Does Masked Language Model Pre-training with Artificial Data Improve Low-resource Neural Machine Translation?

Hiroto Tamura Tosho Hirasawa Hwichan Kim Mamoru Komachi
Tokyo Metropolitan University
tamura-hiroto@ed.tmu.ac.jp, hirasawa-tosho@ed.tmu.ac.jp
kim-hwichan@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

Pre-training masked language models (MLMs) with *artificial data* has been proven beneficial for several natural language processing tasks such as natural language understanding and summarization; however, it has been less explored for neural machine translation (NMT). A previous study revealed the benefit of transfer learning for NMT in a limited setup, which differs from MLM. In this study, we prepared two kinds of artificial data and compared the translation performance of NMT when pre-trained with MLM. In addition to the random sequences, we created artificial data mimicking token frequency information from the real world. Our results showed that pre-training the models with artificial data by MLM improves translation performance in low-resource situations. Additionally, we found that pre-training on artificial data created considering token frequency information facilitates improved performance.

1 Introduction

Transfer learning is an effective method for improving the performance of various natural language processing tasks (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). This has been proven for neural machine translation (NMT) in low-resource situations (Zoph et al., 2016; Dabre et al., 2017; Qi et al., 2018). General explanations attribute the performance improvements in various downstream tasks to the transfer of linguistic traits (e.g., frequency, co-occurrence, and structure of words) in pre-training data (Lin et al., 2019; Tenney et al., 2019; Manning et al., 2020). Meanwhile, some studies have focused on identifying the specific traits in pre-training data that improve downstream task performance by employing artificial data for pre-training (Krishna et al., 2021; Chiang and Lee, 2022; Ri and Tsuruoka, 2022).

With regard to NMT, Aji et al. (2020) showed that pre-training a Transformer model on random sequences (see §2.2) brings better translation performance in low-resource situations. Their pre-training tasks included 1) autoencoding: translating one token into the same, and 2) substitution: translating one token into another; however, their solutions were uncommon for pre-training NMT models. Thus, the improvement of the translation performance when performing pre-training through the masked sequence-to-sequence model (Song et al., 2019; Lewis et al., 2020; Raffel et al., 2020) with artificial data was not addressed.

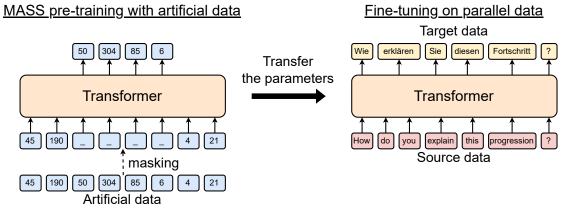


Figure 1: Experimental flow. We pre-train a Transformer model on the artificial dataset with the MASS objective, initialize the weights of the NMT model with the pre-trained one, and fine-tune it on parallel data.

In this work, we use masked language modeling for an encoder–decoder model called MAsked Sequence-to-Sequence pre-training (MASS; Song et al., 2019) as the pre-training task and investigate the translation performance of the NMT model pre-trained on artificial data in simulated (English→German) and genuine (English→Irish) low-resource situations (Figure 1). Additionally, other than random sequences, we create artificial data containing token frequency information from the real world and examine whether injecting this information into pre-training data affects translation performance. We compare the performance when pre-trained on each dataset with the MASS objective. Furthermore, we perform ablation studies to investigate how each part of the network affects the translation performance when pre-trained on artificial data by transferring or freezing some

Findings of the Association for Computational Linguistics: EACL 2023, pages 2216–2225
May 2-6, 2023 ©2023 Association for Computational Linguistics

parameters of the pre-trained model.

Our findings can be summarized as follows:

- Both in simulated and genuine low-resource situations, MASS pre-training with artificial data improves translation performance compared to the model without pre-training.
- Injecting token frequency information into artificial data further improves translation performance.
- Embeddings pre-trained on the artificial dataset mimicking token frequency information obtain useful representations for translation performance.

2 Pre-training Data

2.1 Real-world data

In this study, real-world data includes natural language data and natural language data undergoing some operations (e.g., token shuffling). The sentence examples for each pre-training dataset are in the Appendix (Table 3).

English We use the WMT News Crawl dataset of 2007, and its first 1M sentences for pre-training[1]. English is the source language in both low-resource situations.

English shuf We shuffle subwords from the "English" dataset throughout the corpus,[2] preserving sentence lengths. The generated sentences do not contain information about the structure or co-occurrence of tokens in a sentence. However, at the corpus level, the frequency information of the tokens is preserved.

German To examine the performance when pre-trained on the target language, we employ the German dataset, which is the target language in a simulated low-resource situation.[3] As with the "English" dataset, we use the WMT News Crawl dataset of 2007, and its first 1M sentences for pre-training.

2.2 Artificial data

All of the artificial data used in this study consists of integer tokens. No preprocesses, such as subword segmentation, are applied to artificial data. The vocabulary of each pre-training dataset contains only integers ranging from 0 to $(\mathit{vocabulary\ size\ for\ the\ downstream\ task} - 1)$. The number of sentences is 1M, and sentence lengths are the same as in the "English" dataset.

Random Integers are sampled independently from a uniform distribution to form sentences. This dataset contains no linguistic traits.[4]

Zipf Each integer is sampled independently from the Zipfian distribution[5]:

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)} \qquad (1)$$

where N is the vocabulary size, k is the frequency rank of the token, and s is the exponent value that characterizes the distribution. Here, we set s as 1.0, approximately consistent with the rank–frequency distribution in human-generated languages (Zipf, 1949). Unlike the "Random" dataset, this dataset contains token frequency information (linguistic trait), but no other traits.

3 Experimental Setup

Simulated low-resource situation We use 30k and 100k paired sentences randomly sampled from WMT14 English→German[1] (Europarl v7, Common Crawl, and News Commentary; approximately 4.5M sentences in total) to compare how pre-training with artificial data affects translation performance on different sizes. We use `newstest2013` of WMT as the validation set and `newstest2016` as the test set. We calculate case-sensitive BLEU using SacreBLEU[6,7] (Post, 2018) for evaluation.

Genuine low-resource situation We use English→Irish data in the COVID-19 domain from LoResMT21 (Ortega et al., 2021). The numbers of examples in the training/validation/test sets are 8,112/502/500, respectively. We report

[1] https://www.statmt.org/wmt14/translation-task.html

[2] Although we shuffled the tokens *after* subword segmentation in this work, we leave shuffling *before* subword segmentation, which preserves the order of subwords in the word unit, for future work.

[3] As pre-training on the target language was less effective in translation performance than on the source language in a simulated low-resource situation, we excluded the experiments when pre-trained on Irish, which is the target language in a genuine low-resource situation.

[4] Sentence length may be considered a linguistic trait; however, we discarded it in this work.

[5] A smaller integer is assigned a larger probability.

[6] https://github.com/mjpost/sacrebleu

[7] Signature: BLEU+case.mixed+lang.en-de+numrefs.1 +smooth.exp+test.{wmt13,wmt16}+tok.13a+version.1.5.1

	En→De				**En→Ga**	
	Data size = 30k		**Data size = 100k**			
Pre-training data	**valid**	**test**	**valid**	**test**	**valid**	**test**
N/A (baseline)	4.4 ± 0.65	4.2 ± 0.95	16.3 ± 0.31	20.1 ± 0.51	4.4 ± 0.32	8.4 ± 0.60
Real-world data						
English	**11.4 ± 0.12**	**14.2 ± 0.15**	**17.0 ± 0.10**	**21.6 ± 0.17**	**8.5 ± 0.83**	**14.1 ± 0.72**
German	11.1 ± 0.40	13.3 ± 0.68	16.6 ± 0.23	20.8 ± 0.21	N/A	N/A
English shuf	11.1 ± 0.00	13.5 ± 0.12	15.7 ± 0.15	19.6 ± 0.23	4.7 ± 0.80	11.5 ± 0.64
Artificial data						
Random	10.6 ± 0.10	13.0 ± 0.10	16.0 ± 0.00	19.9 ± 0.17	6.5 ± 0.64	10.6 ± 1.00
Zipf	10.7 ± 0.15	13.5 ± 0.10	15.7 ± 0.15	19.6 ± 0.21	8.0 ± 0.53	12.8 ± 0.40

Table 1: BLEU scores of English→German (En→De) and English→Irish (En→Ga) translation models for each pre-training dataset. We report the mean and standard deviation of three runs.

case-insensitive BLEU[8] for evaluation.

The preprocessing and training settings for both situations are in the Appendix.

Vocabulary assignment The parallel data consists of natural language tokens, whereas the artificial data contains only integer tokens. Consequently, the vocabularies learned from artificial and parallel data exhibit no overlap, which is a bottleneck in transferring the embedding layers. To solve this issue, we adopt frequency assignment (Aji et al., 2020), which sorts integer tokens and natural language tokens based on their frequency in each training dataset, respectively, and assigns the integer token to the natural language token having the same frequency rank. For the vocabulary used in pre-training with real-world data, we use the one created from parallel data.

4 Results

4.1 Simulated low-resource situation

Table 1 shows the English→German translation performance for the 30k and 100k parallel dataset sizes. For the sake of simplicity, we refer to the model without pre-training as the *baseline*, and the group comprising "English shuf", "Random" and "Zipf" datasets as *non-natural language*.

Data size: 30k All pre-trained models outperform the baseline. However, the models pre-trained with the "German" dataset and non-natural languages are inferior to the "English" model. Although the "German" model is pre-trained on a natural language, it performs as well as the "English shuf" model. The "English shuf" and "Zipf" models gain comparable performance on the test set, and both outperform the "Random" model. This indicates that token frequency information in pre-training data contributes to improved performance. Translation examples are in the Appendix (Table 5).

Data size: 100k "English" and "German" models outperform the baseline; the other models degrade from the baseline performance. In contrast to the case of the 30k-sized dataset, where token frequency information contributes to the performance gain, the scores of the models pre-trained on the non-natural languages are all comparable.

4.2 Genuine low-resource situation

From Table 1, all pre-trained models outperform the baseline, and the "English" model achieves the highest score. As both the "English shuf" and "Zipf" models outperform the "Random" model on the test set, we conclude that token frequency information in pre-training data is advantageous when the parallel data size is quite small, considering the results with a data size of 30k in the simulated low-resource situation.

5 Analysis

We investigate which parts of the NMT model pre-trained on artificial data contribute to improved performance and verify whether the effect of each part on translation performance differs from the case where pre-trained on real-world data.

Specifically, we divide the model parameters into four components: embeddings (emb), encoders (enc), cross-attentions (x-attn), and decoders except for cross-attentions (dec), and perform two ablation studies. Firstly, we transfer a part of the components from a pre-trained model and fine-tune the model (Table 2a). This is done to iden-

[8] Signature: BLEU+case.lc+lang.en-ga+numrefs.1 +smooth.exp+tok.13a+version.1.5.1

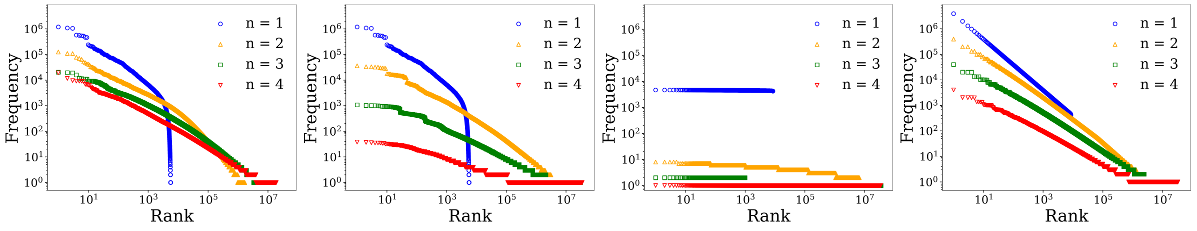
	Components				Pre-training data			
Row	emb	enc	x-attn	dec	English	English shuf	Random	Zipf
1					4.2 ± 0.95	4.2 ± 0.95	4.2 ± 0.95	4.2 ± 0.95
2	✓				6.2 ± 0.64	5.6 ± 0.20	4.9 ± 0.26	5.8 ± 0.15
3		✓	✓	✓	11.5 ± 0.25	12.7 ± 0.12	12.7 ± 0.23	12.7 ± 0.17
4	✓	✓	✓	✓	14.2 ± 0.15	13.5 ± 0.12	13.0 ± 0.10	13.5 ± 0.10

(a) When transferring only some components from each pre-trained model. "✓" denotes that the corresponding component is transferred. The full version is in the Appendix (Table 6).

	Components				Pre-training data			
Row	emb	enc	x-attn	dec	English	English shuf	Random	Zipf
1					14.2 ± 0.15	13.5 ± 0.12	13.0 ± 0.10	13.5 ± 0.10
2		×			10.7 ± 0.15	11.7 ± 0.30	12.2 ± 0.06	11.6 ± 0.10
3				×	12.5 ± 0.32	11.2 ± 0.35	10.7 ± 0.47	11.5 ± 0.10

(b) When freezing each component of the fully transferred model. "×" denotes that the corresponding component is frozen. The full version is in the Appendix (Table 7).

Table 2: BLEU scores on the test set of English→German translation models in two ablation studies.



(a) English (b) English shuf (c) Random (d) Zipf

Figure 2: Rank–frequency distribution of token n-grams ($n = 1, 2, 3, 4$) in each pre-training dataset.

tify whether the information from each component obtained in pre-training encourages better training in the fine-tuning step. Secondly, we transfer all components and freeze one specific component during fine-tuning (Table 2b). This is done to identify whether each component's information obtained in pre-training is sufficient to perform the translation task. We conducted experiments on English→German pair with a parallel data size of 30k. For the pre-training datasets, we employed those except for the "German" dataset, as the performance achieved when pre-trained on this dataset was inferior to that when pre-trained on the "English" dataset.

Token frequency information imparts embeddings with beneficial information for translation From Table 2a, it can be seen that when pre-training with the "English" dataset, transferring emb improves performance (row 2). Similarly, even when pre-training with the "English shuf" and "Zipf" datasets, we observe that transferring emb contributes to improved performance, although tokens in these datasets are independent of each other. On the other hand, when pre-training with the "Random" dataset, the performance gains by transferring emb are negligible (rows 3 and 4).

Both "English shuf" and "Zipf" datasets contain token frequency information from the real world; that is, the distribution of token frequency follows Zipf's law. This brings about multiple occurrences of the same n-gram to a certain extent, even when the tokens in pre-training data are shuffled (Tanaka-Ishii, 2021). Figure 2 shows the rank–frequency distribution of n-grams ($n = 1, 2, 3, 4$) in each pre-training dataset.[9] Even though tokens are sampled independently to form a sequence for the "English shuf" and "Zipf" datasets, we can observe a power trend in the frequency of n-grams at $n = 2, 3, 4$. Therefore, we consider that the presence of multiple instances of the same n-gram, which results in the emergence of local contexts within a sentence, and embeddings pre-trained on this dataset obtain beneficial information for translation performance.

[9] After subword segmentation, we drew the rank–frequency distribution of the "English" dataset. The sharp decline in the curve at $n = 1$ (Figure 2a) is due to the subword segmentation, where the number of merge operations is 8,000.

Encoders pre-trained on artificial data obtain enhanced representations to understand the inputs From Table 2b, we can observe that the superiority tendency in scores among the pre-training dataset is reversed between the cases in freezing enc (row 2) and dec (row 3). The score tendency when freezing enc is "Random" > "English shuf" ≈ "Zipf" > "English", while for when freezing dec, we observe "English" > "English shuf" ≈ "Zipf" > "Random".

We attribute this tendency to the mechanism of the MASS pre-training depending on data property. When pre-training with the "English" dataset, the dec predicts a masked span of an English text that contains linguistic traits like structure, which makes the dec's prediction easier. Therefore, the dec can make predictions without requiring much information from the enc, which makes the enc understand the input sequence moderately. This explains why the best score is achieved when freezing dec and the worst score is achieved when freezing enc compared to other datasets. However, when pre-training with other datasets in which the tokens in a sequence are independent of each other, it is challenging for the dec to predict a masked span autoregressively. This incentivizes the dec to extract beneficial information for predictions from the enc; that is, the dec relies more on enc's information. This encourages the enc to understand the input sequence more, and transferring enc enhanced to capture the input meaning results in higher translation performance. This consideration is consistent with the assertion of Sánchez-Cartagena et al. (2021).

6 Conclusion

In this work, we chose MASS for the pre-training task and explored the effects on translation performance in low-resource situations when pre-training the NMT model on artificial data. Both in simulated (English→German) and genuine (English→Irish) low-resource situations, pre-training with artificial data improved the performance, and further improvements could be obtained by injecting token frequency information when the parallel data size was very small. Through ablation studies, we found that token frequency information generates contexts within a dataset, and pre-training on such datasets enables embeddings to obtain beneficial information for translation performance. In addition, pre-training on artificial datasets in which tokens are independent of each other enhances the capability of encoders to understand inputs, resulting in improved translation performance.

Limitations

Natural language The languages we used for parallel data (English, German, and Irish) are alphabetical. This aspect affects the learning behavior of a translation model, because we jointly learn BPE on both the source and target languages and share all the embedding parameters during pre-training and fine-tuning. Therefore, it is unclear whether the MASS pre-training with artificial data contributes to the gains in translation performance when using non-alphabetic languages such as Japanese and Chinese as the source or target languages.

Artificial data The tokens in artificial data we used in this study are independent of each other; they do not possess linguistic traits like co-occurrence and structure. Ri and Tsuruoka (2022) showed that a Transformer-based causal language model trained on artificial data containing information of co-occurrence and structure between tokens results in lower perplexity than the model trained on artificial data without such information. The model pre-trained on artificial data that contains linguistic information, such as co-occurrence and structure, may behave similarly to that pre-trained on the "English" dataset.

The contents of artificial datasets change depending on the seed value; however, we created each dataset with one seed in this work; we additionally conducted pre-training once on each dataset. Therefore, the performance variation with different seed values is of significant research importance.

Acknowledgements

We thank our reviewers for their thoughtful and instructive comments. This work was partly supported by TMU research fund for young scientists.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In Neural Machine Translation, What Does Transfer Learning Transfer? In *ACL*.

Cheng-Han Chiang and Hung-yi Lee. 2022. On the Transferability of Pre-trained Language Models: A Study from Artificial Datasets. In *AAAI*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation. In *PACLIC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL*.

Kundan Krishna, Jeffrey Bigham, and Zachary C. Lipton. 2021. Does Pretraining for Summarization Require Knowledge Transfer? In *Findings of EMNLP*.

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 *ACL*, pages 7871–7880.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *BlackboxNLP*.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *PNAS*, 117(48):30046–30054.

John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu, editors. 2021. *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *WMT*.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *WMT*.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation? In *NAACL*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *EMNLP*.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models. In *ACL*.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach. In *EMNLP*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *ICML*.

Kumiko Tanaka-Ishii. 2021. *Statistical Universals of Language: Mathematical Chance and Human Choice*. Springer.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *ACL*.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *EMNLP*.

A Appendix

A.1 Detailed experimental setup

Preprocessing settings In a simulated low-resource situation, we normalized punctuations and tokenized the text with Moses[10] (Koehn et al., 2007) scripts, and subworded the output with BPE (Sennrich et al., 2016) jointly learned on parallel data. The vocabulary size of BPE was 8,000 for both 30k and 100k sizes.

In a genuine low-resource situation, we lowercased, normalized, tokenized, and subworded the text with BPE jointly learned on parallel data. The vocabulary size of BPE was 8,000.

Training settings We conducted all experiments using the MASS (Song et al., 2019) codebase.[11] For the training procedure, we pre-trained the model with the MASS objective, initialized the NMT model with the weights of the pre-trained model, and fine-tuned it on parallel data (Figure 1). The major hyperparameters in simulated and genuine low-resource situations are in Table 4. The number of pre-training updates was 100k. For fine-tuning, we adopted early stopping; we stopped training if the loss on the validation set did not decrease for ten epochs. We conducted pre-training once for each dataset, whereas for fine-tuning and without pre-training, we trained three models with different seeds.

A.2 Evaluation with other metrics

We evaluated translation performance with chrF (Popović, 2015) and COMET (Rei et al., 2020). For chrF, we used SacreBLEU to calculate scores.[12] For COMET, we selected `wmt22-comet-da` as an evaluation model to measure scores.[13] In a genuine low-resource situation, we performed evaluations on lowercased texts with both metrics. Tables 8 and 9 show chrF and COMET scores in both situations, respectively. Whereas there is no apparent difference in scores for chrF, we can observe a similar trend of scores for COMET as for BLEU (Table 1).

A.3 Comparison to other pre-training methods

Following Aji et al.'s (2020) work, we examined the translation performance when pre-trained with autoencoding (AE) and one-to-one substitution (SBST) in both low-resource situations. Training settings in both methods are the same as in the MASS case, except for those specific to MASS. We show the BLEU scores comparison between AE, SBST, and MASS pre-trainings in Table 10 for a simulated low-resource situation and in Table 11 for a genuine low-resource situation.

[10] https://github.com/moses-smt/mosesdecoder

[11] https://github.com/microsoft/MASS

[12] Signature: chrF2+lang.en-de+numchars.6+space.false +test.{wmt13,wmt16}+version.1.5.1

[13] https://github.com/Unbabel/COMET

Pre-training data	Sentence
Real-world data	
English	In that time he had not thought once about new vision .
German	Ich weiß nicht , wie gut er einmal werden kann .
English shuf	name the p@@ from N@@ ding ia &apos; w@@ and ordin@@ Ad@@
Artificial data	
Random	8246 1658 3000 1199 7351 8414 2680 3917 7361 4130 2285 1561
Zipf	5 415 31 66 6 237 330 5 258 27 186 71

Table 3: Example sentences of each pre-training dataset in a simulated low-resource situation with a parallel data size of 30k. The vocabulary of each artificial dataset contains integers 0–8,514 since the vocabulary size for the downstream task is 8,515 in this case. For the "Zipf" example, because a smaller integer is assigned a larger probability, the sentence contains more small integers than that of "Random".

	Simulated low-resource			**Genuine low-resource**		
Parameter	**w/o PT**	**PT**	**FT**	**w/o PT**	**PT**	**FT**
encoder layers		6			4	
decoder layers		6			4	
hidden size		512			256	
feed-forward size		2,048			2,048	
attention heads		8			8	
learning rate	5e-4	1e-4	1e-4	1e-4	1e-4	1e-4
dropout	0.3	0.1	0.3	0.3	0.1	0.3
word mask	N/A	0.5	N/A	N/A	0.5	N/A
warmup steps		4,000		2,000	4,000	2,000
batch size		4,096 × 8 tokens			4,096 × 8 tokens	
beam size		4			4	

Table 4: Hyperparameters used in simulated (English→German) and real (English→Irish) low-resource situations. "word mask" is the ratio that controls the masking length of an input sequence in MASS pre-training. "PT" denotes pre-training and "FT" denotes fine-tuning.

	Example 1
Source	But it's a different story among the American public overall.
Reference	Aber es ist eine andere Geschichte in der amerikanischen Öffentlichkeit insgesamt.
Pre-training data	
N/A	Aber es handelt sich um eine andere Seite der amerikanischen Öffentlichkeit.
English	Aber es ist eine andere Geschichte unter der amerikanischen Öffentlichkeit.
German	Aber es handelt sich um eine andere Story zwischen der amerikanischen Öffentlichkeit.
English shuf	Aber es ist eine andere Geschichte der amerikanischen Öffentlichkeit.
Random	Aber es handelt sich um eine andere Geschichte der amerikanischen Öffentlichkeit.
Zipf	Aber es ist eine andere Geschichte unter der amerikanischen Öffentlichkeit.
	Example 2
Source	Here are the different ways to send in your contributions:
Reference	Hier sind die verschiedenen Möglichkeiten, Ihre Beiträge zu senden:
Pre-training data	
N/A	Hier finden Sie verschiedene Beiträge in Ihren Beiträge.
English	Hier sind die unterschiedlichen Möglichkeiten, Ihre Beiträge zu stellen:
German	Hier gibt es die unterschiedlichen Möglichkeiten, in Ihren Beiträge hinzuzufügen:
English shuf	Hier sind die verschiedenen Möglichkeiten, Ihre Beiträge zu senden:
Random	Hier finden Sie die verschiedenen Möglichkeiten, sich in Ihrem Beiträge zu senden:
Zipf	Hier sind die unterschiedlichen Möglichkeiten, um Ihre Beiträge zu senden:

Table 5: English→German translation examples on the test set for each pre-training dataset with the parallel data size of 30k.

	Components				**Pre-training data**			
Row	**emb**	**enc**	**x-attn**	**dec**	**English**	**English shuf**	**Random**	**Zipf**
1					4.2 ± 0.95	4.2 ± 0.95	4.2 ± 0.95	4.2 ± 0.95
2		✓			8.0 ± 0.49	8.3 ± 0.49	9.9 ± 0.21	8.7 ± 0.21
3			✓		0.8 ± 0.00	5.8 ± 0.36	5.7 ± 0.26	5.3 ± 0.31
4				✓	3.8 ± 0.00	5.9 ± 0.12	5.8 ± 0.06	5.5 ± 0.21
5		✓	✓		10.3 ± 0.31	10.6 ± 0.25	11.8 ± 0.26	11.3 ± 0.15
6		✓		✓	8.3 ± 0.31	10.6 ± 0.35	11.1 ± 0.30	10.3 ± 0.30
7			✓	✓	5.5 ± 1.42	8.5 ± 0.20	8.0 ± 0.36	8.5 ± 0.42
8		✓	✓	✓	11.5 ± 0.25	12.7 ± 0.12	12.7 ± 0.23	12.7 ± 0.17
9	✓				6.2 ± 0.64	5.6 ± 0.20	4.9 ± 0.26	5.8 ± 0.15
10	✓	✓			12.2 ± 0.15	10.6 ± 0.10	9.8 ± 0.10	10.8 ± 0.06
11	✓		✓		7.7 ± 1.70	7.5 ± 0.06	7.1 ± 0.17	7.9 ± 0.20
12	✓			✓	5.4 ± 0.57	7.6 ± 0.20	6.4 ± 0.17	8.6 ± 0.32
13	✓	✓	✓		14.0 ± 0.15	12.2 ± 0.10	12.4 ± 0.31	12.2 ± 0.42
14	✓	✓		✓	11.7 ± 0.26	11.9 ± 0.45	11.3 ± 0.35	11.8 ± 0.26
15	✓		✓	✓	9.2 ± 0.32	10.0 ± 0.40	8.9 ± 0.12	11.3 ± 0.12
16	✓	✓	✓	✓	14.2 ± 0.15	13.5 ± 0.12	13.0 ± 0.10	13.5 ± 0.10

Table 6: BLEU scores of English→German translation models when transferring only some components from each pre-trained model. The parallel data size is 30k, and the evaluation was performed on the test set. We report the mean and standard deviation of three runs. "✓" denotes that the corresponding component is transferred.

	Components				**Pre-training data**			
Row	**emb**	**enc**	**x-attn**	**dec**	**English**	**English shuf**	**Random**	**Zipf**
1					14.2 ± 0.15	13.5 ± 0.12	13.0 ± 0.10	13.5 ± 0.10
2	×				5.5 ± 0.06	10.3 ± 0.17	10.7 ± 0.10	10.8 ± 0.47
3		×			10.7 ± 0.15	11.7 ± 0.30	12.2 ± 0.06	11.6 ± 0.10
4			×		12.8 ± 0.23	12.5 ± 0.15	12.1 ± 0.10	12.7 ± 0.25
5				×	12.5 ± 0.32	11.2 ± 0.35	10.7 ± 0.47	11.5 ± 0.10

Table 7: BLEU scores of English→German translation models when freezing each component of the fully transferred model. The parallel data size is 30k, and we performed the evaluation on the test set. We report the mean and standard deviation of three runs. "×" denotes that the corresponding component is frozen.

	En→De				**En→Ga**	
	Data size = 30k		**Data size = 100k**			
Pre-training data	**valid**	**test**	**valid**	**test**	**valid**	**test**
N/A (baseline)	0.29 ± 0.02	0.28 ± 0.02	0.45 ± 0.01	0.49 ± 0.01	0.28 ± 0.01	0.33 ± 0.01
Real-world data						
English	**0.40 ± 0.00**	**0.43 ± 0.00**	**0.46 ± 0.00**	**0.50 ± 0.00**	0.36 ± 0.02	**0.39 ± 0.02**
German	0.39 ± 0.01	0.41 ± 0.01	**0.46 ± 0.01**	0.49 ± 0.00	N/A	N/A
English shuf	**0.40 ± 0.00**	**0.43 ± 0.00**	0.45 ± 0.00	0.49 ± 0.00	0.31 ± 0.02	0.37 ± 0.01
Artificial data						
Random	**0.40 ± 0.00**	**0.43 ± 0.00**	0.45 ± 0.01	0.49 ± 0.01	0.32 ± 0.02	0.35 ± 0.02
Zipf	**0.40 ± 0.00**	**0.43 ± 0.00**	0.45 ± 0.00	0.49 ± 0.00	**0.37 ± 0.02**	**0.39 ± 0.01**

Table 8: chrF scores of English→German (En→De) and English→Irish (En→Ga) translation models for each pre-training dataset. For En→Ga, evaluation is conducted on lowercased texts. We report the mean and standard deviation of three runs.

	En→De				En→Ga	
	Data size = 30k		Data size = 100k			
Pre-training data	valid	test	valid	test	valid	test
N/A (baseline)	−1.31 ± 0.07	−1.39 ± 0.07	−0.23 ± 0.02	−0.25 ± 0.03	−1.03 ± 0.03	−0.76 ± 0.06
Real-world data						
English	**−0.71 ± 0.01**	**−0.73 ± 0.02**	**−0.18 ± 0.01**	**−0.18 ± 0.01**	**−0.75 ± 0.04**	**−0.44 ± 0.06**
German	−0.73 ± 0.02	−0.74 ± 0.03	−0.23 ± 0.02	−0.24 ± 0.02	N/A	N/A
English shuf	−0.77 ± 0.01	−0.80 ± 0.01	−0.29 ± 0.03	−0.31 ± 0.03	−0.97 ± 0.04	−0.67 ± 0.05
Artificial data						
Random	−0.81 ± 0.01	−0.84 ± 0.01	−0.29 ± 0.01	−0.32 ± 0.02	−1.02 ± 0.06	−0.81 ± 0.06
Zipf	−0.77 ± 0.00	−0.81 ± 0.00	−0.29 ± 0.02	−0.31 ± 0.02	−0.86 ± 0.05	−0.61 ± 0.06

Table 9: COMET scores of English→German (En→De) and English→Irish (En→Ga) translation models for each pre-training dataset. For En→Ga, evaluation is conducted on lowercased texts. We report the mean and standard deviation of three runs.

	En→De					
	Data size = 30k			Data size = 100k		
Pre-training data	AE	SBST	MASS	AE	SBST	MASS
N/A (baseline)	4.2 ± 0.95	4.2 ± 0.95	4.2 ± 0.95	**20.1 ± 0.51**	**20.1 ± 0.51**	20.1 ± 0.51
Real-world data						
English	**13.7 ± 0.15**	**13.3 ± 0.06**	**14.2 ± 0.15**	19.6 ± 0.21	19.0 ± 0.17	**21.6 ± 0.17**
German	13.6 ± 0.00	12.8 ± 0.06	13.3 ± 0.68	19.1 ± 0.21	18.8 ± 0.06	20.8 ± 0.21
English shuf	13.6 ± 0.12	12.8 ± 0.26	13.5 ± 0.12	19.4 ± 0.12	19.0 ± 0.29	19.6 ± 0.23
Artificial data						
Random	13.2 ± 0.20	10.7 ± 0.31	13.0 ± 0.10	19.0 ± 0.12	19.6 ± 0.10	19.9 ± 0.17
Zipf	13.3 ± 0.10	10.3 ± 0.26	13.5 ± 0.10	19.1 ± 0.17	18.4 ± 0.06	19.6 ± 0.21

Table 10: Comparison of BLEU scores for English→German (En→De) translation models by pre-training objectives. “AE” denotes autoencoding, and “SBST” denotes one-to-one substitution. We report the mean and standard deviation of three runs on the test set.

	En→Ga		
Pre-training data	AE	SBST	MASS
N/A (baseline)	8.4 ± 0.60	8.4 ± 0.60	8.4 ± 0.60
Real-world data			
English	**10.0 ± 0.01**	8.4 ± 0.91	**14.1 ± 0.72**
English shuf	8.8 ± 1.25	7.8 ± 0.42	11.5 ± 0.64
Artificial data			
Random	6.9 ± 1.08	7.1 ± 0.81	10.6 ± 1.00
Zipf	8.7 ± 1.40	**9.0 ± 0.32**	12.8 ± 0.40

Table 11: Comparison of BLEU scores for English→Irish (En→Ga) translation models by pre-training objectives. “AE” denotes autoencoding, and “SBST” denotes one-to-one substitution. We report the mean and standard deviation of three runs on the test set.