Rethinking the Exploitation of Monolingual Data for Low-Resource Neural Machine Translation

Jianhui Pang*
NLP^{2}CT Lab, University of Macau
nlp2ct.pangjh3@gmail.com

Baosong Yang**
Alibaba Group
yangbaosong.ybs@alibaba-inc.com

Derek Fai Wong**
NLP^{2}CT Lab, University of Macau
derekfw@um.edu.mo

Yu Wan
Alibaba Group
wanyu.wy@alibaba-inc.com

Dayiheng Liu
Alibaba Group
liudayiheng.ldyh@alibaba-inc.com

Lidia Sam Chao
NLP^{2}CT Lab, University of Macau
lidiasc@um.edu.mo

Jun Xie
Alibaba Group
qingjing.xj@alibaba-inc.com

* This research was accomplished when Jianhui Pang was interning at Alibaba DAMO Academy.
**Baosong Yang and Derek Fai Wong are co-corresponding authors.

Action Editor: Min Zhang. Submission received: 10 February 2023; revised version received: 30 May 2023; accepted for publication: 15 June 2023.

https://doi.org/10.1162/coli_a_00496

The utilization of monolingual data has been shown to be a promising strategy for addressing low-resource machine translation problems. Previous studies have demonstrated the effectiveness of techniques such as back-translation and self-supervised objectives, including masked language modeling, causal language modeling, and denoise autoencoding, in improving the performance of machine translation models. However, the manner in which these methods contribute to the success of machine translation tasks and how they can be effectively combined remains an under-researched area. In this study, we carry out a systematic investigation of the effects of these techniques on linguistic properties through the use of probing tasks, including source language comprehension, bilingual word alignment, and translation fluency. We further evaluate the impact of pre-training, back-translation, and multi-task learning on bitexts of varying sizes. Our findings inform the design of more effective pipelines for leveraging monolingual data in extremely low-resource and low-resource machine translation tasks. Experiment results show consistent performance gains in seven translation directions, which provide further support for our conclusions and understanding of the role of monolingual data in machine translation.

1. Introduction

The Neural Machine Translation (NMT) model has shown significant improvement in translation tasks when trained on large-scale parallel data. However, low-resource machine translation remains a challenging area of research in the field. To address this, recent studies have explored the utilization of easily collected monolingual data in training low-resource NMT models through techniques such as pre-training (PT), back-translation (BT), and multi-task learning (MTL) (Luong et al. 2016; Edunov et al. 2018; Wang, Zhai, and Hassan 2020). These methods have respectively demonstrated impressive performance, highlighting the potential of exploiting monolingual data for enhancing low-resource machine translation tasks.

The utilization of monolingual data in the low-resource NMT is an active research area. BT synthesizes pseudo bitexts by translating target monolingual text into source language text (Sennrich, Haddow, and Birch 2016b). PT, on the other hand, leverages a large amount of monolingual data through self-supervised objectives such as masked language modeling (MLM), causal language modeling (CLM), and denoise autoencoding (DAE) (Devlin et al. 2019; Liu et al. 2020; Radford et al. 2018, 2019). MTL, meanwhile, combines self-supervised objectives with the translation task to additionally learn monolingual knowledge (Luong et al. 2016; Wang, Zhai, and Hassan 2020; Gulcehre et al. 2015). While these methods have demonstrated promising results, there is a lack of clarity on how they contribute to the improvement of NMT models and how they interact with each other. This article aims to address these gaps by investigating the impact and interplay of these methods on low-resource translation tasks.

In this article, we aim to: (1) examine the impact of BT and the three self-supervised objectives on three key aspects of a translation model, as determined through probing tasks, including source language understanding, bilingual word alignment, and translation fluency, which are three relevant tasks to NMT (Snover et al. 2009; Garg et al. 2019; Zhu et al. 2020; Kong et al. 2021) and proved as three crucial factors for enhancing translation quality in our experiments, and (2) compare the translation performance of MTL methods, PT methods, and BT with various bitext scale settings, including both extremely low-resource and low-resource settings. We specifically consider three widely used self-supervised objectives, namely, MLM, CLM, and DAE. Our results demonstrate that the combination of the three self-supervised objectives

enhances source language understanding, improves bilingual word alignment, and results in competitive translation fluency. Furthermore, BT has a positive effect on both bilingual word alignment and translation fluency. Meanwhile, MTL effectively trains the translation model with self-supervised objectives and shows potential for mitigating the catastrophic forgetting effect (Thompson et al. 2019; Wang, Zhai, and Hassan 2020). By incorporating self-supervised objectives, the multi-task fine-tuning method further improves translation quality. Based on these findings, we explore more effective strategies for utilizing source-side and target-side monolingual data in low-resource translation tasks. Experiment results on seven translation directions, including similar and distant pairs, echo our claims and understandings.

Our main contributions in this article are: (1) We perform a comprehensive investigation on the impact of self-supervised objectives and back-translation on a translation model by conducting extensive linguistic probing tasks, providing a fine-grained analysis of the effects on source language understanding, bilingual word alignment, and translation fluency; (2) We compare and evaluate the translation performance of pre-training (PT), back-translation (BT), and multi-task learning (MTL) methods with various bitext scales, including extremely low-resource and low-resource settings; and (3) Based on the results and our analysis, we further validate our observations by designing pipelines that successfully improve the performance of extremely low-resource and low-resource translation tasks.

2. Preliminaries

2.1 Background

2.1.1 Back-Translation. Back-Translation (BT) is a data synthesis alternative (Bertoldi and Federico 2009; Bojar and Tamchyna 2011) then proposed for improving the NMT model (Sennrich, Haddow, and Birch 2016a). BT requires a reversed NMT model to translate target-side monolingual sentences to source-side sentences and generates pseudo pairs. Previous works have demonstrated the effectiveness of BT in improving translation quality, as it enriches the dataset and provides additional alignment examples for the NMT model (Edunov et al. 2018; Caswell, Chelba, and Grangier 2019; Liu et al. 2021). In this article, we aim to further understand the linguistic effects of BT on the translation model through a comprehensive investigation utilizing three linguistic probing tasks and to evaluate the effectiveness of BT synthetic data for extremely low-resource and low-resource tasks. We note that forward-translation is another data synthesis method, of which the generated target-side sentences are too noisy to enhance the low-resource tasks (Tars, Tättar, and Fišel 2021).

2.1.2 Masked Language Modeling. The masked language model (MLM) is a self-supervised pre-training objective, first introduced by Devlin et al. (2019), designed to train the encoder of a model on monolingual data. The objective of MLM is to randomly mask a percentage of input tokens and then require the model to predict these masked tokens. Previous work has demonstrated the benefits of MLM pre-training on translation tasks (Rothe, Narayan, and Severyn 2019; Zhu et al. 2020). Recently, Wang, Zhai, and Hassan (2020) extended the use of MLM to multi-task learning, resulting in improved translation performance in a multilingual setting. These findings emphasize the importance of source language understanding for improving translation quality.

2.1.3 Causal Language Modeling. The pre-training of the language model via CLM involves maximizing the likelihood of generating text in an auto-regressive manner

(Radford et al. 2018). In the translation domain, pre-trained language models such as GPT2 have been used to initialize the decoder of a transformer model, excluding the cross-attention layer. However, Rothe, Narayan, and Severyn (2020) found that a model initialized with CLM (GPT2) performed comparatively poorly in translation tasks when compared with a model initialized with MLM (BERT). Another recent study (Baziotis, Haddow, and Birch 2020) leverages the pre-trained CLM as an informative prior by adding a regularization term that encourages the output of the translation model to conform to the distributions generated by the language model. Despite these findings, further analysis is needed to better understand the impact of CLM on NMT models.

2.1.4 Denoise Autoencoding. DAE is a sequence-to-sequence (seq-to-seq) self-supervised objective, which trains a model to recover a noisy sentence. DAE has been applied to various tasks in the NLP community, including pre-trained models, multi-task learning, and so on (Kim, Geng, and Ney 2019; Lewis et al. 2020; Liu et al. 2020; Wang, Zhai, and Hassan 2020). To improve DAE, these studies put in efforts to propose various noisy functions for better training a model, showing that DAE effectively learns from a large scale of monolingual data. We are going to study the insight linguistic effects of DAE on a translation model. Following the noise setting of Lample et al. (2017), we adopt three types of noise functions. (1) *Word Shuffle*: slightly shuffles the original sentence. We apply a random permutation σ to the original sentence in condition of $\forall i \in 1, n, |\sigma(i) - i| \leq k$ where n is the sentence length and k is set to 3 by default. (2) *Word Drop*: we drop every word in the original sentence with a constant probability p_{wd}, which we set to 0.1 by default. (3) *Word Blank*, we replace every word in the original sentence to *unk* token with a constant probability p_{wb}, which we set to 0.1 by default.

2.2 Rethinking Monolingual Exploitation

In the field of NLP, three prevalent pre-trained models are BERT (Devlin et al. 2019), GPT2 (Radford et al. 2019), and mBART (Liu et al. 2020), which are trained on the MLM, CLM, and DAE objectives, respectively. These models have demonstrated their ability to effectively learn language knowledge from monolingual data. Previous studies have shown that these pre-trained models can improve few-shot tasks such as low-resource translation (Liu et al. 2020; Rothe, Narayan, and Severyn 2020; Wang, Zhai, and Hassan 2020). Despite the widespread use of these models, there is currently a lack of fair comparisons between them in the context of identical monolingual data and model settings. Another method for leveraging monolingual data in translation models is BT, which uses data synthesis to improve the translation model. However, the difference between the effects of BT and self-supervised objectives on translation models is not yet fully understood.

The encoder and decoder are two essential components in a translation model. The encoder is responsible for understanding the source sentence and generating sentence features, while the decoder generates the target sentence given the source sentence features. In recent studies, the utilization of pre-trained encoders, such as BERT, has been shown to improve translation quality through model initialization and incorporation (Rothe, Narayan, and Severyn 2019; Zhu et al. 2020). These results highlight the significance of source language understanding in facilitating translation tasks. The decoder, on the other hand, attends to the output of the encoder and generates the translation. Research has shown that a translation model can generate accurate word alignment using the cross-attention matrix (Garg et al. 2019; Chen et al. 2020), indicating that word alignment is closely related to the translation task. Additionally, translation

fluency is a crucial metric for evaluating translation quality (Snover et al. 2009). In conclusion, source language understanding, bilingual alignment, and translation fluency are considered as three essential factors for the performance of a translation model. Thus, we propose three probing tasks to evaluate the impact of BT and three self-supervised objectives on these crucial aspects, respectively.

2.3 Experiment Setup

2.3.1 Data. We utilize 5 million monolingual data for each language from the publicly available News-Crawl corpus.[1] Our translation experiments are conducted on widely used translation benchmark datasets, including the WMT and OPUS-100 datasets. We adopt three WMT benchmarks, including the WMT14 English-German (EN-DE), WMT18 English-Turkish (EN-TR), and WMT16 English-Romanian (EN-RO) datasets. To analyze the effects of different methods under different resource settings, we further randomly sample 7 subsets from the EN-DE dataset, including 5k, 10k, and 15k for extremely low-resource settings, and 50k, 100k, 200k, and 500k for low-resource settings, following Gu et al. (2018b). Specifically, the EN-DE 10k dataset contains 10k sentence pairs as a training set, and the other datasets are similarly structured. We randomly sample 10k sentence pairs for the EN-RO dataset to conduct extremely low-resource translation tasks. For the OPUS datasets, we randomly sample 10k and 100k bitexts on four translation pairs from the OPUS-100 dataset (Zhang et al. 2020), which include English-French (EN-FR), English-Russian (EN-RU), English-Arabic (EN-AR), and English-Chinese (EN-ZH) language pairs, and the validation and test sets remain unchanged. We denote each dataset using the translation direction and the number of sentence pairs, such as EN-DE 100k. The details are as follows:

- The EN-DE of 5k, 10k, 15k, 50k, 100k, 200k, and 500k: These seven datasets are subsets of the commonly used WMT14 English-Germany benchmark, which includes 5k, 10k, 15k, 50k, 100k, 200k, and 500k training pairs. We use *newstest2013* and *newstest2014* as the evaluation set and testing set, respectively. For a fair comparison in probing tasks, all these datasets share the same 32k bpe vocabulary following Edunov et al. (2018).
- EN-RO 10k and EN-RO: The evaluation set and testing set are *newstest2015* and *newstest2016*. EN-RO refers to the commonly used benchmark WMT16 English-Romania. EN-RO 10k is a subset of EN-RO, which contains 10k parallel training data. Specifically, we learn 32k bpe vocabulary for EN-RO 10k with monolingual data following Gu et al. (2018a).
- EN-TR: This dataset is a low-resource translation benchmark, WMT18 English-Turkish. The evaluation set and testing set are *Newstest2017* and *Newstest2018*, respectively.
- EN-FR, EN-RU, EN-AR, and EN-ZH: The training sets of these for translation direction are sampled from OPUS-100, and the validation and test sets are unchanged. We sample the sentence with lengths between 10

1 `https://data.statmt.org/news-crawl/`.

and 80. For our experiments, we sample 10k for the extremely low-resource setting and 100k for the low-resource setting.

2.3.2 Probing Task. The three probing tasks, which are source language understanding, bilingual word alignment, and translation fluency, are conducted on finetuned translation models of EN-DE 100k and EN-DE 200k. The details are as followed:

1. For the task of source language understanding, we utilize the SentEval framework, which includes ten classification tasks for a comprehensive analysis (Conneau and Kiela 2018). SentEval is structured based on different linguistic properties, including surface information (Surf.), syntactic information (Sync.), and semantic information (Semc.). More details regarding SentEval can be found in Appendix A. For the downstream tasks, we freeze the parameters of the encoder and add a trainable classification head. We use Accuracy as the evaluation metric.
2. For the bilingual word alignment task, we utilize the alignment test set as provided in Vilar, Popović, and Ney (2006). Our models are not fine-tuned on the alignment task training set. The alignments are induced using the cross-attention weight, as described in Garg et al. (2019). The alignment error rate (AER) is used as the evaluation metric, as adopted in previous work such as Zhang and Zong (2016) and Chen et al. (2020).
3. For the translation fluency task, we assess the quality of translations by computing their perplexity (PPL) using the German-GPT2 language model (Schweter 2020). The fluency of the translations is reported as the average PPL over all the sentences.

2.3.3 Settings. In order to make a fair comparison in our experiments, we use the Transformer base architecture as described in Vaswani et al. (2017). We leverage the *subword-nmt* toolkit to learn bpe subwords and create a joint dictionary for all datasets.[2] We also include language tags in each sentence during training and inference, following the methodology in Liu et al. (2020). We use the Adam optimizer (Kingma and Ba 2015) with specified hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ to optimize the model parameters. The pre-trained model is trained with mini-batches of $32K$ target-language tokens for 150,000 steps. The early-stop strategy is adopted for training the translation tasks with a patience of 10. All experiments were conducted on 4 Nvidia Tesla V100 32GB GPUs, using the *fairseq* toolkit.[3] The noisy functions of the MLM and DAE objectives are applied with default settings as described in previous works (Lample et al. 2017; Devlin et al. 2019). The settings of the four technologies are:

- For BT, we train a reverse translation model to generate 5 million BT synthetic bitexts for each language direction.

2 https://github.com/rsennrich/subword-nmt.
3 https://github.com/facebookresearch/fairseq.

- The MLM objective is achieved by adding an additional output layer Φ_{head} to the encoder and trains the encoder on source-side monolingual data to predict masked tokens.
- The CLM objective is performed by ignoring the cross-attention modules and trains all other decoder modules on target-side monolingual data.
- The DAE objective is achieved by training both the encoder and decoder on source-side and target-side monolingual data.

Given a source and target monolingual sentence, x and y, the model $\Theta = \{\Theta_{enc}, \Theta_{dec}\}$ can be pre-trained with the following joint objective:

$$\mathcal{J} = \underbrace{\log P(x|\hat{x}; \Theta_{enc}, \Phi_{head})}_{\text{MLM}} + \underbrace{\log P(y|\hat{y}; \Theta_{dec})}_{\text{CLM}} + \underbrace{\log P(x|\hat{x}; \Theta) + \log P(y|\hat{y}; \Theta)}_{\text{DAE}} \quad (1)$$

where Φ_{head} is a linear layer sharing the same parameter with the embedding layer following Devlin et al. (2019), and $\hat{x}$ and $\hat{y}$ are the noise version of x and y, respectively. In practical use, each above objective can be controlled by adding and removing its training terms, such as removing CLM and DAE for MLM pre-training. In the fine-tuning stage, we can add the translation objective for the golden training pairs and the BT synthetic data, respectively.

2.3.4 Evaluation. For the evaluation of EN-DE machine translation performance, we follow Vaswani et al. (2017) to evaluate the tokenized BLEU score for all models via *compound_split_bleu.perl*.[4] In the case of EN-RO machine translation, we adopt the tokenization and normalization procedures described in Sennrich, Haddow, and Birch (2016a) by using the Moses script.[5] For EN-TR, the BLEU score is measured by the detokenized case-sensitive SacreBLEU (Kudo and Richardson 2018).[6] For EN-FR, EN-RU, EN-AR, and EN-ZH, the tokenized BLEU score is computed using SacreBLEU.

3. Understanding Monolingual Exploitation

In this section, we first conduct probing tasks on understanding the effects of BT, MLM, CLM, and DAE on translation models, respectively. Then, in Section 3.4, we further compare these methods on translation tasks with various bitext sizes, including extremely low-resource settings and low-resource settings. In this article, we denote the joint objective of MLM, CLM, and DAE as MLM+CLM+DAE.

3.1 Source Language Understanding

The probing tasks of SentEval are conducted on the translation model encoder for fine-grained analysis, in order to explore two questions: (1) How do existing methods influence the translation model encoder on linguistic properties? and (2) Are self-supervised objectives complementary to each other? For probing tasks, we initialize the

4 https://github.com/facebookresearch/fairseq/blob/main/scripts/compound_split_bleu.sh.
5 https://github.com/rsennrich/wmt16-script.
6 BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0.

Table 1
Classification accuracy on SentEval probing tasks of evaluating the linguistic properties. "+BT" means training a translation model with BT synthetic parallel data. "MLM+CLM+DAE" means the joint objective of the three self-supervised objectives.

Method	EN-DE 100k			EN-DE 200k		
	Surf.	**Sync.**	**Semc.**	**Surf.**	**Sync.**	**Semc.**
base	54.93	49.71	68.47	47.62	51.00	69.39
+BT	55.99	47.28	60.08	32.37	45.35	66.61
MLM	68.56	56.74	68.90	57.45	54.46	69.29
CLM	60.79	52.46	69.73	52.62	53.25	70.55
DAE	61.86	49.08	70.10	64.43	50.29	70.92
MLM+CLM+DAE	**69.37**	**54.96**	**73.15**	**83.53**	**59.97**	**72.96**

(a) Average classification accuracy of EN-DE 100k and EN-DE 200k.

Method	**Surf.**		**Sync.**			**Semc.**				
	Seln	WC	TDep	ToCo	BShif	Tense	SubN	ObjN	SoMo	CoIn
base	46.21	63.64	33.22	62.78	53.14	82.67	77.14	76.87	50.03	57.10
+BT	67.37	44.61	30.46	61.37	50.01	68.74	68.50	63.28	49.87	50.01
MLM	70.65	66.47	37.67	70.80	61.75	83.36	77.41	76.50	50.42	56.83
CLM	70.05	51.53	37.03	67.10	53.26	82.81	78.58	77.85	50.01	59.38
DAE	58.40	65.32	33.20	61.17	52.87	86.26	78.76	79.52	50.53	57.10
MLM+CLM+DAE	72.07	66.67	38.49	68.02	58.36	86.98	84.05	83.21	53.57	57.94

(b) Detailed classification accuracy of EN-DE 100k.

Method	**Surf.**		**Sync.**			**Semc.**				
	Seln	WC	TDep	ToCo	BShif	Tense	SubN	ObjN	SoMo	CoIn
base	29.04	66.19	32.09	64.86	56.04	84.09	78.62	77.61	51.04	55.59
+BT	17.32	47.42	24.66	56.06	55.34	80.72	75.32	73.33	49.47	54.19
MLM	65.65	49.25	34.7	66.98	61.71	85.06	76.96	77.34	49.97	57.13
CLM	45.47	59.76	34.58	69.26	55.92	84.86	80.23	79.23	50.31	58.10
DAE	70.14	58.72	35.26	62.84	52.77	86.21	79.13	81.15	50.97	57.15
MLM+CLM+DAE	87.72	79.34	42.18	75.70	62.02	88.09	82.74	82.38	51.97	59.64

(c)Detailed classification accuracy of EN-DE 200k.

probing model with a translation model encoder and an additional classification head. Then, we further fine-tune the classification head on 10 classification probing tasks of SentEval by freezing the classification head, respectively. The hyperparameters of these tasks are the same as the configuration of Conneau and Kiela (2018). The models are trained until early stopping based on the validation loss, with a patience of 10. The 10 probing tasks are divided into three categories (Conneau and Kiela 2018): (1) surface tasks (Surf.) evaluate surface properties in the sentence embedding; (2) syntactic tasks (Sync.) are designed to evaluate the capabilities of the encoder on capturing the syntactic information; (3) semantic tasks (Semc.) assess the ability of the encoder to understand a sentence in the semantic level. The detailed information is listed in Appendix A. The average results for these three categories are reported in Table 1a.

Observation 1. The results of our experiments demonstrate that the self-supervised objectives respectively improve source language understanding and are complementary to each other. The

joint objective of MLM+CLM+DAE further yields performance gains. The results show that MLM achieves the highest accuracy improvements in the surface and syntactic properties, with increases of up to 3.73 and 7.03 in EN-DE 100k, and up to 9.83 and 3.46 in EN-DE 200k, respectively, compared with the base model. These indicate that the MLM objective is beneficial for capturing the surface and syntactic information. Among the three objectives, DAE demonstrates the best performance in the semantic property, with an accuracy of 70.10, which benefits from the use of both source-side and target-side monolingual data. CLM shows improvement in these probing tasks despite only being trained on target-side monolingual data and initializing the shared embedding layer. It is known that the MLM pre-trained model learns to extract the surface, syntactic, and semantic information via predicting the masking words (Jawahar, Sagot, and Seddah 2019), and our experiment finds that its translation model demonstrates superiority in surface and syntactic tasks compared with those of CLM and DAE. For the semantic tasks, the translation model of DAE outperforms both those of MLM and CLM, while CLM achieves second place. As a reason, DAE trains an encoder-to-decoder model to generate a sentence, which may require the encoder to generate deeper hidden features. The encoder of CLM is randomly initialized and trained on translation tasks, which directly captures deeper linguistic properties, such as semantic information (Hao et al. 2019; Xu et al. 2019). By simply combining them, the joint objective of MLM+CLM+DAE informs significant improvements in surface, syntactic, and semantic properties for the translation model, with gains of up to 14.44, 5.25, and 4.68 in EN-DE 100k, and up to 35.91, 8.97, and 3.57 in EN-DE 200k, compared with the base model, respectively. Overall, the results of our experiments suggest that these self-supervised objectives are complementary to each other in terms of improving source language understanding.

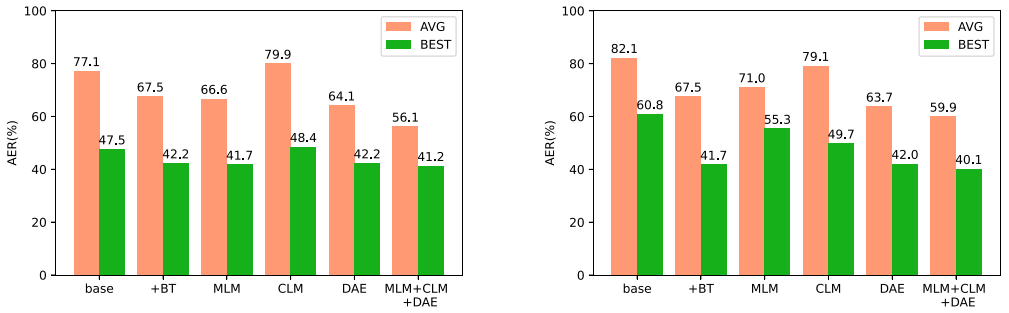
Observation 2. The experiment results indicate that training a translation model on BT synthetic data negatively impacts the syntactic and semantic understanding of the encoder. Although there is a slight improvement in the surface property for EN-DE 100k, the results of training a translation model with both golden bitexts and BT synthetic data suggest a detrimental effect on the syntactic and semantic understanding of the model encoder for both EN-DE 100k and EN-DE 200k. Specifically, the average accuracy of Sync. degrades to 47.28, while that of Semc. degrades to 60.08 for EN-DE 100k, and those degrade to 45.35 and 66.61 for EN-DE 200k, respectively. These highlight the significant impact of the noise present in BT data on the encoder's ability to understand source sentences. Previous research has demonstrated that the BT synthetic data is of noise and could potentially interfere with the encoder's encoding process, especially for the translation task with a small dataset (Edunov et al. 2018).

In general, using monolingual data through self-supervised objectives can improve a model's ability to capture the linguistic properties of source sentences, while training with BT data informs some negative impacts. Nevertheless, both pre-training and back-translation play crucial roles in enhancing the quality of translations for low-resource tasks as demonstrated in previous studies (Edunov et al. 2018; Liu et al. 2020, 2021). For translation tasks, BT synthesizes pseudo bitexts and has been shown to alleviate the data scarcity problem (Sennrich, Haddow, and Birch 2016b; Edunov et al. 2018). Therefore, although BT introduces some noisy data, it may improve other properties of a translation model. To further investigate the difference between pre-training and back-translation, we conduct experiments on two other key factors of translation models, which are bilingual word alignment and translation fluency, in the following.

3.2 Bilingual Word Alignment

Bilingual word alignment, a task of extracting word-level alignments from two bilingual parallel sentences, is closely related to machine translation, as demonstrated by Ganchev, Graça, and Taskar (2008). Recent studies (Garg et al. 2019; Chen et al. 2020) have utilized transformer translation models to extract alignments, indicating the positive correlation between the two tasks. In order to understand the contribution of the three pre-training methods (MLM, CLM, and DAE) and BT to the translation model, we use bilingual word alignment as a probing task. To present a concrete comparison, we present both the average (**AVG**) and best (**BEST**) alignment performance across layers, as the alignments may differ greatly among layers (Garg et al. 2019). The results are shown in Figure 1.

Observation 3. Pre-training a model with MLM+CLM+DAE achieves both the lower AVG and BEST AER scores for the bilingual word alignment task. The simultaneous pre-training of the model with the three self-supervised objectives results in an AVG AER score of 56.1 and a BEST AER score of 41.2 for EN-DE 100k, and an AVG AER score of 59.9 and a BEST AER score of 40.1 for EN-DE 200k, outperforming the base, BT, and three other individual objectives. DAE shows a better AVG AER score compared with MLM and CLM, indicating better overall alignment accuracy across different layers. DAE is a seq-to-seq self-supervised objective that trains both the encoder and decoder, including cross-attention modules. The cross-attention module queries the encoder output features and generates new features for the next layer (Vaswani et al. 2017). In alignment tasks, we extract the alignment hypothesis from the cross-attention matrix, emphasizing the importance of accurate attention in the cross-attention layer for the interaction between source and target features (Gheini, Ren, and May 2021). Additionally, MLM outperforms the base model on both AVG and BEST metrics for EN-DE 100k and EN-DE 200k, respectively, demonstrating that MLM has effectively learned source monolingual knowledge. CLM presents some side effects in the translation model trained on EN-DE 100k, which tend to disappear as the dataset size increases to EN-DE 200k. As a result, the joint objective of MLM+CLM+DAE in the pre-trained model results in improved performance for three properties compared to other methods, suggesting the complementarity of these self-supervised objectives.



(a) Results of models trained on EN-DE 100k (b) Results of models trained on EN-DE 200k

Figure 1
Alignment Error Rate (AER) on the English-German test (Vilar, Popović, and Ney 2006). "**AVG**" and "**BEST**" are average AER and the best AER across all layers. Detailed information and results are presented in Table B.1.

Table 2
Back-translation synthesis parallel data examples. "**S**" and "**T**" stand for pseudo source sentences and golden target sentences, respectively. We manually recover subwords and mark the alignments in bold and italics.

S	It is pop to ***America*** before the Constitution wildfires.
T	Diese Bamberger Besonderheit wird bis nach ***Amerika*** exportiert.
S	In addition, ***but*** then.
T	Daneben, ***aber*** immerhin.
S	Harvard professor ***Martin*** SVioling suggests posts in a ***good*** slogan.
T	CEO ***Martin*** Senn spricht von einem ***guten*** Ergebnis

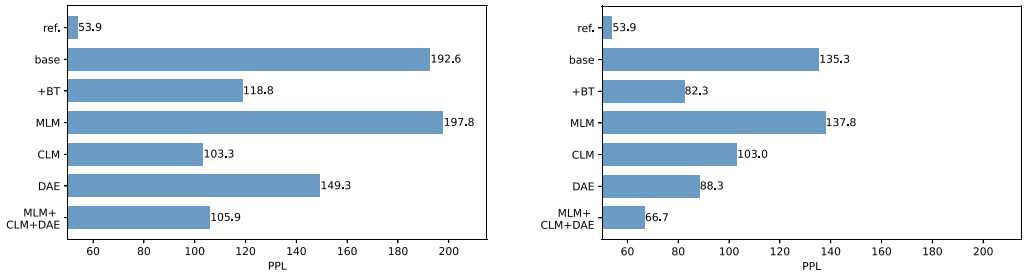
Observation 4. The utilization of BT enhances the performance of the bilingual word alignment task to a certain degree. This is achieved through enriching the parallel corpus with the synthetic data generated by a reversed translation model. To assess the quality of the generated pseudo-data, we conduct a case study, the results of which are presented in Table 2. The synthetic instances contain aligned words, which may enrich parallel corpus and improve word alignment information as well. Therefore, although the generated data is noisy and harmful for SentEval probing tasks, it probably enhances word alignment information, which positively affects the alignment performance, as evidenced by the BEST AER score of 42.2 for EN-DE 100k and the 41.7 for EN-DE 200k. These results align with the findings discussed in Section 3.1, where it is concluded that although BT introduces noise to source language understanding, it enhances the bilingual word alignment ability of the translation model.

The results presented in Figure 1 inform that pre-training with the CLM objective has some side effects for the bilingual word alignment tasks, while both MLM and DAE pre-trained models result in improvement. Further comparison is made by evaluating the joint objective of MLM+DAE as depicted in Table B.1. This combination is comparable with the joint objective of MLM+CLM+DAE, which achieves a lower BEST AER score but a higher AVG AER score.

3.3 Translation Fluency

Early work has explored the importance of generation fluency for translation tasks (Snover et al. 2009), indicating that translation fluency is a crucial factor for improving translation quality, especially for low-resource settings (Xia et al. 2019; Ranathunga et al. 2021). In this section, we evaluate the translation fluency by computing the sentence perplexity (PPL) using the German-GPT2 (Schweter 2020) and reporting the average PPL in Figure 2. Accordingly, we have:

Observation 5. Learning generating target-side monolingual sentences apparently improve translation fluency. In our experiments, the sentence PPL of the golden reference is 53.9. The BT synthesis parallel data consists of pseudo-source and golden target sentences, which are trained during the fine-tuning stage. The CLM objective is trained to causally generate target-side sentences in the pre-training stage. Both approaches achieve a significant improvement in translation fluency, with CLM achieving the lowest PPL score of 103.3 for the EN-DE 100k setting and the BT method achieving the lowest PPL score



(a) Results of models trained on 100k (b) Results of models trained on 200k

Figure 2
Perplexity (PPL) on translation sentences. For both reference and hypothesis, we use German-GPT2 (Schweter 2020) to compute the average PPL.

of 82.3 for the EN-DE 200k. By incorporating these self-supervised objectives, the joint objective of MLM+CLM+DAE achieves 105.9 and 66.7 PPL scores for both translation tasks, respectively, representing an 86.7 and a 68.6 score decrease compared with the base model.

3.4 Bitext Scale Effect

The former observations indicate that the pre-trained model with the joint observation of MLM+CLM+DAE obtains competitive performance on three probing tasks, compared to individually pre-trained methods. In this section, we investigate the impact of bitext scales on the performance of the self-supervised objectives and BT in both the extremely low-resource and low-resource translation task settings. The joint objective of MLM+CLM+DAE is applied to translation tasks, both in the multi-task learning and pre-training scenario, with various bitext scale settings. Meanwhile, we investigate the impact of multi-task fine-tuning methods using target-side monolingual data, such as CLM and target-side DAE (tDAE), as well as the efficiency of BT on various bitext scales. The findings are presented in Figure 3 and Table 3.

Observation 6. The pre-trained model with the joint objective of MLM+CLM+DAE consistently improves the translation quality both for extremely low-resource tasks and low-resource tasks. In Figure 3, the PT method presents a dominant performance in extremely low-resource tasks and achieves competitive results on low-resource tasks compared to other methods such as base, BT, and MTL. For example, in Table 3, the PT method obtains 11.35, 13.39, and 14.33 score gains for 5k, 10k, and 15k settings, respectively, which is far better performance than the other three methods. This indicates that the PT method adequately trains the model on monolingual data and provides a better model initialization, which especially benefits extremely low-resource translation tasks. In contrast, results show that the BT method does not effectively improve the translation quality in extremely low-resource scenarios, such as datasets with 5k, 10k, and 15k parallel sentences. Furthermore, the MTL method, which trains both the parallel corpus with the translation objective and monolingual data with self-supervised objectives simultaneously, may introduce a detrimental effect for the extremely low-resource scenario because the parallel corpus is much smaller than the size of monolingual data.

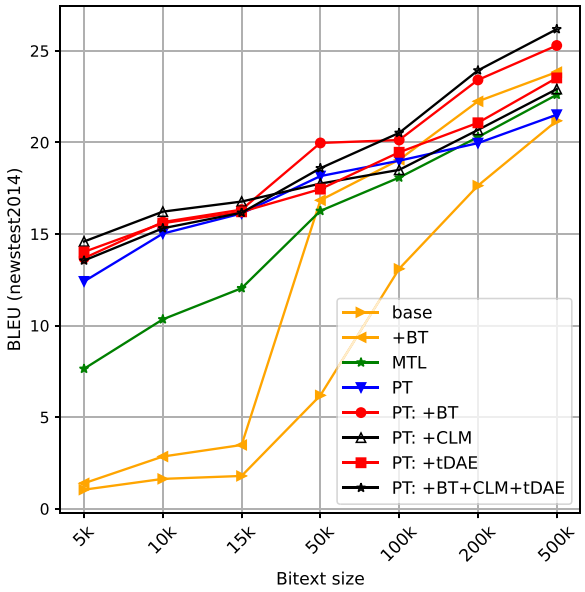


Figure 3
BLEU score with various bitext sizes. "tDAE" denotes DAE on target-side monolingual data, "PT" means the model is first pre-trained with the joint objective of MLM+CLM+DAE. Detailed information is presented in Table 3.

Table 3
BLEU score on *Newstest2014 English-to-Germany*. "MTL" denotes multi-task learning with MLM, CLM, and DAE objectives. "PT" means pre-training a model with MLM, CLM, and DAE objectives.

Method	Extremely Low-Resource			Low-Resource			
	5k	**10k**	**15k**	**50k**	**100k**	**200k**	**500k**
base	1.04	1.63	1.79	6.19	13.10	17.64	21.18
+BT	1.39	2.85	3.48	16.84	19.05	22.24	23.84
MTL	7.64	10.34	12.04	16.26	18.08	20.29	22.60
PT	12.41	15.02	16.12	18.16	19.00	19.96	21.51
+BT	13.68	15.65	16.33	19.97	20.13	23.41	25.29
+CLM	14.59	16.22	16.77	17.75	18.50	20.68	22.91
+tDAE	14.02	15.59	16.21	17.45	19.46	21.07	23.52
+BT+CLM+tDAE	13.55	15.31	16.16	18.59	20.52	23.93	26.18

Observation 7. BT yields limited improvement in terms of translation quality for extremely low-resource tasks, but gradually demonstrates its superiority and outperforms other methods as bitext size increases. In comparison to the base model, training with BT synthetic data only leads to marginal improvement in BLEU scores for the extremely low-resource translation tasks, which are 1.39, 2.85, and 3.48 for tasks of 5k, 10k, and 15k bitext size, respectively. This is due to the fact that the reversed translation models in these cases tend to produce unreliable source sentences, whose pseudo bitexts are too detrimental to the translation models (Gu et al. 2018a; Edunov et al. 2018). Furthermore, for these

cases, training with BT synthetic data following pre-training still yields a lower performance compared to training using the target-side generation knowledge via multi-task fine-tuning with the CLM objective (see the curve of "PT: +CLM" in Figure 3). However, as the bitext size increases to 50k, 100k, 200k, and 500k in our experiments, the BT method gradually demonstrates improved performance and outperforms some other methods, indicating that high-quality synthetic data plays a crucial role in boosting the performance of a translation model in the general low-resource setting.

Observation 8. The results suggest that the MTL method has limited improvement on the translation model compared with other methods in most cases, but begins to exhibit superiority in larger bitext scenarios such as 200k and 500k. In most cases ($<$ 200k), the performance gains obtained from the MTL method are inferior to those of the PT methods, which adopt the same joint objective of MLM+CLM+DAE in the pre-training stage. However, as the bitext size increases, the performance gap between the MTL and PT methods gradually decreases, and the MTL method even outperforms the PT method in the 200k and 500k scenarios. These results hint that the MTL method has the potential to further enhance translation quality in the fine-tuning stage.

Observation 9. Multi-task fine-tuning with self-supervised objectives further improves translation qualities for both the extremely low-resource task and the low-resource tasks. In Table 3, the results reveal that fine-tuning with the CLM objective successfully improves performance in extremely low-resource settings. In cases of low-resource tasks with more than 50k bitexts, a more pronounced improvement in translation quality is observed when fine-tuning with both the CLM and tDAE objectives, as well as training additional BT synthetic data. This indicates that multi-task fine-tuning is of great potential to mitigate the issue of catastrophic forgetting.

Overall, to build a translation model with limited parallel corpus, we can first pre-train the encoder-to-decoder model with the joint objective of MLM+CLM+DAE for obtaining linguistic properties. Then, in the fine-tuning stage, multi-task fine-tuning further makes use of the monolingual data to improve the translation quality.

4. Pipelines on the Exploitation of Monolingual Data

In this section, we design two pipelines for both the extremely low-resource and low-resource translation tasks, respectively. Then we conduct evaluation experiments on public-available datasets, including similar and distant translation pairs, to validate our claims and findings.

4.1 Pipelines

4.1.1 Pipeline on Extremely Low-Resource Setting. According to observations 1, 3, and 6, we find that pre-training the encoder-to-decoder model with a joint objective of MLM+CLM+DAE can lead to the following outcomes: (1) improved linguistic properties of the source sentences, (2) enhanced bilingual word alignment performance, and (3) competitive translation fluency. This provides a decent model initialization for extremely low-resource translation tasks, where the PT approach shows significant improvements in translation quality as shown in Table 3. Thus, the initial step in our pipeline involves pre-training the model with the joint objective of MLM+CLM+DAE. In the fine-tuning stage, the translation model is further improved by multi-task learning with the CLM objective, as informed in Observations 5 and 9.

4.1.2 Pipeline on Low-Resource Setting. For low-resource translation tasks, we adopt a pre-training strategy that leverages the joint objective of MLM+CLM+DAE. In the fine-tuning stage, we utilize multi-task fine-tuning to further leverage target-side monolingual data, incorporating BT, CLM, and tDAE objectives, as informed in Observations 4, 5, and 6. The observations in Section 3 indicate that BT can enhance the word alignment performance and generation fluency of a translation model, while CLM and tDAE have a positive impact as well. In contrast to the extreme low-resource setting, the BT synthetic data is less noisy and substantially improves the translation model of the low-resource tasks, as observed in Observation 7.

To demonstrate our findings, we include two popular and effective existing methods as our baselines, including fine-tuning on MLM pre-trained and DAE pre-trained models (Devlin et al. 2019; Liu et al. 2020). For a sound comparison, on the one hand, we combine all the techniques, including MLM, CLM, and DAE, in both the pre-training and the fine-tuning stages, which is denoted as *All PT-and-FT*. On the other hand, we include the small transformer architecture, which is a base architecture with three encoder layers and one decoder layer, named *base-Enc3Dec1*. During the training process, we set the max-token of a batch to 2,048 and 8,192 for extremely low-resource and low-resource translation tasks, respectively, and stop training when the model no longer reduces the validation loss within 10 times.

4.2 Results

The experiment results are presented in Tables 4a and 4b. Our pipelines demonstrate consistent improvement over the base model and existing methods in the case of extremely low-resource translation tasks. For low-resource tasks, the pipeline achieves competitive results on EN-DE 100k and EN-TR, and comparable performance on EN-RO. These findings align with our observations and demonstrate the effectiveness of our pipelines on various bitexts and translation pairs, including distant language pairs such as EN-AR and EN-ZH. On the one hand, the small transformer model of base-Enc3Dec1 obtains better performance in EN-RU, EN-AR, and EN-ZH translation tasks compared with the base model, but they fail to leverage BT synthetic data since the model capacity is too limited to learn these abundant noisy sentence pairs. On the other hand, the small translation model also shows poor performance in extremely low-resource translation tasks as the base model, which highlights the challenge of data scarcity.

In general, for extremely low-resource translation tasks, the base model achieves a BLEU score of less than 5, except for the EN-FR language pair, as well as the small translation models. This highlights the challenge of training a translation model with only a small number of training pairs. To address this challenge, we first use the joint objective of MLM+CLM+DAE to train the model on monolingual data for better model initialization, thereby improving its abilities in source language understanding, bilingual word alignment, and generating fluent sentences. Another intuitive method is to enrich the data scale for extremely low-resource tasks. However, the use of BT synthetic data in these settings has a limited impact on translation quality, as demonstrated in Observation 7. As an alternative, incorporating additional target-side sentences through multi-task learning with a CLM objective during the fine-tuning stage has been shown to lead to improved performance in extreme settings, as discussed in Observations 5 and 9. Therefore, in Table 4b, our pipelines obtain impressed translation quality for both close and distant translation pairs, except that the method of DAE with BT data obtains a slight improvement of 0.1 for the close translation task, EN-FR. Additionally, the models

Table 4
Results of our pipelines on multiple datasets. The bitext size of extremely low-resource is 10k and that of low-resource is 100k if not specified, respectively.

Method	Extremely Low-Resource				Low-Resource		
	EN-DE 5k	EN-RO 5k	EN-DE 10k	EN-RO 10k	EN-DE 100k	EN-TR	EN-RO
base	1.04	3.50	1.63	4.80	13.10	10.00	33.90
+BT	1.39	7.10	2.85	8.20	19.05	12.60	35.00
Small Transformer Model							
base-Enc3Dec1	0.70	4.50	1.74	4.90	10.78	5.40	27.70
+BT	0.00	2.90	0.00	1.80	0.00	3.40	2.90
Existing Pre-Training Method							
MLM	1.95	6.90	3.59	13.30	15.94	11.60	34.90
+BT	2.89	9.10	5.30	13.10	18.80	13.10	38.40
DAE	12.84	25.20	14.55	27.60	17.46	11.10	35.60
+BT	13.74	26.20	15.71	29.10	20.50	13.30	38.90
Combine All Techniques							
All PT-and-FT	11.85	18.55	14.90	19.95	14.93	12.21	29.10
+BT	12.56	18.08	13.48	19.82	17.97	12.56	31.00
Our Pipelines	**14.59**	**27.50**	**16.22**	**30.40**	**20.52**	**13.90**	**38.90**

(a) The BLEU score on WMT datasets.

Method	Extremely Low-Resource				Low-Resource			
	EN-FR	EN-RU	EN-AR	EN-ZH	EN-FR	EN-RU	EN-AR	EN-ZH
base	6.30	2.90	1.20	2.80	26.70	12.20	5.50	13.40
+BT	11.20	5.00	2.80	6.60	31.60	18.50	11.70	19.90
Small Transformer Model								
base-Enc3Dec1	9.50	2.90	0.70	1.20	26.00	16.50	9.90	15.40
+BT	8.80	2.50	0.60	1.40	25.30	11.10	8.50	13.10
Existing Pre-Training Method								
MLM	13.40	5.00	2.20	5.90	31.50	20.00	11.50	18.90
+BT	16.60	8.20	5.60	10.40	32.60	22.00	13.80	22.00
DAE	26.50	13.70	8.50	10.20	33.00	21.10	12.60	19.90
+BT	**26.90**	14.20	8.70	12.50	33.10	23.50	16.40	22.40
Combine All Techniques								
All PT-and-FT	24.10	12.80	7.20	10.70	34.00	23.30	17.50	24.80
+BT	24.00	12.90	6.80	10.50	34.60	23.50	13.60	21.60
Our Pipelines	26.80	**16.50**	**9.40**	**16.70**	**34.90**	**23.70**	**18.10**	**23.80**

(b) The BLEU score on OPUS-100 sampled datasets.

of *Combine All Techniques* present moderate results compared with other methods, which demonstrates the reasonableness of our designed pipelines. Additionally, MLM does not significantly improve the performance of extremely low-resource tasks, with a BLEU score of 3.59 in En-De 10k and 13.30 in En-Ro 10k, while DAE achieves a score of 14.55 in En-De 10k and 27.60 in En-Ro 10k. The performance gap between MLM and DAE indicates that relying solely on source monolingual data to improve translation quality in extreme settings is insufficient.

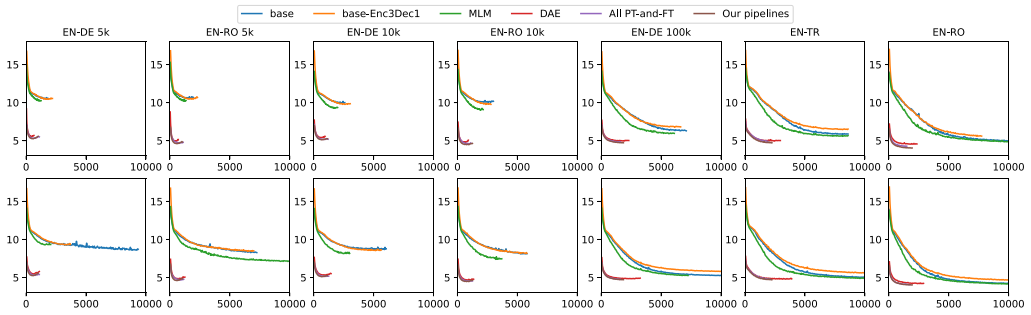
For low-resource machine translation tasks, incorporating both BT synthetic data and golden parallel data into the training process generally results in improved translation quality. This supports the finding of Observation 7 that the quality of BT data tends

to be improved with an increase in bitext size. Therefore, except for the extremely low-resource settings, BT is an effective method for improving translation quality generally, although it requires an additional reverse translation model for data synthesis. Experimental results demonstrate that as the bitext size increases, the performance gaps between our proposed pipelines and existing methods tend to diminish, in agreement with the prior research which suggests that pre-training does not significantly contribute to translation performance in high-resource settings (Liu et al. 2020). It is worth noting that the model of ALL PT-and-FT obtains an impressive result of 24.80 BLEU score in EN-ZH low-resource translation task without additional training with BT synthetic data, which reveals the potential of the combination of MLM, CLM, and DAE techniques.

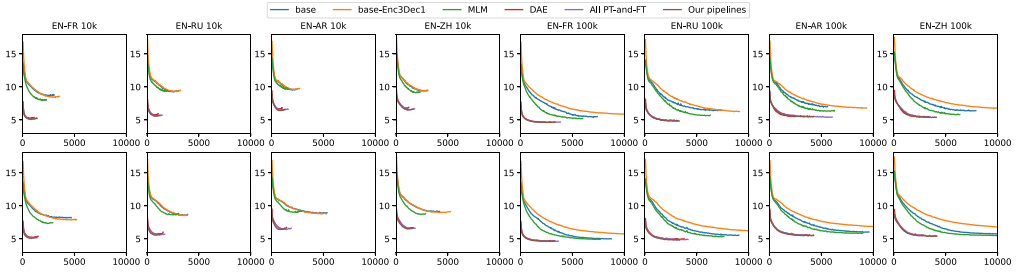
Table 4b presents the results of evaluating both similar and distant translation pairs. The results demonstrate the efficacy of our pipelines on distant translation pairs, particularly in extreme settings. Specifically, our pipelines improve by 2.3, 0.7, and 4.2 BLEU score for the EN-RU, EN-AR, and EN-ZH translation tasks of 10k bitext size, respectively. Despite the marginal differences in alphabets between English (EN), Russian (RU), Arabic (AR), and Chinese (ZH), our pipelines still achieve remarkable translation performance, demonstrating the generalization capability of our pipelines.

4.3 Analysis

During the training progress, the models are validated every 50 updates, and Figure 4 shows the validation loss with the update steps for each model in Table 4. We find



(a) Validation loss on WMT datasets of Table 4a.



(b) Validation loss on OPUS-100 sampled datasets of Table 4b.

Figure 4
Training curves with update steps. In each sub-figure, the first and the second rows show the models training without and with BT data, respectively. Each column corresponds to each translation task in Table 4 in order, respectively.

that (1) the models only trained with extremely low-resource datasets converge quickly and easily suffer from the overfitting problem; (2) the models pre-trained with the MLM+CLM+DAE joint objective generally converge to a stage with a lower validation loss and the DAE pre-trained models take the second place. These phenomena highlight that the model initialization is crucial for improving the extremely low-resource and low-resource translation tasks. Also, as the dataset size increases, the models converge to a certain range. For example, the models of EN-RO generally reach around 5 of validation loss in Figure 4a. This echoes the finding that the influence of the pre-training methods is on the ebb as the dataset size increases.

5. Conclusion

In this article, we undertake a systematic investigation of the effects of existing methods such as Back-Translation (BT), Masked Language Modeling (MLM), Causal Language Modeling (CLM), and Denoise Autoencoding (DAE) through three fine-grained probing tasks that assess source language understanding, bilingual word alignment, and translation fluency. The outcomes of the probing tasks inform that MLM, CLM, and DAE are complementary to each other in improving the linguistic properties of a translation model. Then, we analyze the impact of Pre-Training, Back-Translation, and Multi-Task Learning on translation tasks of various sizes. Results show that multi-task fine-tuning further improves both the extremely low-resource task and low-resource tasks. With the same model capacity and data scale, we summarize nine observations about how these existing methods affect the translation model and provide fine-grained analysis. Our pipelines show an effective improvement for resource-poor translation tasks, offering an alternative to relying solely on costly parallel data. Inspired by recent research (Xue et al. 2021; Muennighoff et al. 2022; Touvron et al. 2023), we would like to extend our work to the scenario with large-scale model capacity and data for improving the quality of the translation model for both low-resource and high-resource translation tasks.

Appendix A: Detailed Information of SentEval

The SentEval probing tasks include 10 classification tasks for English sentences. All ten datasets contain 100k training instances, 10k validation instances, and 10k test instances. The details could be found on the official Web site;[7] we simply list them as follows:

1. Sentence Length (Seln): All sentences have been binned into 6 possible categories with lengths ranging in the following intervals: −0: (5–8), 1: (9–12), 2: (13–16), 3: (17–20), 4: (21–25), 5: (26–28). The classification model is trained to predict the length of a given sentence.
2. Word Content (WC):
   The targets are 1,000 lower-cased words. Then the classification model is trained to predict which word the given sentence contains among these 1,000 words.

7 https://github.com/facebookresearch/SentEval/tree/main/data/probing.

3. Tree Depth (TDep): The classification model is asked to predict the maximum depth of the sentence's syntactic tree. The tree depth range is from 5 to 12, 7 classes in total.
4. Top Constituents (ToCo): This task requires the classifier to predict the sentence structures, including 19 common structures and one OTHER class.
5. Bigram Shift (BShif): This is a binary classification task, which requires predicting whether two consecutive words within the sentence have been inverted.
6. Pas Present (Tense): This requires the model to predict whether the main verb is present or past tense.
7. Subj Number (SubN): A binary classification task to predict the number of the subject of the main clause. The first class means singular, while the second class means plural or mass.
8. Obj Number (ObjN): A binary classification task to predict the number of the object of the main clause. Class labels are the same as the task of SubN.
9. Odd Man Out (SoMo): This is a binary classification task to predict whether a noun or a verb of a sentence was replaced with another form.
10. Coordination Inversion (CoIn): This binary task asks whether the order of two coordinated clausal conjoint has been inverted.

Appendix B: Detailed Results of Bilingual Word Alignment

The detailed results are listed in Table B.1. Accordingly, MLM and DAE consistently improve the alignment performance for both data settings, while CLM shows a slight side effect for EN-DE 100k. As a result, the method of MLM+DAE slightly improves the BEST AER score by removing the CLM objective, but the joint objective of MLM+CLM+DAE still demonstrates the lowest AVG AER.

Table B.1
Alignment Error Rate (AER) on English-to-German (Vilar, Popović, and Ney 2006). Numbers 1 to 6 indicate the decoder layer and so as others. "AVG" and "BEST" are average AER and best AER across all layers.

Method	AER(%)							
	1	2	3	4	5	6	AVG.	BEST
base	93.5	94.6	94.2	82.6	50.4	47.5	77.1	47.5
+BT	89.0	87.8	84.7	55.3	42.2	45.7	67.6	42.2
MLM	92.3	86.2	84.2	48.4	41.7	46.6	66.6	41.7
CLM	94.3	94.6	93.0	84.6	64.3	48.4	79.9	48.4
DAE	86.6	88.9	74.6	48.4	42.2	43.9	64.1	42.2
MLM+DAE	85.4	87.7	58.6	49.2	41.1	45.4	61.2	41.1
MLM+CLM+DAE	88.4	86.0	49.5	47.5	41.2	44.1	56.1	41.2

(a) Results of models trained on EN-DE 100k

Method	AER(%)							
	1	2	3	4	5	6	AVG.	BEST
base	93.8	94.5	93.7	84	65.7	60.8	82.1	60.8
+BT	88.5	87	84.3	55.7	41.7	47.3	67.4	41.7
MLM	90.4	86.4	81.6	56.9	55.3	55.5	71.0	55.3
CLM	93.2	94.4	92.0	84.2	61.2	49.7	79.1	49.7
DAE	84.2	87.8	74.7	47.4	42.0	46.2	63.7	42.0
MLM+DAE	85.1	86.0	61.1	47.1	39.8	46.8	61.0	39.8
MLM+CLM+DAE	84.5	84.7	59.5	47.5	40.1	42.9	59.9	40.1

(b) Results of models trained on EN-DE 200k

Acknowledgments

This work was supported in part by the Science and Technology Development Fund, Macau SAR (grant nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ) and the Multi-year Research Grant from the University of Macau (grant no. MYRG2020-00054-FST), and Alibaba Group through Alibaba Research Intern Program.

The authors would like to thank all the anonymous reviewers and the chief editor for their insightful comments.

References

Baziotis, Christos, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*. https://doi.org/10.18653/v1/2020.emnlp-main.615

Bertoldi, Nicola and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189. https://doi.org/10.3115/1626431.1626468

Bojar, Ondřej and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336.

Caswell, Isaac, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63. https://doi.org/10.18653/v1/W19-5206

Chen, Yun, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576. https://doi.org/10.18653/v1/2020.emnlp-main.42

Conneau, Alexis and Douwe Kiela. 2018. SentEval: An evaluation toolkit for

universal sentence representations. *CoRR*, abs/1803.05449.

Devlin, Jacob, 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, Volume 1 (Long and Short Papers)*, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Edunov, Sergey, 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*, pages 489–500. https://doi.org/10.18653/v1/D18-1045

Ganchev, Kuzman, João V. Graça, and Ben Taskar. 2008. Better alignments = better translations? In *Proceedings of ACL-08: HLT*, pages 986–993.

Garg, Sarthak, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer 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)*, pages 4453–4462. https://doi.org/10.18653/v1/D19-1453

Gheini, Mozhdeh, Xiang Ren, and Jonathan May. 2021. Cross-attention is all you need: Adapting pretrained transformers for machine translation. *arXiv preprint arXiv:2104.08771*. https://doi.org/10.18653/v1/2021.emnlp-main.132

Gu, Jiatao, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018a. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354. https://doi.org/10.18653/v1/N18-1032

Gu, Jiatao, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018b. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631. https://doi.org/10.18653/v1/D18-1398

Gulcehre, Caglar, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Hao, Jie, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. 2019. Modeling recurrence for transformer. 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)*, pages 1198–1207. https://doi.org/10.18653/v1/N19-1122

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657. https://doi.org/10.18653/v1/P19-1356

Kim, Yunsu, Jiahui Geng, and Hermann Ney. 2019. Improving unsupervised word-by-word translation with language model and denoising autoencoder. *CoRR*, abs/1901.01590. https://doi.org/10.18653/v1/D18-1101

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kong, Xiang, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. 2021. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624. https://doi.org/10.18653/v1/2021.eacl-main.138

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*, pages 66–71. https://doi.org/10.18653/v1/D18-2012

Lample, Guillaume, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lewis, Mike, 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*, pages 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Liu, Xuebo, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the complementarity between pre-training and back-translation for neural machine translation. *arXiv:2110.01811*. https://doi.org/10.18653/v1/2021.findings-emnlp.247

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. https://doi.org/10.1162/tacl_a_00343

Luong, Minh Thang, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Muennighoff, Niklas, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111. https://doi.org/10.18653/v1/2023.acl-long.891

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Ranathunga, Surangika, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *CoRR*, abs/1907.12461.

Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *arXiv preprint arXiv:1907.12461*. https://doi.org/10.1162/tacl_a_00313

Schweter, Stefan. 2020. German GPT-2 model. https://doi.org/10.5281/zenodo.4275046

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376. https://doi.org/10.18653/v1/W16-2323

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 86–96. https://doi.org/10.18653/v1/P16-1009

Snover, Matthew, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. https://doi.org/10.3115/1626431.1626480

Tars, Maali, Andre Tättar, and Mark Fišel. 2021. Extremely low-resource machine translation for closely related languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 41–52.

Thompson, Brian, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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)*, pages 2062–2068. https://doi.org/10.18653/v1/N19-1209

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMa: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Processing Systems, volume 30, Curran Associates, Inc.

Vilar, David, Maja Popović, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments? In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*.

Wang, Yiren, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034. https://doi.org/10.18653/v1/2020.emnlp-main.75

Xia, Mengzhou, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796. https://doi.org/10.18653/v1/P19-1579

Xu, Mingzhou, Derek F. Wong, Baosong Yang, Yue Zhang, and Lidia S. Chao. 2019. Leveraging local and global patterns for self-attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3069–3075. https://doi.org/10.18653/v1/P19-1295

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148

Zhang, Jiajun 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*, pages 1535–1545. https://doi.org/10.18653/v1/D16-1160

Zhu, Jinhua, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wen gang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. *arXiv preprint arXiv:2002.06823*.