backslash y_n^{mask}). \quad (1)$$

CMLM can be directly used to train a standard Seq2Seq model with a bidirectional encoder, a unidirectional decoder, and a cross attention. However, it is not restricted to the autoregressive feature on the decoder side because of the independence between masked words. Therefore, following practices of NAT, we use CMLM to pre-train a Seq2Seq model with a bidirectional decoder, as shown in Figure 1.

Although bilingual sentence pairs can be directly used to train the model together with the conventional CMLM (Ghazvininejad et al., 2019), it is challenging for sentence pairs created from monolingual corpora because of identical source and target sentences. Therefore, we introduce a two-step masking strategy to enhance model training on both bilingual and monolingual corpora.

2.2 Aligned Code-Switching & Masking

We use *aligned code-switching & masking* strategy to replace the source word or phrase with a new word in another language, and then mask the corresponding target word. Different from the previous code-switching methods (Yang et al., 2020; Lin et al., 2020) where source words always are randomly selected and replaced directly, our method consists of three steps:

1. **Aligning**: We utilize a multilingual translation dictionary to get a set of aligned words $\Lambda = \{\cdots, (x_m^i, y_n^j), \cdots\}$ between the source X_m and target Y_n. The word pair (x_m^i, y_n^j) denotes that the i-th word in X_m and j-th word in Y_n are translations of each other. For sentence pairs created from monolingual corpora, words in an aligned word pair are identical.

2. **Code-Switching Replace (CSR)**: Given an aligned word pair $(x_m^i, y_n^j) \in \Lambda$, we first select a new word $\hat{x}_k^i$ in the language L_k that can be used to replace x_m^i in the source sentence X_m,

$$\hat{x}_k^i = F_m(x_m^i)$$

where $F_m(x)$ is a multilingual dictionary lookup function for a word x in the language L_m, $\hat{x}_k^i$ is a randomly selected word from the

dictionary, which is a translation of x_m^i in the language L_k.

3. **Code-Switching Masking (CSM)**: If the source word x_m^i in the aligned pair (x_m^i, y_n^j) is replaced by $\hat{x}_k^i$, we also mask y_n^j in Y_n by replacing it with a universal mask token. Then, CeMAT will be trained to predict it in the output layers of the bidirectional decoder.

For aligning and CSR, we only use available multilingual translation dictionary provided by MUSE (Lample et al., 2018). Figure 2 shows the process of *aligned code-switching & masking*. According to the given dictionary, "dance" and "tanzen" are aligned, then a new French word "danse" is selected to replace "dance", and "tanzen" replaced by "[mask]" (marked as red color).

During training, at most 15% of the words in the sentence will be performed by CSR and CSM. For monolingual data, we set this ratio to 30%. We use

$$(\mathrm{CSR}(X_m), \mathrm{CSM}(Y_n))$$

to denote the new sentence pair after *aligned code-switching & masking*, which will be further dynamically dual-masked at random.

2.3 Dynamic Dual-Masking

Limited by the dictionary, the ratio of aligned word pairs is usually small. In fact, we can only match aligned pairs for 6% of the tokens on average in the bilingual corpora. To further increase the training efficiency, we perform *dynamic dual-masking* (DM) on both bilingual and monolingual data.

- Bilingual data: We first sample a masking ratio υ from a uniform distribution between $[0.2, 0.5]$, then randomly select a subset of target words which are replaced by "[mask]". Similarly, we select a subset on the source texts and mask them with a ratio of μ in a range of $[0.1, 0.2]$. Figure 2 shows an example of *dynamic dual-masking* on bilingual data. We set $\upsilon \geq \mu$ to force the bidirectional decoder to obtain more information from the encoder.

- Monolingual data: Since the source and target are identical before masking, we sample $\upsilon = \mu$ from a range $[0.3, 0.4]$ and mask the same subset of words on both sides. This will avoid the decoder directly copying the token from the source.

Follow practices of pre-trained language models, 10% of the selected words for masking remain unchanged, and 10% replaced with a random token. Words replaced by the *aligned code-switching & masking* will not be selected to prevent the loss of cross-lingual information. We use

$$(\mathrm{DM}(\mathrm{CSR}(X_m)), \mathrm{DM}(\mathrm{CSM}(Y_n)))$$

to denote the new sentence pair after dynamic dual-masking, which will be used for pre-training.

2.4 Multilingual Pre-training Objectives

We jointly train the encoder and decoder on MLM and CMLM tasks. Given the sentence pair

$$(\hat{X}_m, \hat{Y}_n) = (\mathrm{DM}(\mathrm{CSR}(X_m)), \mathrm{DM}(\mathrm{CSM}(Y_n)))$$

from the masked corpora $\hat{D}$, the final training objective is formulated as follows:

$$\begin{aligned}\mathcal{L} = -\sum_{(\hat{X}_m, \hat{Y}_n) \in \hat{D}} \lambda \sum_{y_n^j \in y_n^{mask}} \log P(y_n^j | \hat{X}_m, \hat{Y}_n) \\ +(1-\lambda) \sum_{x_m^i \in x_m^{mask}} \log P(x_m^i | \hat{X}_m)\end{aligned} \tag{2}$$

where y_n^{mask} are the set of masked target words, x_m^{mask} are the set of masked source words, and λ is a hyper-parameter to balance the influence of both tasks. In our experiments, we set $\lambda = 0.7$.

3 Pre-training Settings

Pre-training Data We use the English-centric multilingual parallel corpora of PC32[2], and then collect 21-language monolingual corpora from common crawl[3]. In this paper, we use ISO language code[4] to identify each language. A "[*language code*]" token will be prepended to the beginning of the source and target sentence as shown in Figure 2. This type of token helps the model to distinguish sentences from different languages. The detailed correspondence and summary of our pre-training corpora can be seen in Appendix A.

Data pre-processing We directly learn a shared BPE (Sennrich et al., 2016b) model on the entire data sets after tokenization. We apply Moses tokenization (Sennrich et al., 2016b) for most languages, and for other languages, we use KyTea[5]

[2] https://github.com/linzehui/mRASP
[3] https://commoncrawl.org/
[4] https://www.loc.gov/standards/iso639-2/php/code_list.php
[5] http://www.phontron.com/kytea/

Lang-Pairs	En-Kk		En-Tr		En-Et		En-Fi		En-Lv		En-Cs	En-De	En-Fr	Avg
Source	WMT19		WMT17		WMT18		WMT17		WMT17		WMT19	WMT19	WMT14	
Size	91k(low)		207k(low)		1.94M(medium)		2.66M(medium)		4.5M(medium)		11M(high)	38M(extr-high)	41M(extr-high)	
Direction	→	←	→	←	→	←	→	←	→	←	→	→	→	
Direct	0.2	0.8	9.5	12.2	17.9	22.6	20.2	21.8	12.9	15.6	16.5	30.9	41.4	17.1
mBART	2.5	7.4	17.8	22.5	21.4	27.8	22.4	28.5	15.9	19.3	18.0	30.5	41.0	21.2
mRASP	8.3	12.3	20.0	23.4	20.9	26.8	24.0	28.0	21.6	**24.4**	19.9	35.2	**44.3**	23.8
CeMAT	**8.8**	**12.9**	**23.9**	**23.6**	**22.2**	**28.5**	**25.4**	**28.7**	**22.0**	24.3	**21.5**	**39.2**	43.7	25.0
Δ	+8.6	+12.1	+14.4	+11.4	+4.3	+5.9	+5.2	+6.9	+9.1	+8.7	+5.0	+8.3	+2.3	+7.9

Table 2: Comprehensive comparison with mRASP and mBART. Best results are highlighted in **bold**. CeMAT outperforms them on AT for all language pairs but two directions. Even for extremely high-resource scenarios(denoted as "extr-high"), we observe gains of up to +8.3 BLEU on En→De language pair.

for Japanese and jieba[6] for Chinese, and a special normalization for Romanian (Sennrich et al., 2016a). Following Liu et al. (2020), we balance the vocabulary size of languages by up/down-sampling text based on their data size when learning BPE.

Model and Settings As shown in Figure 1, we apply a bidirectional decoder so that it can utilize left and right contexts to predict each token. We use a 6-layer encoder and 6-layer bidirectional decoder with a model dimension of 1024 and 16 attention heads. Following Vaswani et al. (2017), we use sinusoidal positional embedding, and apply layer normalization for word embedding and pre-norm residual connection following Wang et al. (2019a).

Our model is trained on 32 Nvidia V100 GPUs for 300K steps, The batch size on each GPU is 4096 tokens, and we set the value of update frequency to 8. Following the training settings in Transformer, we use Adam optimizer ($\epsilon = 1e-6, \beta_1 = 0.9, \beta_2 = 0.98$) and polynomial decay scheduling with a warm-up step of 10,000.

4 Autoregressive Neural Machine Translation

In this section, we verify CeMAT provides consistent performance gains in low to extremely high resource scenarios. We also compare our method with other existing pre-training methods and further present analysis for better understanding the contributions of each component.

4.1 Fine-Tuning Objective

The AT model consists of an encoder and a unidirectional decoder. The encoder maps a source sentence X_m into hidden representations which are then fed into the decoder. The unidirectional decoder predicts the t-th token in a target language L_n conditioned on X_m and the previous target tokens $y_n^{<t}$. The training objective of AT is to minimize the negative log-likelihood:

$$\mathcal{L}(\theta) = \sum_{(X_m, Y_n) \in D(m,n)} \sum_{t=1}^{|Y_n|} -\log P(y_n^t | X_m, y_n^{<t}; \theta) \quad (3)$$

4.2 Experimental Settings

Benchmarks We selected 9 different language pairs and then use CeMAT to fine-tune on them. They are divided into four categories according to their data size: low-resource ($<$ 1M), medium-resource ($>$ 1M and $<$ 10M), high-resource ($>$ 10M and $<$ 25M), and extremely high-resource ($>$ 25M). See Appendix B for more details.

Configuration We adopt a dropout rate of 0.1 for extremely high-resource En→Fr, En→De (WMT19); for all other language pairs, we set the value of 0.3. We fine-tune AT with a maximum learning rate of $5e-4$, a warm-up step of 4000 and label smoothing of 0.2. For inference, we use beam search with a beam size of 5 for all translation directions. For a fair comparison with previous works, all results are reported with case-sensitive and tokenized BLEU scores.

4.3 Results and Analysis

Main Results We fine-tune AT systems initialized by our CeMAT on 8 popular language pairs, which are the overlapping language pairs in experiments of mBART (Liu et al., 2020) and mRASP (Lin et al., 2020). Table 2 shows the results. Compared to directly training AT models, our systems with CeMAT as initialization obtain significant improvements on all four scenarios. We observe gains of up to +14.4 BLEU and over +11.4 BLEU on three of the four tasks on low-resource scenarios, i.e., En↔Tr. Without loss of generality, as the scale of the dataset increases, the benefits of pre-training

[6] https://github.com/fxsjy/jieba

models are getting smaller and smaller. However, we can still obtain significant gains when the data size is large enough (extremely high-resource: $>$ 25M), i.e. +8.3 and +2.3 BLEU for En→De and En→Fr respectively. This notable improvement shows that our model can further enhance extremely high-resource translation. Overall, we obtain performance gains of more than +8.0 BLEU for most directions, and finally observe gains of +7.9 BLEU on average on all language pairs.

We further compare our CeMAT with mBART (Liu et al., 2020) and mRASP (Lin et al., 2020), which are two pre-training methods of current SOTA. As illustrated in Table 2, CeMAT outperforms mBART on all language pairs with a large margin (+3.8 BLEU on average), for extremely high-resource, we can obtain significant improvements when mBART hurts the performance. Compared to mRASP, we achieve better performance on 11 out of the total 13 translation directions, and outperforms this strong competitor with an average improvement of +1.2 BLEU on all directions.

Comparison with Existing Pre-training Models We further compare our CeMAT with more existing multilingual pre-trained models on three popular translation directions, including WMT14 En→De, WMT16 En↔Ro. Results are shown in Table 3. Our CeMAT obtains competitive results on these languages pairs on average, and achieves the best performance on En→Ro.

Our model also outperforms BT (Sennrich et al., 2016a), which is a universal and stable approach to augment bilingual with monolingual data. In addition, when combining back-translation with our CeMAT on Ro →En, we obtain a significantly improvement from 36.8 to 39.0 BLEU, as shown in Table 3. This indicates that our method is complementary to BT.

The Effectiveness of Aligned Code-Switching and Masking We investigate the effectiveness of *aligned code-switching & masking* as shown in Table 4. We find that utilizing *aligned code-switching & masking* can help CeMAT improve the performance for all different scenarios with gains of +0.5 BLEU on average, even though we can only match the aligned word pairs for 6% of the tokens on average in the bilingual corpora. We presume the method can be improved more significantly if we adopt more sophisticated word alignment methods.

The Effectiveness of Dynamic Masking In the pre-training phase, we use a dynamic strategy when doing dual-masking on the encoder and decoder respectively. We verify the effectiveness of this dynamic masking strategy. As illustrated in Table 4 and Appendix C, we achieve significant gains with margins from +0.4 to +4.5 BLEU, when we adjusted the ratio of masking from a static value to a dynamically and randomly selected value. The average improvement on all language pairs is +2.1 BLEU. This suggests the importance of dynamic masking.

Lang-Pairs Size	En → De 4.5M	En → Ro 597K	Ro → En 597K	Ro → En (+BT)
Direct	29.3	34.3	34.0	36.8
mBART	-	37.7	37.8	38.8
mRASP	30.3	37.6	36.9	38.9
MASS	28.9	–	–	39.1
XLM	28.8	–	35.6	38.5
mBERT	28.6	–	–	–
CeMAT	30.0	38.0	37.1	39.0

Table 3: Comparison with recent multilingual pre-training models on WMT14 En→De, WMT16 En↔Ro. We reach comparable results on all three directions. When combining back-translation, we further obtain gains of +2.2 BLEU on Ro→En.

5 Non-autoregressive Neural Machine Translation

In this section, we will verify the performance of our CeMAT on the NAT, which generates translations in parallel, on widely-used translation tasks.

5.1 Fine-Tuning Objective

As illustrated in Figure 1, NAT also adopts a Seq2Seq framework, but consists of an encoder and a bidirectional decoder which can be used to predict the target sequences in parallel. The training objective of NAT is formulated as follows:

$$\mathcal{L}(\theta) = \sum_{(X_m, Y_n) \in D(m,n)} \sum_{t=1}^{|Y_n|} -\log P(y_n^t | X_m; \theta) \tag{4}$$

In this work, we follow Ghazvininejad et al. (2019), which randomly sample some tokens y_n^{mask} for masking from target sentences and train the model by predicting them given source sentences

Lang-Pairs	En-Kk		En-Tr		En-Et		En-Fi		En-Lv		Avg
Direction	→	←	→	←	→	←	→	←	→	←	
CeMAT	8.8	12.9	23.9	23.6	22.2	28.5	25.4	28.7	22.0	24.3	22.0
. w/o *Aligned CS masking*	8.0	12.3	23.6	23.1	22.1	28.0	24.8	28.1	21.4	24.1	21.5
. w/o *Aligned CS masking & Dynamic*	7.2	8.7	21.2	20.4	20.8	26.8	24.4	27.5	16.9	20.2	19.4

Table 4: Verification of the effectiveness of different techniques. ". w/o *Aligned CS masking*" denotes that we pre-train CeMAT without *aligned code-switching & masking* algorithm. ". w/o *Aligned CS masking & Dynamic*" means that we further abandon the use of dynamic setting for dual-masking, where we only use a fixed masking ratio with 0.15 for the encoder and decoder. More details can be found in Appendix C. We can see two methods are all critical components.

and remaining targets. The training objective is:

$$\mathcal{L}(\theta) = \sum_{(X_m, Y_n) \in D(m,n)} \sum_{y_n^j \in y_n^{mask}} -\log P(y_n^j | X_m, Y_n \backslash y_n^{mask}; \theta) \quad (5)$$

During decoding, given an input sequence to translate, the initial decoder input is a sequence of "[mask]" tokens. The fine-tuned model generates translations by iteratively predicting target tokens and masking low-quality predictions. This process can make the model re-predict the more challenging cases conditioned on previous high-confidence predictions.

5.2 Experimental Settings

NAT Benchmark Data We evaluate on three popular datasets: WMT14 En↔De, WMT16 En↔Ro and IWSLT14 En↔De. For a fair comparison with baselines, we only use the bilingual PC32 corpora to pre-train our CeMAT. We only use knowledge distillation (Gu et al., 2018) on WMT14 En↔De tasks.

Baselines We use our CeMAT for initialization and fine-tune a Mask-Predict model (Ghazvininejad et al., 2019) as in Section 4. To better quantify the effects of the proposed pre-training models, we build two strong baselines.

Direct. We directly train a Mask-Predict model with randomly initialized parameters.

mRASP. To verify that our pre-trained model is more suitable for NAT, we use a recently pre-trained model mRASP (Lin et al., 2020) to fine-tune on downstream language pairs.

Configuration We use almost the same configuration as the pre-training and AT except the following differences. We use learned positional embeddings (Ghazvininejad et al., 2019) and set the max-positions to 10,000.

5.3 Main Results

The main results on three language pairs are presented in Table 5. When using CeMAT to initialize the Mask-Predict model, we observe significant improvements (from +0.9 to +5.3 BLEU) on all different tasks, and finally obtain gains of +2.5 BLEU on average. We also achieve higher results than the AT model on both En→De (+2.8 BLEU) and De→En (+0.9 BLEU) directions on IWSLT14 datasets, which is the extremely low-resource scenarios where training from scratch is harder and pre-training is more effective.

As illustrated in Table 5, on all different tasks, CeMAT outperforms mRASP with a significant margin. On average, we obtain gains of +1.4 BLEU over mRASP. Especially under low-resource settings on IWSLT14 De→En, we achieve a large gains of +3.4 BLEU over mRASP. Overall, mRASP shows limited improvement (+0.4 to +1.9 BLEU) compared to CeMAT. This also suggests that although we can use the traditional pre-training method to fine-tune the NAT task, it does not bring a significant improvement like the AT task because of the gap between pre-training and fine-tuning tasks.

We further compare the dynamic performance on three language pairs during iterative decoding, as shown in Appendix D. We only need 3 to 6 iterations to achieve the best score. During the iteration, we always maintain rapid improvements. In contrast, mRASP obtains the best result after 6 to 9 iterations. We also observe a phenomenon that the performance during iterations is also unstable on both mRASP and Mask-Predict, but CeMAT appears more stable. We conjecture that our pre-trained model can learn more related information between words in both the same and different languages. This ability alleviated the drawback of NAT assumptions: the individual token predictions are conditionally independent of each other.

Source	IWSLT14		WMT16		WMT14		Avg
Lang-Pairs	En→De	De→En	En→Ro	Ro→En	En→De	De→En	
Transformer (Vaswani et al., 2017)	23.9	32.8	34.1	34.5	28.0	32.7	31.0
Mask-Predict (Ghazvininejad et al., 2019)	22.0	28.4	31.5	31.7	26.1	29.0	28.1
mRASP (Lin et al., 2020)	23.9	30.3	32.2	32.1	26.7	29.8	29.2
CeMAT (Ours)	**26.7**	**33.7**	**33.3**	**33.0**	**27.2**	**29.9**	30.6

Table 5: Comprehensive comparison with two strong baselines. "mRASP" denotes using mRASP to initialize Mask-Predict, "CeMAT (Ours)" denotes using our CeMAT to initialize. We obtain consistent and significant improvements on all language pairs, outperforming AT on IWSLT14 tasks. Best non-autoregressive results are highlighted in **bold**.

6 Related Work

Multilingual Pre-training Task Conneau and Lample (2019) and Devlin et al. (2019) proposed to pre-train a cross-lingual language model on multi language corpora, then the encoder or decoder of model are initialized independently for fine-tuning. Song et al. (2019), Yang et al. (2020) and Lewis et al. (2020) directly pre-trained a Seq2Seq model by reconstructing part or all of inputs and achieve significant performance gains. Recently, mRASP (Lin et al., 2020) and CSP (Yang et al., 2020) apply the code-switching technology to simply perform random substitution on the source side. Another similar work, DICT-MLM (Chaudhary et al., 2020) introduce multilingual dictionary, pre-training the MLM by mask the words and then predict its cross-lingual synonyms. mRASP2 (Pan et al., 2021) also used code-switching on monolingual and bilingual data to improve the effectiveness, but it is essentially a multilingual AT model.

Compared to previous works: 1) CeMAT is the first pre-trained Seq2Seq model with a bidirectional decoder; 2) We introduce aligned code-switching & masking, different from traditional code-switching, we have two additional steps: align between source and target, and CSM; 3) We also introduce a dynamic dual-masking method.

Autoregressive Neural Machine Translation Our work is also related to AT, which adopts an encoder-decoder framework to train the model (Sutskever et al., 2014). To improve the performance, back-translation, forward-translation and related techniques were proposed to utilize the monolingual corpora (Sennrich et al., 2016a; Zhang and Zong, 2016; Edunov et al., 2018; Hoang et al., 2018). Prior works also attempted to jointly train a single multilingual translation model that translates multi-language directions at the same time (Firat et al., 2016; Johnson et al., 2017; Aharoni et al., 2019; Wu et al., 2021). In this work, we focus on pre-training a multilingual language model, which can provide initialization parameters for the language pairs. On the other hand, our method can use other languages to further improve high-resource tasks.

Non-autoregressive Neural Machine Translation Gu et al. (2018) first introduced a transformer-based method to predict the complete target sequence in parallel. In order to reduce the gap with the AT model, Lee et al. (2018) and Ghazvininejad et al. (2019) proposed to decode the target sentence with iterative refinement. Wang et al. (2019b) and Sun et al. (2019) utilized auxiliary information to enhance the performance of NAT. One work related to us is Guo et al. (2020), which using BERT to initialize the NAT. In this work, CeMAT is the first attempt to pre-train a multilingual Seq2Seq language model on NAT task.

7 Conclusion

In this paper, we demonstrate that multilingually pre-training a sequence-to-sequence model but with a bidirectional decoder produces significant performance gains for both Autoregressive and Non-autoregressive Neural Machine Translation. Benefiting from conditional masking, the decoder module, especially the cross-attention can learn the word representation and cross-lingual representation ability more easily. We further introduce the *aligned code-switching & masking* to align the representation space for words with similar semantics but in different languages, then we use a *dynamic dual-masking* strategy to induce the bidirectional decoder to actively obtain the information from the source side. Finally, we verified the effectiveness of these two methods. In the future, we will investigate more effective word alignment method for *aligned code-switching & masking*.

8 Acknowledgments

We would like to thank anonymous reviewers for their helpful feedback. we also thank Wenyong Huang, Lu Hou, Yinpeng Guo, Guchun Zhang for their useful suggestion and help with experiments.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3874–3884. Association for Computational Linguistics.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. DICT-MLM: improved multilingual pre-training using bilingual dictionaries. *CoRR*, abs/2010.12566.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics.

Orhan Firat, KyungHyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *CoRR*, abs/1601.01073.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6111–6120. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020. Incorporating BERT into parallel sequence decoding with adapters. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 18–24. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1173–1182. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 2649–2663. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 244–258. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhi-Hong Deng. 2019. Fast structured decoding for sequence models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3011–3020.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019a. Learning deep transformer models for machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1810–1822. Association for Computational Linguistics.

Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019b. Non-autoregressive machine translation with auxiliary regularization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5377–5384. AAAI Press.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7291–7305. Association for Computational Linguistics.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP: code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2624–2636. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1535–1545. The Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Statistics of the Pre-Training Data.

We present dataset statistics for pre-training corpora in Table 6.

B Statics of Five Different Scenarios

We present dataset statistics for fine-tuning corpora in Table 7.

C Detailed Ablation Experiments

We show more detailed results of the ablation experiments on two language pairs in Table 8.

D Performance with Iterations for NAT

We present the dynamic performance on three language-pair datasets during iterative decoding in Figure 3, 4, 5, 6, 7 and 8.

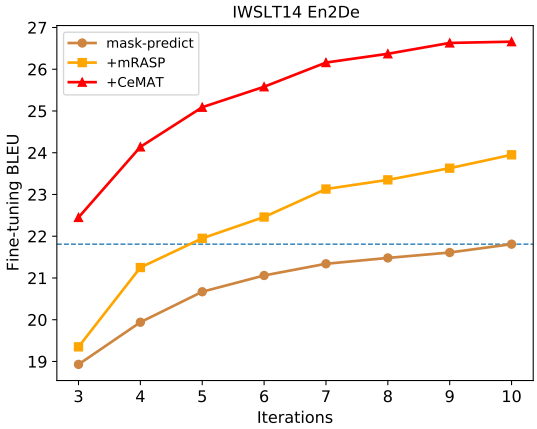

Figure 3: The performance of IWSLT14 En2De when decoding with different number of iterations.

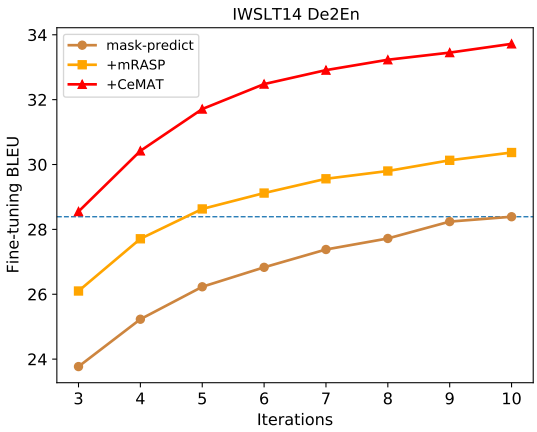

Figure 4: The performance of IWSLT14 De2En when decoding with different number of iterations

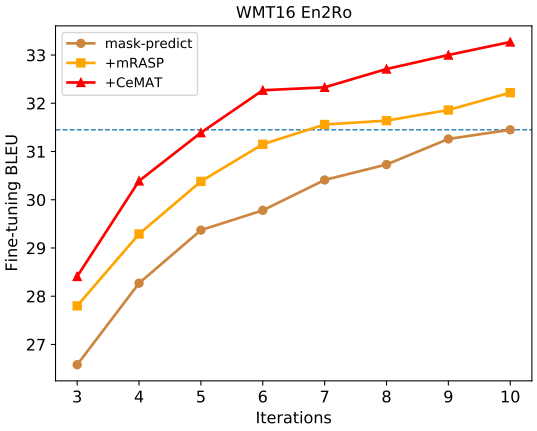

Figure 5: The performance of WMT16 En2Ro when decoding with different number of iterations.

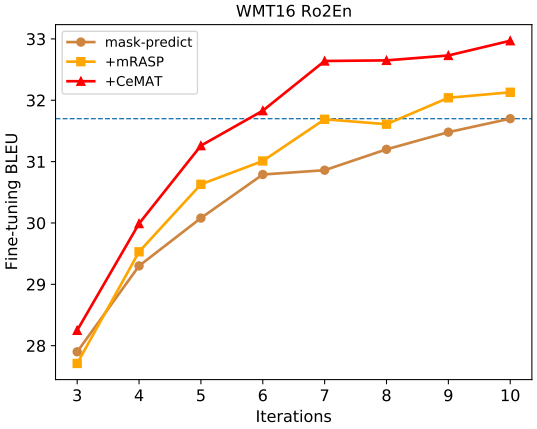

Figure 6: The performance of WMT16 Ro2En when decoding with different number of iterations.

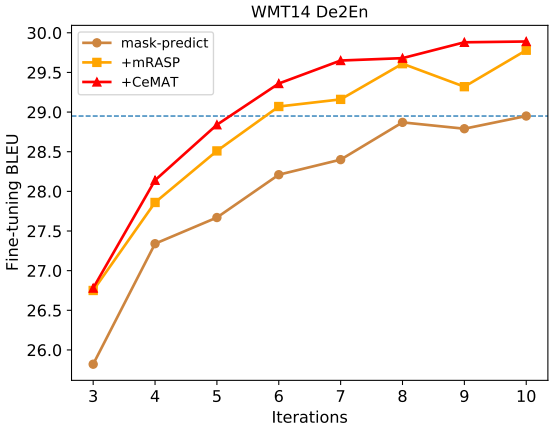

Figure 7: The performance of WMT14 De2En when decoding with different number of iterations.

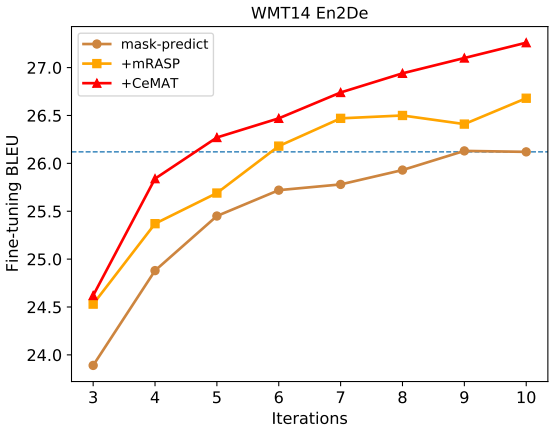

Figure 8: The performance of WMT14 En2De when decoding with different number of iterations.

ISO	Language	Bilingual	Monolingual	ISO	Language	Bilingual	Monolingual
Gu	Gujarati	11K	815K	Ko	Korean	1.4M	–
Be	Belarusian	24K	–	Ms	Malay	1.6M	–
My	Burmese	28K	–	Ru	Russian	1.8M	9.9M
Mn	Mongolian	28K	–	Fi	Finnish	2M	9.9M
Af	Afrikaans	40K	–	Ja	Japanese	2M	3.4M
Eo	Esperanto	66K	–	It	Italian	2M	9.9M
Kk	Kazakh	122K	1.8M	Es	Spanish	2.1M	9.9M
Sr	Serbian	133K	3.7M	Et	Estonian	2.2M	5.3M
Mt	Maltese	174K	–	Lt	Lithuanian	2.3M	2.8M
Ka	Kannada	198K	–	Lv	Latvian	3.0M	11.3M
He	Hebrew	330K	–	Bg	Bulgarian	3.1M	9.9M
Tr	Turkish	383K	9.9M	Vi	Vietnamese	3.1M	–
Ro	Romanian	770K	20M	De	German	4.6M	15M
Cs	Czech	814K	9.9M	Zh	Chinese	21M	4.4M
Ar	Arabic	1.2M	–	Fr	French	36M	15M
El	Greek	1.3M	8.3M	En	English	–	15M
Hi	Hindi	1.3M	9.9M				

Table 6: A list of 32 Enlish-centric language pair datasets. Among them, 21 languages have corresponding monolingual data. In this work, we using the ISO code represent the language name, and put them at the beginning of the source and target.

Lang-Pairs	Source	Size	Category
En-Kk	WMT19	97K	low-resource
De-En	IWSLT14	159K	low-resource
En-Tr	WMT17	207K	low-resource
En-Ro	WMT16	597K	low-resource
En-Et	WMT18	1.9M	medium-resource
En-Fi	WMT17	2.7M	medium-resource
En-Lv	WMT17	4.5M	medium-resource
En-De	WMT14	4.5M	medium-resource
En-Cs	WMT19	11M	high-resource
En-De	WMT19	38M	extremely high-resource
En-Fr	WMT14	41M	extremely high-resource

Table 7: The statistical information of the language pairs on *low- / medium- / high- / extremely high*-resource for the machine translation task.

Lang-Pairs	Kk-En		Et-En		Avg
Direction	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	
w/ *Bilingual*	7.8	5.5	24.4	19.1	14.2
w/ *Monolingual*	5.4	5.4	23.5	18.9	13.3
w/ *Bi- & Monolingual*	9.0	5.6	25.2	19.0	14.7
w/o *Aligned CS masking*	8.4	5.1	24.3	18.2	14.0
w/o *Dynamic (masking:0.15)*	7.3	4.4	23.5	17.7	13.2
w/o *Dynamic (masking:0.35)*	8.8	5.6	23.7	18.1	14.1

Table 8: Verification of the effectiveness of different techniques on two language pairs: Kk-En and Et-En. "w/ *Bilingual*" denotes that we use only bilingual data when pre-training CeMAT; "w/ *Monolingual*" denotes that we use only monolingual data when pre-training CeMAT; "w *Bi- & Monolingual*" denotes that when pre-training CeMAT, we use both bilingual and monolingual data; "w/o *Aligned CS masking*" denotes that we pre-train CeMAT without *aligned code-switching & masking* algorithm; "w/o *Dynamic (masking:0.15)*" means that we use a fixed masking ratio with 0.15 for dual-masking; "w/o *Dynamic (masking:0.35)*" means that we use a fixed masking ratio with 0.35 for dual-masking to make a more fair comparison with dynamic masking. To save computational resources, we use Transformer-base to obtain all the results of this experiment.