MLM WITH GLOBAL CO-OCCURRENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

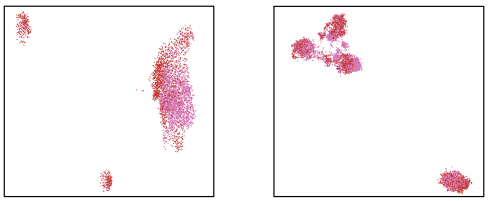
Global co-occurrence information is the primary source of structural information on multilingual corpora, which potentially gives useful structural similarities across languages to the model for cross-lingual transfer. In this work, we push MLM (masked language modeling) pre-training further to leverage global co-occurrence information on multilingual corpora. The result is MLM-GC (MLM with Global Co-occurrence) pre-training that the model learns local bidirectional information from MLM and global co-occurrence information from a log-bilinear regression. We find that analogical co-occurred words across languages have similar co-occurrence counts/frequencies (normalized) giving weak but stable self-supervision for cross-lingual transfer. Experiments show that MLM-GC pre-training substantially outperforms MLM pre-training for 4 downstream multilingual/cross-lingual tasks and 1 additional monolingual task, showing the advantages of capturing embedding analogies.

1 INTRODUCTION

MLM attempts to understand bidirectional information (Devlin et al., 2019) surrounding the masked tokens without specifically setting the receptive field. Empirical studies (Lample et al., 2018a; Conneau et al., 2020a;c) show multilinguality and cross-linguality emerge from MLM pre-training on multilingual corpora without any supervision. The model is trained/pre-trained as a generator that yields masked token probabilities over the vocabulary. As an alternative, we present MLM-GC (MLM with Global Co-occurrence) with the combined objective of the generator and a global log-bilinear regression for multilingual pre-training. Our starting point is from two observations on multilingual MLM pre-training.

Language's structural information is every property of an individual language that is invariant to the script of the language. Conneau et al. (2020c); Karthikeyan et al. (2020); Sinha et al. (2021); Pires et al. (2019) show that structural similarities across languages can contribute to cross-lingual transfer. *n-gram* or co-occurrence information is the primary source of structural information available to all methods. Some methods like span-based masking (Devlin et al., 2019; Joshi et al., 2020; Levine et al., 2021) now exist to leverage this information for masking in *monolingual* MLM pre-training, aiming at improving context understanding. However, in *multilingual* MLM pre-training, the question still remains as to how meaning is generated from these statistics on multilingual corpora, how the structural similarities could be learned from that meaning across languages, and how cross-lingual transfer might be improved from that meaning.

Furthermore, we run naive MLM pre-training on $\{En, De\}$ corpora in preliminary experiments. Since MLM pre-training can form a cross-lingual embedding space (Lample & Conneau, 2019; Artetxe et al., 2020a), ideally, analogical embeddings across languages should be clustered because they potentially represent similar structural information and general meanings. However, the t-SNE projection in Figure 1 (a) reports that they are not successfully clustered in the box. We suspect that naive MLM pre-training on multilingual corpora cannot satisfactorily capture and understand analogical embeddings and morphological variants across languages, which results in limitations of cross-lingual transfer. Intuitively, leveraging global co-occurrence information might solve this problem and then improve cross-lingual transfer on multilingual corpora, given that: *1)* it is a proven idea to search embedding analogies on monolingual corpora (Pennington et al., 2014); *2)* analogical words and co-occurred words across languages may have similar frequencies/counts on the multilingual corpora as indicated by Zipf's law (Ha et al., 2002; Søgaard, 2020), which allows for better cross-lingual and multilingual representations.



(a) MLM (b) MLM-GC

Figure 1: Preliminary experiment of MLM pre-training on $\{De, En\}$ corpora and t-SNE visualization. *light*: De, *dark*: En. Analogical embeddings are not clustered well after MLM pre-training. By contrast, the clustering phenomenon is observed after MLM-GC pre-training. Further discussions and full-sized figures are in the Cross-lingual Embedding experiment and Appendix §B.

To this end, we present MLM-GC to utilize global co-occurrence information. MLM-GC builds on MLM with an extra objective of global log-bilinear regression that minimizes the error between scores of the model's pre-softmax linear transformation and the matrix of global co-occurrence counts. Since MLM only needs to predict masked tokens, we factorize both the matrices, only using vectors relating to the masked tokens. The result is MLM-GC pre-training with a combined objective of MLM and the global log-bilinear regression in pre-training. The model is encouraged to learn bidirectional information from MLM and the global co-occurrence information from the global log-bilinear regression. On multilingual corpora, MLM-GC pre-training can improve cross-lingual transfer because analogical co-occurred words across languages might have similar co-occurrence counts/frequencies allowing for cross-lingual representations.

We have four contributions. **1)** We present MLM-GC pre-training for multilingual tasks. The model is additionally supervised by global co-occurrence information on multilingual corpora. **2)** MLM-GC pre-training outperforms MLM pre-training on 4 multilingual/cross-lingual tasks. The objective of MLM-GC can be adapted to encoder-decoder-based MLM models, e.g., MASS (Song et al., 2019) and encoder-based MLM models, e.g., XLM (Lample & Conneau, 2019). MLM-GC pre-training can also work on monolingual corpora for language understanding tasks. **3)** We report negative results of clustering embedding analogies after naive MLM pre-training. By contrast, the model is encouraged to understand embedding analogies across languages after MLM-GC pre-training, which is potentially useful for cross-lingual and multilingual tasks. **4)** We find that global co-occurrence counts contribute to structural similarities across languages for cross-lingual transfer. Our empirical study shows that analogical co-occurred words across languages have similar co-occurrence counts/frequencies (normalized) giving weak but stable self-supervision for cross-lingual transfer.

2 Related Work and Comparison

Structural Similarity and Zipf's Law Morphologies, word-order, word frequencies, and co-occurred word frequencies are all parts of structure of a language and invariant to the script of the language. Zipf's law (Zipf, 1949; 2013) indicates that words or phrases appear with different frequencies, and one may suggest analogical words or phrases appear with relatively similar frequencies in other languages. In multilingual MLM pre-training, Conneau et al. (2020c); Karthikeyan et al. (2020); Pires et al. (2019); Karthikeyan et al. (2020); Sinha et al. (2021) shed light on studying structural information and find that structural similarities across languages are essential for cross-linguality and multilinguality, where in this case, structural similarities might mean similar frequencies as Zipf's law indicated. We follow this line, consider structural similarities from co-occurrence counts, and provide an empirical study to observe how the model learns structural similarities from global co-occurrence counts (§3.3) on multilingual corpora. Meanwhile, GloVe (Pennington et al., 2014) report that co-occurrence counts or frequencies can provide regularities for embeddings to understand word analogies and morphological variations for monolingual tasks. We extend the scope of GloVe to contextualized representations and *multilingual* tasks, helping the model understand analogical embeddings across languages in pre-training.

N-gram, Co-occurrence, and Regularity in MLM pre-training Studying co-occurrence or *n-gram* is not a novel idea in MLM pre-training. Whole Word Masking (Devlin et al., 2019), Span-

BERT (Joshi et al., 2020), and PMI-Masking (Levine et al., 2021) suggest *n-gram* spans across several sub-tokens for masking to improve context understanding in monolingual tasks because the model may only learn from easier multi-tokens instead of usefully hard context, where easier multi-tokens are in a subset of the context and result in sub-optimization. By contrast, we show that co-occurrence counts can refine contextualized representations for improving context understanding and allow for cross-lingual representations, suggesting a new objective for MLM pre-training instead of a new masking scheme to capture global co-occurrence information in multilingual pre-training. On the other hand, for cross-lingual transfer, the contextualized representations could be further regularized and refined by aligning cherry-picked pairs after MLM pre-training on multilingual corpora (Ren et al., 2019; Chaudhary et al., 2020; Wang et al., 2020; Cao et al., 2020; Aldarmaki & Diab, 2019; Artetxe et al., 2020a; Ai & Fang, 2021). Compared to that, MLM-GC pre-training does not require dictionaries, translation tables, or statistical machine translation models.

3 APPROACH

3.1 GLOBAL REGRESSION MODELING IN MONOLINGUAL EMBEDDING SPACE

GloVe (Pennington et al., 2014) present a log-bilinear regression model:

$$\mathcal{L} = \sum_{i,j=1}^{V} f(X_{w_i w_j})(E_{w_i}^T E_{w_j} + b_{w_i} + b_{w_j} - \log X_{w_i w_j})^2, \tag{1}$$

where $f(x) = \begin{cases} (x/x_{max})^\alpha, x < x_{max} \\ 1, otherwise \end{cases}$, V is the vocabulary, E_w is the embedding of token w, b_{w_i} and b_{w_j} are bias to restore the symmetry, X stands for the matrix of token-token co-occurrence counts, entries $X_{w_i w_j}$ tabulate the number of times token w_j occurs in the context of token w_i, and x_{max} is empirically set to 100. The global regression model derives from the skip-gram $softmax$ probability $Q_{w_i w_j} = \frac{exp(E_{w_i}^T E_{w_j})}{\sum_{k=1}^{V} exp(E_{w_i}^T E_{w_k})}$ and factorizes the log of the global co-occurrence matrix, where $Q_{w_i w_j}$ is a model for the probability that w_j appears in the context of w_i. The model is encouraged to understand the correspondence between two embeddings E_{w_i} and E_{w_j} from the co-occurrence counts $X_{w_i w_j}$ on the corpora.

3.2 GLOBAL CO-OCCURRENCE MODELING IN MULTILINGUAL MLM PRE-TRAINING

In MLM pre-training, when w_t at the position t is replaced by the artificial masking token $[\mathcal{M}]_t$, the output distribution for w_t is obtained by applying a pre-softmax linear transformation $O \in R^{d \times V}$ from *the final hidden state or the contextualized representation* $H_{[\mathcal{M}]_t}$ to the output vocabulary size V, followed by a $softmax$ operation which generates an output matrix normalized over its rows. Specifically, $Q_{[\mathcal{M}]_t w_t} = \frac{exp(H_{[\mathcal{M}]_t}^T O_{w_t})}{\sum_{k=1}^{V} exp(H_{[\mathcal{M}]_t}^T O_{w_k})}$ is the model for the probability of w_t in the context of $H_{[\mathcal{M}]_t}$, where O_{w_t} and O_{w_k} are a vector factorized from O, i.e., self-recognizing. Since w_t at position t is replaced by $[\mathcal{M}]_t$, the model is encourage to consider bidirectional information for outputting $H_{[\mathcal{M}]_t}$ (Devlin et al., 2019). Considering the sub-optimization and limitation of capturing co-occurrence or bidirectional information (Devlin et al., 2019; Joshi et al., 2020; Levine et al., 2021), we further consider a model for the probability of a neighboring token w_n to be considered as the bidirectional information. In this way, the probability of w_n in the context $H_{[\mathcal{M}]_t}$ is similar to $Q_{w_i w_j}$ in the global regression model. Specifically, for w_n, $Q_{[\mathcal{M}]_t w_t}$ could be extended to:

$$Q_{[\mathcal{M}]_t w_n} = \frac{exp(H_{[\mathcal{M}]_t}^T O_{w_n})}{\sum_{k=1}^{V} exp(H_{[\mathcal{M}]_t}^T O_{w_k})}. \tag{2}$$

For all the neighboring tokens $w_{t \pm n}$ of the input sentence at position $[t-n, ..., t) \cup (t, ..., t+n]$, i.e., excluding position t, we have the model $Q_{[\mathcal{M}]_t w_{t \pm n}}$. Then, we employ the new global log-bilinear regression model in MLM pre-training. Formally, given the factorized $O_{w_{t \pm n}}$ and $X_{w_t w_{t \pm n}}$ from O and X respectively, we have the model:

$$\mathcal{L}_{GC} = \frac{1}{2n} \sum_{n} f(X_{w_t w_{t \pm n}})(\frac{H_{[\mathcal{M}]_t}^T O_{w_{t \pm n}}}{\sqrt{d}} - \log X_{w_t w_{t \pm n}})^2, \tag{3}$$

where d is the model dimension. Compared to Eq.1, we add scaling $\sqrt{d}$ and weight $\frac{1}{2n}$ to make training stable, where $\sqrt{d}$ is inspired by scaled dot-product attention (Vaswani et al., 2017) to prevent the dot products get large. They serve as hyperparameters for $\mathcal{L}_{GC}$.

To obtain the matrix of token-token co-occurrence counts on multilingual corpora for multilingual tasks, we follow GloVe's suggestion that a distance weight scheme is employed. Specifically, in a context window of size $2n+1$, we calculate the token-token co-occurrence counts for positions $[k-n, ..., k, ..., k+n]$ with the rule $[c_{lang}/n+1, ..., c_{lang}/2, 0, c_{lang}/2, ..., c_{lang}/n+1]$ over the share vocabulary, which means we do not calculate the unigram counts or self-co-occurrence X_{kk} for the centric token w_k at position k. Meanwhile, we are aware that the probability is fake and is equivalent to token-token co-occurrence counts, similar to GloVe, and not all the languages have the same amount of samples in the corpora (e.g., low-resource v.s. high resource). Considering this, we use the language-wise constant $c_{lang} = C_{En}/C_{lang}$, where C_{En} is the total number of tokens in English corpora, and C_{lang} is the total number of tokens in the language $lang$. As Zipf's law holds for frequencies, c_{lang} extends Zipf's law to counts when computing token-token co-occurrence counts on the multilingual corpora, i.e., *co-occurrence counts are normalized by* c_{lang}.

3.3 MULTILINGUAL MLM-GC PRE-TRAINING

In pre-training, we have a combined objective of MLM and global co-occurrence modeling, attempting to train the model to understand the masked tokens from bidirectional information and linguistic structures surrounding the masked tokens from global co-occurrence counts, and the resulting trainer is our MLM-GC. Formally, we have the model:

$$\begin{aligned} \mathcal{L}_{MLM-GC} &= \mathcal{L}_{MLM} + \mathcal{L}_{GC} \\ &= \sum_t (-\log Q_{[\mathcal{M}]_t w_t} + \frac{1}{2n} \sum_n f(X_{w_t w_{t\pm n}}) (\frac{H^T_{[\mathcal{M}]_t} O_{w_{t\pm n}}}{\sqrt{d}} - \log X_{w_t w_{t\pm n}})^2). \end{aligned} \tag{4}$$

In the early experiment, we add a hyperparameter $\lambda \in \{0.1, 0.5, 1, 2\}$ to $\lambda\mathcal{L}_{GC}$. We find $\lambda = 1$ is a general choice for experiments. On the other hand, we find $warm\_up$ (Vaswani et al., 2017) of lr, $\sqrt{d}$, and $\frac{1}{2n}$ (Eq. 3) are significant. The model may collapse to $\mathcal{L}_{GC}$ without $warm\_up$, $\sqrt{d}$, or $\frac{1}{2n}$ because $\mathcal{L}_{GC}$ converges too fast and is unstable. See the early experiment in Appendix E.

Improved Contextualized Representation $Q_{[\mathcal{M}]_t w_{t\pm n}}$ considers the correspondence in the context $[t-n, ..., t) \cup (t, ..., t+n]$ with an explicit objective. In this way, the model is encouraged to learn from usefully hard context instead of easier multi-tokens under the supervision from co-occurrence information, where easier multi-tokens are in a subset of the context and result in sub-optimization (Levine et al., 2021) as discussed in §Related Work.

Improved Cross-lingual Transfer With the objective of $\mathcal{L}_{GC}$, we aim at associating $H^T_{[\mathcal{M}]_t} O_{w_{t\pm n}}$ with $H^T_{[M]_{\tilde{t}}} O_{\tilde{w}_{\tilde{t}\pm n}}$ of different languages if $X_{w_t w_{t\pm n}} = X_{\tilde{w}_{\tilde{t}} \tilde{w}_{\tilde{t}\pm n}}$, where co-occurred words $w_t w_{t\pm n}$ and co-occurred words $\tilde{w}_{\tilde{t}} \tilde{w}_{\tilde{t}\pm n}$ are analogical in different languages. In this way, it underlies the basic assumption that analogical co-occurred words across languages have similar co-occurrence counts (normalized by c_{lang}), i.e., $w_t w_{t\pm n}$ and $\tilde{w}_{\tilde{t}} \tilde{w}_{\tilde{t}\pm n}$ are analogical co-occurred words $\implies$ $X_{w_t w_{t\pm n}} = X_{\tilde{w}_{\tilde{t}} \tilde{w}_{\tilde{t}\pm n}}$. Although Zipf's law supports this assumption (Ha et al., 2002; Søgaard, 2020) in linguistics, we are still interested in the questions: *how it reflects on the multilingual corpora we use* and *whether analogical pair of* $w_t w_{t\pm n}$ *and* $\tilde{w}_{\tilde{t}} \tilde{w}_{\tilde{t}\pm n}$ $\implies$ $X_{w_t w_{t\pm n}} = X_{\tilde{w}_{\tilde{t}} \tilde{w}_{\tilde{t}\pm n}}$. To answer this question, we extract all (1100k) the pairs of parallel co-occurred words in En and De from the open-source translation tables (OPUS, Wikipedia v1.0)$\diamond$, e.g., "ist die" (De) and "is the" (En), and compute co-occurrence counts on $\{De, En\}$ Wikipedia dumps (the same dataset we use in our experiment of the Cross-lingual Embedding task). For any pair, we compute the absolute difference $|log(De) - log(En)|$, the sum $log(De) + log(En)$ (sorted into bins), and the ratio $|log(De) - log(En)|/(log(De) + log(En))$ for statistics in Figure 2. The figure tells us that the absolute difference avg and the ratio avg for all the pairs are relatively small and have narrow confidence (95%) intervals. Although the absolute difference avg is proportional to the sum, the ratio avg has no proportional relationship with the sum and is small throughout all the bins. Note that some pairs have low translation scores resulting in large absolute differences. The absolute difference avg is not 0, i.e., exact match for any pair. However, it still gives weak (not 0) but stable

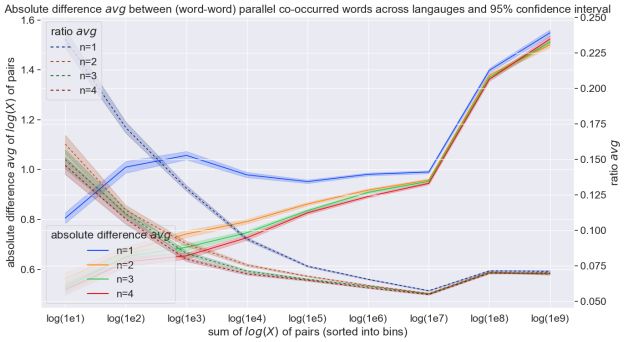


Figure 2: Study of co-occurrence counts of pairs across languages on $\{De, En\}$ corpora.

(relatively small with high confidence) supervision for cross-lingual transfer and confirms that analogical co-occurred words across languages have similar (but not identical) co-occurrence counts. Meanwhile, the model is encouraged to distinguish relevant information from irrelevant information and to discriminate between the two relevant information across languages from co-occurrence counts and refine contextualized representations accordingly, which is beneficial for cross-lingual transfer. For example, in our experiment ($n = 2$), given the translation pair "ist die" (De) and "is the" (En), the relevant pair "ist die" and "is a" (En), the irrelevant pair "ist die" and "locally known" (En), we find $|log(ist\ die) - log(is\ the)| = 0.67 < |log(ist\ die) - log(is\ a)| = 1.73 < |log(ist\ die) - log(locally\ known)| = 5.45 < |log(ist\ die) - log(En\ avg)| = 5.58$, where $log(En\ avg)$ is the avg of En co-occurrence counts.

Efficiency *1)* $\mathcal{L}_{GC}$ does not hurt training efficiency because of the factorization of O. In our experiment, for the same configuration, $\mathcal{L}_{MLM-GC}$ and $\mathcal{L}_{MLM}$ need $\approx$ 720 ms and $\approx$ 670 ms per training step time, respectively. *2)* Computing the co-occurrence matrix is laborious on large corpora. However, it requires a single pass through the entire corpora to collect the statistics, which is a one-time up-front cost and is easy to obtain new information from new corpora for updating. *3)* For memories, the co-occurrence matrix is huge, e.g., $\approx$ 11 G for a 60k BPE vocabulary with float 32. However, it is somewhat trivial because the memory is allocated on CPUs not GPUs. This can be automatically finished by DL platforms like Tensorflow. Also, the matrix can be formatted to float 16 or even float 8 by pre-$logging$ the co-occurrence counts, which will significantly reduce the memory. *4)* Meanwhile, we save the token-token co-occurrence matrix as dictionaries $\{(w_i, w_j)$: token-token co-occurrence counts$\}$ so that querying the co-occurrence counts for $X_{w_i w_j}$ is O(1).

Tokenization Sub-token-level vocabularies may impact the co-occurrence counts. In extreme cases, several connective tokens of co-occurrence may only come from one word. However, BERT (Devlin et al., 2019) reports that whole-word prediction or masking is beneficial. Similarly, even in this case, the model can be improved from the co-occurrence counts in MLM-GC pre-training.

4 Experiment

All the links of datasets, libraries, scripts, and tools marked with $\diamond$ are listed in Appendix G. A preview version of the code is submitted, and we will open the source code on GitHub.

4.1 MLM Instance, Configuration, Data Preprocessing and Pre-training

We use XLM (Lample & Conneau, 2019) and MASS (Song et al., 2019) as the MLM instances, where XLM is a token-based encoder model, and MASS is a span-based encoder-decoder model (see Appendix §C.1 for more details about MLM instances). The Transformer configuration is identical to XLM and MASS, where the dimensions of word embeddings, hidden states, and filter sizes are 1024, 1024, and 4096 respectively (**default**). Meanwhile, to be fair, we reimplement all the baseline models on our machine with our configurations, using official XLM$\diamond$, Tensor2Tensor$\diamond$,

and HuggingFace$\diamond$ as references. We compare the results of our reimplementation with the reported results on the same test set to ensure the difference less than 2% in overall performance (Appendix F). For the window or context size $2n + 1$ of the co-occurrence counts and Eq.4, **we set** $n = 2$ **for all the experiments**, which is decided by our *dev experiment*.

Data preprocessing is identical to XLM and Mass. Specifically, we employ fastBPE$\diamond$ to learn BPE (Sennrich et al., 2016b) with a sampling criteria from Lample & Conneau (2019) for all the experiments. To tokenize $\{Zh, Th, Ne\}$, we use Stanford Word Segmenter$\diamond$, PyThaiNLP$\diamond$, and Indic-NLP Library$\diamond$, respectively. For the others, we use the Moses tokenizer$\diamond$ with default rules.

Our code is implemented on Tensorflow 2.6 (Abadi et al., 2016) with 4 NVIDIA Titan Xp 12G GPU. We use Adam optimizer Kingma & Ba (2015) with $\beta_1 = 0.9$,$\beta_2 = 0.999$, $\epsilon = 1e-8$, $warm\_up = 10000$ (Vaswani et al., 2017) and $lr = 1e-4$. We set dropout regularization with a drop rate $rate = 0.1$. We accumulate gradients of 2 mini-batches per pre-training step. Since we have only 4 GPUs, this operation emulates 8 GPUs.

4.2 Multilingual Task

Readers can refer to Appendix §C.2 or references for more introductions to these tasks.

Cross-lingual Embedding We attempt SemEval'17$\diamond$ (Camacho-Collados et al., 2018) and MUSE$\diamond$ tasks (Lample et al., 2018a) that measure similarities between two paired words to generally evaluate the degree of the isomorphism of languages' embedding spaces. Meanwhile, as discussed in Lample & Conneau (2019); Wang et al. (2020) and our preliminary experiment, the performance of the isomorphism is potentially proportional to the performance of cross-lingual transfer learning tasks. Therefore, we treat this experiment as our *dev experiment* to search n.

UNMT UNMT (unsupervised neural machine translation) (Lample & Conneau, 2019; Lample et al., 2018b; Song et al., 2019; Liu et al., 2020) tackles bilingual translation (Bahdanau et al., 2015; Vaswani et al., 2017) on non-parallel bilingual corpora without any cross-lingual signal.

Cross-lingual Classification We consider XNLI$\diamond$ (Conneau et al., 2020b) on 15 languages (including English) under the cross-lingual transfer setting. The model is pre-trained on multilingual corpora and fine-tuned on the English dataset, aiming at zero-shot classification for other languages.

Cross-lingual Question Answering We attempt MLQA$\diamond$ (Lewis et al., 2020b) on 7 languages (including English). This task requires identifying the answer to a question as a span in the corresponding paragraph. The model is pre-trained on multilingual corpora and fine-tuned on the English dataset, then attempting zero-shot prediction for other languages.

4.3 Secondary Monolingual Task

Recall that, as presented in Eq.2, $H_{[\mathcal{M}]_t}$ is the contextualized representation or the final hidden state. Therefore, MLM-GC pre-training is general and can work for other MLM instances such as BERT (Devlin et al., 2019), mBART (Liu et al., 2020), SpanBERT (Joshi et al., 2020), BART (Lewis et al., 2020a), and ALBERT (Lan et al., 2020). Meanwhile, MLM-GC pre-training is substantially better than MLM pre-training beyond multilingual tasks. We provide further experiments on monolingual tasks including SQuAD v1&v2 Rajpurkar et al. (2016) in Appendix §D, building on ALBERT.

5 Result

5.1 Cross-lingual Embedding and Understanding Co-occurrence

Setup We configure an identical MLM instance to XLM with a 12-layer Transformer encoder. However, instead of 80K BPE and 15 languages in the original work, we learn 60K BPE and pre-train the model on Wikipedia dumps$\diamond$ of the 2 languages. After 300K pre-training steps, we extract the embeddings required by the test set from the embedding space of the model. For words split into 2+ sub-tokens, we average all the sub-token embeddings. See details in Appendix C.2.1. As mentioned early, this experiment is our *dev experiment*.

Table 1: Results on SemEval'17 and MUSE tasks. This is our *dev experiment* for n.

#	Model	MUSE task cos similarity	MUSE task L2 distance	SemEval'17 task Pearson correlation
1	MUSE	0.38	5.13	0.65
2	NASARI (SemEval'17 baseline)			0.60
3	XLM (reported on 15 languages)	0.55	2.64	0.69
	12-layer Transformer encoder, 60K BPE and Wikipedia dumps in $\{De, En\}$. See texts for details.			
4	XLM (reimplemented on 2 languages)	0.55	1.87	0.68
5	XLM + OURS $n = 2$ (default)	0.63	1.31	0.72
6	XLM + OURS $n = 1$	0.61	1.73	0.69
7	XLM + OURS $n = 3$	0.62	1.59	0.71
7	XLM + OURS $n = 4$	0.62	1.51	0.71

Performance We follow the instruction to compute the cosine similarity and L2 distance for the MUSE task and Pearson correlation for the SemEval'17 task, reporting the result in Table 1 for $En \leftrightarrow De$ test sets. MLM-GC pre-training outperforms the baseline model in all the metrics with different n. A large n does not consistently improve performance. We suspect that a large n may impact the capacity of the contextualized representation $H_{[\mathcal{M}]_t}$, which makes the model hard to be trained. Furthermore, $n = 2$ shows the best performance, and we may explain that in our comparison of co-occurrence counts (Figure 2), $n = 2$ has slightly smaller absolute difference avg and the ratio avg and narrower confidence (95%) intervals in large-count bins ($> log(1e7)$) contributing to over 45% co-occurrence counts on the multilingual corpora. Since we do not inject any cross-lingual supervision into the embedding space, this test can quantitatively report how MLM-GC refines the language spaces from co-occurrence counts for the isomorphic space and multilinguality.

Visualization We visualize all the embeddings from the MUSE test sets. Since the task is originally designed for word translation including nouns, verbs, and other meaningful words, analogical words should be clustered in the embedding space. As reported in Google's NMT (Johnson et al., 2017), analogical embeddings, morphological variants, and other embeddings of similar linguistic properties and meanings should be clustered in the t-SNE visualization. Then, we employ the t-SNE visualization to observe the clustering phenomenon. Figure 1 shows that embeddings after MLM pre-training are not clustered well. By contrast, we can see a clustering phenomenon after MLM-GC pre-training that indicates the model is encouraged to understand analogical embeddings.

Multilingual Embedding Analogy Besides, we consider the classic analogy test: "English: *King - Man + Woman = Queen* and German: *König-Mann+Frau = Königin*", showing results in Table 2. MLM-GC pre-training consistently improves the performance on monolingual tests (only English or German) and multilingual tests (mixing English with German). Then, we can confirm the effectiveness of our method on multilingual embedding analogy and linguistic structures.

Table 2: Word analogy: King - Man + Woman = Queen (German: König-Mann+Frau = Königin).

	X	cos (X , Queen) XLM	cos (X , Queen) XLM+OURS	cos(X , Königin) XLM	cos(X , Königin) XLM+OURS
mono:	King-Man+Woman	0.44†	0.58†	0.35	0.52
mono:	König-Mann+Frau	0.33	0.51	0.45†	0.58†
multi:	King-Man+Frau	0.34	0.53	0.33	0.50
multi:	King-Mann+Woman	0.45	0.54	0.33	0.51
multi:	King-Mann+Frau	0.42	0.56	0.35	0.52
multi:	König-Man+Woman	0.35	0.46	0.44	0.49
multi:	König-Man+Frau	0.25	0.48	0.40	0.50
multi:	König-Mann+Woman	0.38	0.50	0.43	0.53

5.2 UNMT

Setup&Training We consider three similar language pairs $\{Fr, De, Ro\} \leftrightarrow En$ from WMT$\diamond$ (Bojar et al., 2018) and a dissimilar language pair $En \leftrightarrow Ne$ (Nepali) from FLoRes$\diamond$ (Guzmán et al., 2019). Transformer, configurations, corpora, and BLEU scripts are identical to XLM and Mass. We pre-train the model around 400K iterations on only monolingual corpora of the two languages. In the training phase, we use Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.997$ and $\epsilon = 10^{-9}$, and a dynamic learning rate with $warm\_up = 8000$ ($learning\_rate \in (0, 7e^{-4}]$) is employed. We set dropout with $rate = 0.1$ and label smoothing with $gamma = 0.1$. After MLM-GC pre-training,

Table 3: Results of UNMT. $\star$ is reimplemented.

6-layer Transformer encoder-decoder, 60K BPE for each bilingual UNMT. See Appendix C.2.2.								
Language pair	$Fr \leftrightarrow En$		$De \leftrightarrow En$		$Ro \leftrightarrow En$		$Ne \leftrightarrow En$	
Test set	*newstest2014*		*newstest2016*				FLoRes$\diamond$	
Corpora	News Crawl $\diamond$ from 2007 to 2017						FLoRes$\diamond$	
default *multi-BLEU.perl*$\diamond$								
XLM	33.3	33.4	34.3	26.4	31.8	33.3	0.5	0.1
XLM + word translation tables $\star$			35.1	27.4	33.6	34.4	4.1	2.2
XLM + *n-gram* translation tables	34.9	35.4	35.6	27.7	34.1	34.9	4.8$\star$	2.4$\star$
XLM + OURS	35.2	35.8	36.0	27.8	33.7	35.2	5.2	2.8
MASS	34.9	37.5	35.2	28.3	33.1	35.2		
MASS + OURS	35.9	38.2	36.7	28.7	34.3	36.8	7.1	3.1
default *sacreBleu*$\diamond$:nrefs:1—case:mixed—eff:no—tok:13a—smooth:exp—version:2.0.0								
mBART + CC25 corpora			34.0	29.8	30.5	35.0	10.0	4.4
XLM + OURS	34.9	35.6	35.7	27.6	33.4	34.9	5.0	2.7
MASS +OURS	35.6	38.0	36.4	28.5	34.2	36.5	6.9	2.9

Table 4: Results of cross-lingual classification on XNLI. $\star$ is reimplemented.

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
baseline	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
mBERT	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
12-layer Transformer encoder, 80K BPE, and 15 XNLI languages from Wikipedia dumps downloaded by WikiExtractor$\diamond$. See Appendix C.2.3.																
XLM	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM + PMI-masking $\star$	84.1	78.4	77.8	76.6	75.1	75.5	74.9	69.7	70.8	73.0	70.7	73.4	68.1	66.1	65.3	73.3
XLM + OURS	84.8	79.7	79.1	77.1	76.2	77.0	76.2	71.5	72.2	74.2	72.3	75.2	69.4	68.2	67.5	74.7
+ Parallel Sentences from OPUS$\diamond$																
XLM + TLM	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
XLM + TLM + OURS	85.0	79.9	79.2	78.5	77.1	78.0	76.4	73.1	74.0	76.7	73.9	76.8	70.2	68.8	67.9	75.7

we follow XLM and MASS to train the model for translation from pre-trained weights. After around 400K iterations, we report results. See details in Appendix C.2.2.

Performance In Table 3, we report *multi-BLEU.perl*$\diamond$ to compare with XLM and MASS and *sacreBleu*$\diamond$ to compare with mBART (Liu et al., 2020) so that the evaluation is based on the same BLEU script. MLG-GC pre-training consistently improves the performance of baseline models on all the similar language pairs by $3\% \sim 8\%$ and on the dissimilar pair by $2.0 \sim 5.0$ BLEU. The performance on the dissimilar pair is somewhat comparable to SOTA: mBART and is better than mBART on similar language pairs, but mBART uses CC25 (Wenzek et al., 2020) for pre-training and obtains benefits from more languages (25 languages) and samples. Surprisingly, our method even slightly outperforms two dictionary-based works (Ren et al., 2019; Chaudhary et al., 2020) which require static translation tables from pre-trained n-gram models, golden dictionaries, or bilingual lexicon induction (available on OPUS$\diamond$). Intuitively, as reported in Artetxe et al. (2020b); Kementchedjhieva et al. (2019); Czarnowska et al. (2019); Vania & Lopez (2017), such translation tables are reported to misrepresent morphological variations and are not contextualized properly, which limit the improvements for sentence translation. By contrast, the global co-occurrence information is general for analogical words and morphological variations and allows for cross-lingual representations, which eventually helps the model understand translation knowledge. Meanwhile, we observe substantial gains on MASS + OURS (and ALBERT Lan et al. (2020) in Appendix D), where MASS is based on span masking. As discussed in §Related Work and Introduction, span-based masking (also including Whole Word Masking and PMI-masking) implicitly leverages co-occurrence information for improving context understanding. In addition to the empirical study in Figure 2, the gain further confirms that global co-occurrence information significantly injects some signals for cross-lingual transfer beyond improving context understanding.

5.3 Cross-lingual Classification

Setup&Fine-tuning The model configuration, preprocessing, and corpora are identical to XLM[1]. For the classification objective, we deploy a linear classification layer on top of the encoder. After pre-training, we deploy the randomly initialized linear classifier and fine-tune the encoder and the linear classifier on the En NLI dataset with mini-batch size 16. We use Adam optimizer with $lr = 5 \times 10^{-4}$ and linear decay of lr. After fine-tuning, we make zero-shot prediction for the other 14 languages. See details in Appendix C.2.3.

[1]In the literature, this setup also refers to XLM-15.

Table 5: Results of cross-lingual question answering on MLQA. We report the F1 and EM (exact match) scores for zero-shot prediction. $\star$ is reimplemented.

Model	en	es	de	ar	hi	vi	zh	Avg
mBERT-102	77.7 / 65.2	64.3 / 46.6	57.9 / 44.3	45.7 / 29.8	43.8 / 29.7	57.1 / 38.6	57.5 / 37.3	57.7 / 41.6
12-layer Transformer encoder, 80K BPE, and and 15 XNLI languages from Wikipedia dumps downloaded by WikiExtractor$\diamond$. See Appendix C.2.4.								
XLM	74.9 / 62.4	68.0 / 49.8	62.2 / 47.6	54.8 / 36.3	48.8 / 27.3	61.4 / 41.8	61.1 / 39.6	61.6 / 43.5
XLM + PMI-masking $\star$	76.0 / 63.9	69.2 / 50.2	64.1 / 48.0	55.8 / 38.0	49.8 / 28.5	62.9 / 42.2	63.3 / 40.5	63.1 / 44.4
XLM + OURS	77.5 / 65.6	71.4 / 50.9	65.3 / 48.6	57.1 / 39.6	51.1 / 29.9	64.1 / 43.0	64.5 / 41.7	64.4 / 45.7

Performance We report the result in Table 4. Our method consistently improves baseline models by 4.5% (Avg). As discussed in previous models (Conneau et al., 2020b; Karthikeyan et al., 2020; Wu & Dredze, 2019; Pires et al., 2019; Dufter & Schütze, 2020), multilinguality is essential for this task. Then, we confirm that MLM-GC pre-training including global co-occurrence information improves the cross-linguality and multilinguality learned in pre-training. Furthermore, our method outperforms XLM + PMI-masking (span-based). Similar to the comparison between MASS and MASS + OURS in UNMT, MLM-GC pre-training uses co-occurrence information for better context understanding and cross-lingual transfer, whereas XLM+PMI-masking leverages co-occurrence information for context understanding but performs worse for cross-lingual transfer because of the lack of a mechanism to understand cross-lingual supervision. We also include XLM + TLM (Lample & Conneau, 2019) for comparison. Recall that in UNMT, our method outperforms dictionary-based methods. However, in this experiment, XLM + TLM using parallel sentences in pre-training outperforms MLM-GC, which indicates the knowledge gap between co-occurrence information and parallel sentences for cross-lingual supervision. Besides, when applying MLM-GC pre-training for XLM + TLM, we still observe improvement. We attribute the additional gains to the morphologies or the embedding space that is further refined by co-occurrence information to represent embeddings of similar meanings, e.g., embedding analogy. Intuitively, it indicates that the co-occurrence information gives extra cross-lingual supervision beyond a limited amount of parallel sentences.

5.4 Cross-lingual Question Answering

Setup&Fine-tuning The setup is similar to Cross-lingual Classification. We follow the instruction of SQuAD from BERT, fine-tuning the model with a span extraction loss on the English dataset. We use Adam optimizer with $lr = 5 \times 10^{-5}$ and linear decay of lr. As suggested, we fine-tune the model on the SQuAD v1.1 (Rajpurkar et al., 2016) dataset and then make zero-shot prediction for the 7 languages of MLQA. See details in Appendix C.2.4.

Performance We show the results in Table 5. MLM-GC pre-training substantially improves the Avg performance in both F1 and EM metrics by 4.5 % and 4.8 % respectively. Meanwhile, MLM-GC pre-training yields more improvements for low-resource languages. We attribute all the improvements to the global co-occurrence objective the model learns in MLM-GC pre-training. Intuitively, spans of answers are most likely to consist of nouns and terms and can be easily represented, clustered, and aligned by considering global unigram frequencies and co-occurrence frequencies because these spans or a group of words from multiple languages are analogical. Therefore, considering embedding analogies is potentially beneficial for zero-shot cross-lingual transfer.

6 Conclusion

In this work, we leverage the global co-occurrence information from multilingual corpora in MLM pre-training. The result is MLM-GC pre-training with a combined objective of MLM and global co-occurrence modeling. Our experiments show that MLM-GC pre-training can substantially improve the performance of naive MLM pre-training for 4 multilingual tasks, and additional experiments show it can work for monolingual tasks. the model is encouraged to distinguish relevant information from irrelevant information and to discriminate between the two relevant information across languages from co-occurrence counts (normalized) and refine contextualized representations accordingly. Eventually, the model learns to understand analogical embeddings, morphological variants, and structural similarities across languages from co-occurrence counts. We believe it is an interesting avenue for leveraging the primary source of information on multilingual corpora.

REFERENCES

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, 2016.

Xi Ai and Bin Fang. Empirical regularization for synthetic sentence pairs in unsupervised neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12471–12479, 2021.

Hanan Aldarmaki and Mona Diab. Context-aware cross-lingual mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3906–3911, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1391. URL https://aclanthology.org/N19-1391.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://www.aclweb.org/anthology/2020.acl-main.421.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. A Call for More Rigor in Unsupervised Cross-lingual Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7375–7388. Association for Computational Linguistics, 2020b. doi: 10.18653/v1/2020.acl-main.658. URL https://github.com/google-research/.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pp. 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics.

Ond rej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation*, pp. 272–307, Belgium, Brussels, October 2018. Association for Computational Linguistics.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 15–26. Association for Computational Linguistics, 2018. doi: 10.18653/v1/s17-2002. URL http://alt.qcri.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual Alignment of Contextual Word Representations. In *8th International Conference on Learning Representations, ICLR 2020 - Conference Track Proceedings*, 2020. URL https://arxiv.org/pdf/2002.03518.pdfhttp://arxiv.org/abs/2002.03518.

Pi Chuan Chang, Michel Galley, and Christopher D Manning. Optimizing Chinese word segmentation for machine translation performance. In *3rd Workshop on Statistical Machine Translation, WMT 2008 at the Annual Meeting of the Association for Computational Linguistics, ACL 2008*, pp. 224–232, 2008. ISBN 9781932432091. doi: 10.3115/1626394.1626430. URL https://aclanthology.org/W08-0336.pdf.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries. *arXiv preprint arXiv:2010.12566*, 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://www.aclweb.org/anthology/2020.acl-main.747.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Veselin Stoyanov, Adina Williams, and Samuel R. Bowman. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485. Association for Computational Linguistics, 2020b. ISBN 9781948087841. doi: 10.18653/v1/d18-1269.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020c. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL https://www.aclweb.org/anthology/2020.acl-main.536.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 974–983, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1090. URL https://aclanthology.org/D19-1090.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Philipp Dufter and Hinrich Schütze. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4423–4437, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.358. URL https://aclanthology.org/2020.emnlp-main.358.

Francisco Guzmán, Peng Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The Flores evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 6098–6111, feb 2019. ISBN 9781950737901. doi: 10.18653/v1/d19-1632. URL http://arxiv.org/abs/1902.01382.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. Extension of Zipf's law to words and phrases. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1117.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl_a_00065.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *8th International Conference on Learning Representations, ICLR 2020 - Conference Track Proceedings*, 2020.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3336–3341, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1328. URL https://aclanthology.org/D19-1328.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

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. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-2045.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In *Advances in neural information processing systems*, 2019.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018a.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020 - Conference Track Proceedings*, 2020. URL https://github.com/google-research/ALBERT.http://arxiv.org/abs/1909.11942.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. PMI-Masking: Principled masking of correlated spans. In *9th International Conference on Learning Representations, ICLR 20201- Conference Track Proceedings*, 2021. URL https://github.com/huggingface/tokenizershttp://arxiv.org/abs/2010.01825.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7315–7330, 2020b. doi: 10.18653/v1/2020.acl-main.653. URL https://github.com/facebookresearch/.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://www.aclweb.org/anthology/2020.tacl-1.47.

Mauro Mezzini. Empirical study on label smoothing in neural networks. In *WSCG 2018 - Short papers proceedings*, 2018. doi: 10.24132/csrn.2018.2802.25.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. URL https://aclanthology.org/D14-1162.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? pp. 4996–5001, July 2019. doi: 10.18653/v1/P19-1493. URL https://www.aclweb.org/anthology/P19-1493.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. ISBN 9781945626258. doi: 10.18653/v1/d16-1264.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pp. 784–789, 2018. ISBN 9781948087346. doi: 10.18653/v1/p18-2124. URL https://arxiv.org/pdf/1806.03822.pdf.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. Explicit cross-lingual pre-training for unsupervised machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 770–779, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1071. URL https://www.aclweb.org/anthology/D19-1071.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*, 2021.

Anders Søgaard. Some languages seem easier to parse because their treebanks leak. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 2765–2770, 2020. ISBN 9781952148606. doi: 10.18653/v1/2020.emnlp-main.220. URL https://wals.info.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5926–5936. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/song19d.html.

Clara Vania and Adam Lopez. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pp. 2016–2027, 2017. ISBN 9781945626753. doi: 10.18653/v1/P17-1184. URL https://doi.org/10.18653/v1/P17-1184.

Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Pascal Vincent. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework. In *8th International Conference on Learning Representations, ICLR 2020 - Conference Track Proceedings*, 2020. URL https://github.com/google-research/bert/blob/master/multilingual.mdhttp://arxiv.org/abs/1910.04708.

Guillaume Wenzek, Marie Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4003–4012, 2020. ISBN 9791095546344. URL https://commoncrawl.org/about/.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 833–844, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL https://www.aclweb.org/anthology/D19-1077.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020 - Conference Track Proceedings*, 2020. URL https://github.com/tensorflow/addons/blob/master/tensorflow_addons/.

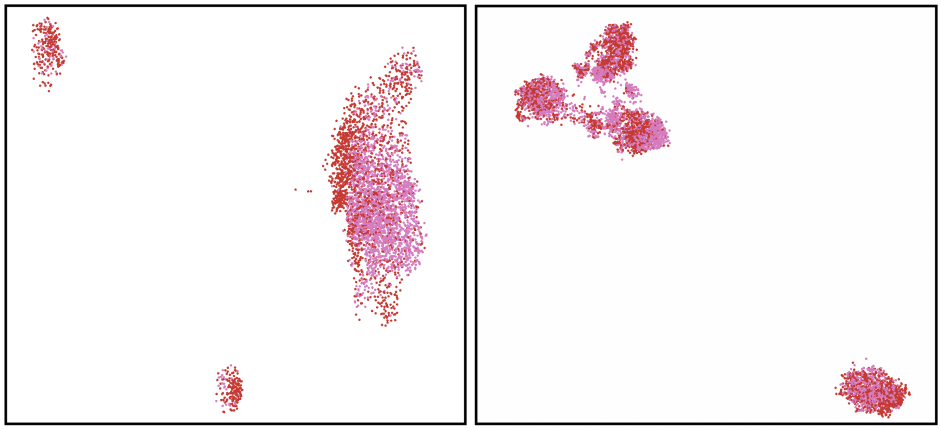
Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 19–27, 2015. URL http://www.cs.utoronto.ca/.

George Kingsley Zipf. Human behavior and the principle of least effort: an introd. to human ecology. 1949.

George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013.

A Appendix

- For the visualization and inspiration, please refer to §B.
- We provide details of our experiment in §C.
- We give supportive results in §D.
- We present early experiments in §E.
- We list all the sources we use for this work in §G.



(a) MLM pre-training (b) MLM-GC pre-training

Figure 3: Preliminary experiment of multilingual MLM pre-training on $\{De, En\}$ corpora and t-SNE visualization for the MUSE task.

B Inspiration from Visualization

t-SNE Visualization We visualize the embedding space of the pre-trained model to observe the alignment of embeddings, considering the test set $De \leftrightarrow En$ from the cross-lingual embedding MUSE (Lample et al., 2018a) task. We present the t-SNE visualization[2] in Figure 3. In previous multilingual models trained on parallel sentences like Johnson et al. (2017), t-SNE visualizations are used to cluster semantically similar representations. In our case, semantically and linguistically similar embeddings should be in the same cluster. In our experiments, embedding pairs are not clustered well after MLM pre-training, which means embedding analogies and morphologies are not represented well. By contrast, after MLM-GC pre-training, embeddings are clustered better, which is beneficial for multilinguality.

Word Analogy and Morphological Variant on Multilingual Corpora As discussed in the main paper, analogical words and morphological variants across languages potentially have similar frequencies on multilingual corpora. We assume that they can further provide multilingual and cross-lingual knowledge for learning. In previous works like GloVe (Pennington et al., 2014), using global co-occurrence information is a proven idea to leverage this information, which inspires us to introduce global co-occurrence information to the contextualized representation outputted by the final layer.

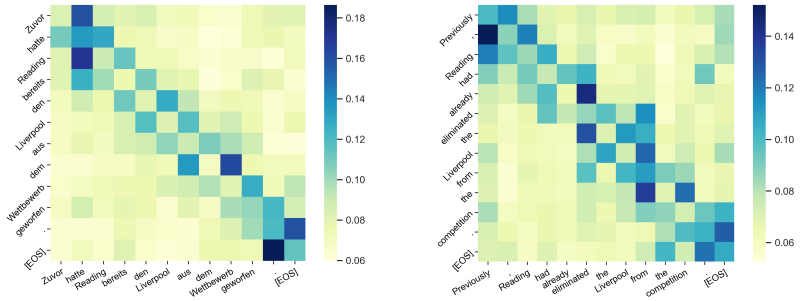
Attention Visualization We visualize the attention weights for each position. As illustrated in Figure 4 for a simple case study, given the parallel sentences in the figure, which have comparable linguistic structures, the model shows similar behavior in information processing. Intuitively, it tells us that the *co-occurrence* information is significant for the model and represents structural similarities, and the model learns to understand *n-co-occurrence* information for restoring inputs in MLM pre-training on multilingual corpora across languages. However, Levine et al. (2021) report negative results of capturing occurrence information due to sub-optimization of token collocations. Regularities for this information are potentially beneficial for downstream tasks (Cao et al., 2020; Artetxe et al., 2020a).

C Experiment

C.1 MLM Instance

We adapt our method to two MLM instances: XLM (Lample & Conneau, 2019) and MASS (Song et al., 2019). We follow the instructions of BERT (Devlin et al., 2019) and these two MLM instances

[2]See introduction from https://distill.pub/2016/misread-tsne/.



(a) De sentence

(b) En sentence

Figure 4: Preliminary experiment of multilingual MLM pre-training on $\{De, En\}$ corpora. Intuitively, the *co-occurrence* information surrounding each position is significant for the model. Multilinguality emerges from the structural similarity, where in this case, the structural similarity means *co-occurrence* information.

that each selected token is replaced with the probabilities $(p[unchanged], p[random], p[mask]) = (0.1, 0.1, 0.8)$.

XLM XLM is similar to BERT (Devlin et al., 2019) but uses text streams of an arbitrary number of sentences. Following the instruction, we randomly select 15% of the tokens from the input sentence for replacing.

MASS MASS is different from XLM and BERT but similar to SpanBERT (Joshi et al., 2020), using spans to replace consecutive tokens. Given an input sentence with length N, we randomly select consecutive tokens with length $N/2$ for replacing.

C.2 Multilingual Task

C.2.1 Cross-lingual Embedding

We are interested in the isomorphism of languages' embedding spaces. To investigate, we attempt SemEval'17⋄ (Camacho-Collados et al., 2018) and MUSE⋄ tasks (Lample et al., 2018a) that measure similarities between two paired words. This test can generally evaluate the degree of the isomorphism of languages' embedding spaces. Meanwhile, as discussed in Lample & Conneau (2019); Wang et al. (2020) and our preliminary experiment, the performance of the isomorphism is potentially proportional to the performance of cross-lingual transfer learning tasks. Therefore, we treat this experiment as our *dev experiment* to search n.

Setup We configure a 12-layer Transformer encoder and use Moses tokenizer⋄ with default rules for tokenization, identical to XLM (Lample & Conneau, 2019). For fast *dev experiment*, we employ fastBPE⋄ to learn 60K BPE (Sennrich et al., 2016b) from concatenated corpora with a sampling criterion from (Lample & Conneau, 2019) and pre-train the model on 2 languages instead of 80BPE and 15 languages in the reported work.

Training In the pre-training phase, we pre-train the model on Wikipedia dumps⋄ of the two languages for 300K steps. After pre-training, we extract the words' embeddings required by the test set from the embedding space of the model. For words split into 2+ sub-tokens, we average all the extracted embeddings of sub-tokens. We then evaluate paired embeddings in cosine similarity, L2 distance, and Pearson correlation.

C.2.2 UNMT

UNMT (unsupervised neural machine translation) (Lample & Conneau, 2019; Lample et al., 2018b; Song et al., 2019; Liu et al., 2020) tackles bilingual translation (Bahdanau et al., 2015; Vaswani et al., 2017) on non-parallel bilingual corpora without having access to any parallel sentence. In

other words, there is no supervision for translation. The model requires pre-training to obtain some initial multilingual knowledge for decent performance.

Setup We configure an identical Transformer model to XLM (Lample & Conneau, 2019) and MASS (Song et al., 2019), which has 6 layers in both the encoder and decoder using default configurations. We consider multiple families of languages. Specifically, we consider similar language pairs $\{Fr, De, Ro\} \leftrightarrow En$, using the same dataset as previous works (Lample & Conneau, 2019). The dataset consists of monolingual corpora $\{Fr, De, En\}$ from WMT 2018$\diamond$ (Bojar et al., 2018) including all available $NewsCrawl$ datasets from 2007 through 2017 and monolingual corpora Ro from WMT 2016$\diamond$ (Bojar et al., 2016) including $NewsCrawl$ 2016. We report the performance for $Fr \leftrightarrow En$ on *newstest2014* and $\{De, Ro\} \leftrightarrow En$ on *newstest2016*. Meanwhile, we share the FLoRes$\diamond$ (Guzmán et al., 2019) task to evaluate on a dissimilar language pair $Ne \leftrightarrow English$ (Nepali). For tokenization, we use the Moses tokenizer$\diamond$ developed by Koehn et al. (2007) with default rules except for Ne that is tokenized by Indic-NLP Library$\diamond$. We employ fastBPE$\diamond$ to learn 60K BPE (Sennrich et al., 2016b) from concatenated corpora of paired languages, using the same sampling criteria in Lample & Conneau (2019). We use learnable language embeddings and position embeddings.

Training In MLM-GC pre-training, the model is pre-trained around 400K iterations on only monolingual corpora of different languages. In the training phase, we use Adam optimizer (Kingma & Ba, 2015) with parameters $\beta_1 = 0.9$,$\beta_2 = 0.997$ and $\epsilon = 10^{-9}$, and a dynamic learning rate with $warm\_up = 8000$ Vaswani et al. (2017) ($learning\_rate \in (0, 7e^{-4}]$) is employed. We set dropout regularization with a drop rate $rate = 0.1$ and label smoothing with $gamma = 0.1$ (Mezzini, 2018). On-the-fly back-translation (Sennrich et al., 2016a) (the inference mode of the model) performs to generate synthetic parallel sentences that can be used for training of translation as NMT (neural machine translation) is trained on genuine parallel sentences in a supervised manner. Meanwhile, UNMT learns an objective of denoising language modeling (Vincent, 2010) to maintain language knowledge in the training phase except for MASS. After around 400K iterations, we report BLEU computed by *multi-BLEU.perl*$\diamond$ and *scareBLEU*$\diamond$ with default rules, according to baseline models. In conclusion, in pre-training, we only have the objective of MLM-GC, and in training, we have the two objectives: *1)* denoising language modeling for XLM or MASS itself and *2)* translation (i.e., NMT), where the translation objective is finished by using synthetic pairs sentences from on-the-fly back-translation.

C.2.3 Cross-lingual Classification

We experiment on XNLI$\diamond$ (Conneau et al., 2020b) that is a general cross-lingual classification task on 15 languages (including English) under the cross-lingual transfer setting. The model takes in two input sentences and is required to classify into one of the three labels: entailment, contradiction, and neutral. The model is fine-tuned on the English dataset and then attempts zero-shot classification for other languages.

Setup Following the previous work[3] (Lample & Conneau, 2019), we use raw sentences including 15 XNLI languages from Wikipedia dumps downloaded by WikiExtractor$\diamond$. We concatenate all the downloaded corpora and then shuffle the concatenated corpus. The model configuration and preprocessing are identical to XLM (Lample & Conneau, 2019) that we use a 12-layer transformer encoder and 80K BPE. For the classification objective, we deploy a linear classification layer on top of the encoder. To tokenize $\{zh, th\}$, we use Stanford Word Segmenter$\diamond$ and PyThaiNLP$\diamond$ respectively. For the others, we use the Moses tokenizer$\diamond$ with default rules. Similar to the Cross-lingual Embedding experiment, we use fastBPE$\diamond$ and the sampling strategy to learn BPE.

Fine-tuning After pre-training on the corpora, we deploy a randomly initialized linear classifier and fine-tune the encoder and the linear classifier on the En NLI dataset with mini-batch size 16. We use Adam optimizer (Kingma & Ba, 2015) with $lr = 5 \times 10^{-4}$ and linear decay of lr. After fine-tuning, we make zero-shot prediction for the other 14 languages. We use categorical cross-entropy with three labels: entailment, contradiction, and neutral.

[3] In the literature, this setup also refers to XLM-15.

Table 6: MLM-GC pre-training for ALBERT. ⋆ *denotes the baseline models that are reimplemented.*

Model	SQuAD1.1 (F1)	SQuAD2.0 (F1)
12-base-ALBERT Lan et al. (2020)	89.3	80.0
⋆12-base-ALBERT	89.4	80.0
12-base-ALBERT + OURS: 12-base-LT	89.7	80.6

C.2.4 Cross-lingual Question Answering

We consider the MLQA⋄ (Lewis et al., 2020b) dataset for a cross-lingual question answering task. Given a question and a passage containing the answers, the aim is to predict the answer text span in the passage. This task requires identifying the answer to a question as a span in the corresponding paragraph. The evaluation data for English and 6 other languages are obtained by automatically mining target language sentences that are parallel to sentences in English from Wikipedia, crowd-sourcing annotations in English, and translating the question and aligning the answer spans in the target languages. Similar to the cross-lingual classification task, the model is fine-tuned on the English dataset, and then attempts zero-shot prediction for other languages.

Setup The setup is similar to Cross-lingual Classification.

Fine-tuning We follow the instruction of SQuAD from BERT (Devlin et al., 2019), fine-tuning the model with a span extraction loss on the English dataset. We use Adam optimizer (Kingma & Ba, 2015) with $lr = 5 \times 10^{-5}$ and linear decay of lr. Meanwhile, as suggested, we fine-tune the model on the SQuAD v1.1 (Rajpurkar et al., 2016) dataset and then make zero-shot prediction for the 7 languages of MLQA. Given a sequence T, we only have a start vector $S \in R^{hidden}$ and an end vector $E \in R^{hidden}$ during fine-tuning. The probability of word i being the start of the answer span is computed as a dot product T_i and S d by a $softmax$ over all of the words in the sequence $p_i = \frac{e^{ST_i}}{\sum_{k \in T} e^{ET_k}}$. Similarly, we can compute the end of the span. The score of a candidate span from position i to position j is defined as $ST_i + ET_j$ and the maximum scoring span where $j \geq i$ is used as a prediction.

D Additional and Supportive Result

D.1 Pre-training for Monolingual Task

Although we derive our method from the observation of multilingual models, MLM-GC pre-training is substantially better than MLM pre-training. We provide further experiments on monolingual tasks including SQuAD v1&v2 (Rajpurkar et al., 2016).

setup For this monolingual task, our configuration is identical to 12-base-ALBERT (Lan et al., 2020). Specifically, We set the model dimension, word embedding dimension, and the maximum number of layers to 768, 128, and 12. As recommended, we generate a masked span for the MLM targets using the random strategy from Joshi et al. (2020), and we use LAMB optimizer⋄ with a learning rate of 0.00176 (You et al., 2020) instead of Adam optimizer. Following the instructions, we pre-train models on BooksCorpus⋄ Zhu et al. (2015) and English Wikipedia⋄ (Devlin et al., 2019) for 140k steps.

Fine-tuning Similar to the cross-lingual question answering task, we fine-tune the pre-trained model on SQuAD(v1.1 and v2.0)⋄ (Rajpurkar et al., 2016; 2018).

Result Table 6 shows that MLM-GC pre-training is substantially better than MLM pre-training, when pre-training 12-base-ALBERT for monolingual tasks. These observations confirm the effectiveness of MLM-GC pre-training on monolingual tasks.

Table 7: Impact of Tokenization Method. $\star$ denotes reimplemented models.

Model	$De \leftrightarrow En$	
baseline (BPE-based) $\star$	33.81	26.32
+ Ours	34.98	27.20
baseline (Word-level) $\star$	33.01	25.79
+ Ours	34.15	26.61

D.2 Impact of Tokenization Method

We are interested in how the tokenization method affects the performance because it potentially affects the token-token co-occurrence counts. For evaluation, we use all the configurations in UNMT and additionally configure a word-level vocabulary for the model. The word-level vocabulary has the same number of tokens as the BPE vocabulary. Table 7 shows that our method can work with different tokenization methods. Our method can generally improve the performance, regardless of the difference between the two baseline models in the same configuration.

E Early Experiment

Table 8: Early Emperiments on SemEval'17 and MUSE tasks.

#	Model	MUSE task		SemEval'17 task
		cos similarity	L2 distance	Pearson correlation
1	MUSE Lample et al. (2018a)	0.38	5.13	0.65
2	NASARI (SemEval'17 baseline)Camacho-Collados et al. (2018)			0.60
3	XLMLample & Conneau (2019)	0.55	2.64	0.69
4	XLM + OURS $\lambda = 1$ (default)	0.63	1.31	0.72
5	XLM + OURS $\lambda = 0.1$,	0.60	2.12	0.65
6	XLM + OURS $\lambda = 0.5$	0.63	1.43	0.71
7	XLM + OURS $\lambda = 2$	0.61	1.86	0.66
8	XLM + OURS $\lambda = 2$ and no $warm\_up$ of learning rate		fail	
9	XLM + OURS $\lambda = 2$ and no $\sqrt{d}$		fail	
10	XLM + OURS $\lambda = 2$ and no $\frac{1}{2n}$		fail	

Setup The setup and training are identical to the experiment of the Cross-lingual Embedding task, except that we add a hyper-parameter $\lambda \in [0.1, 0.5, 1, 2]$ to $\lambda \mathcal{L}_{GC}$.

Performance The result is presented in Table 8. We find λ is not that significant, but $\lambda = 1$ is a general choice for experiments. We add scaling $\sqrt{d}$ and weight $\frac{1}{2n}$ to make training stable (Eq.3), where $\sqrt{d}$ is inspired by scaled dot-product attention to prevent the dot products get large. They serve as hyperparameters for $\mathcal{L}_{GC}$. Meanwhile, we find $warm\_up$ steps (Vaswani et al., 2017) of learning rate is significant. Without $warm\_up$, $\sqrt{d}$, or $\frac{1}{2n}$, the model may collapse to $\mathcal{L}_{GC}$ because $\mathcal{L}_{GC}$ converges too fast and is unstable. In this situation, the model ignores the objective of MLM. Then, the model can only learn co-occurrence information and does not learn the language knowledge.

F Reimplementation

Language pair	$De \leftrightarrow En$	
multi-BLEU.perl$\diamond$ with default rules		
XLMLample et al. (2018b) *reported*	34.30	26.40
XLMLample et al. (2018b) $\star$	33.90	26.30
XLM+OURS	35.95	27.42
multi-BLEU.perl$\diamond$ with default rules		
MASSSong et al. (2019) *reported*	35.20	28.30
MASSSong et al. (2019)$\star$	35.0	28.0
MASS+OURS	36.65	28.62

Table 9: Performance of UNMT. Baseline models ($\star$) are reimplemented with our configurations.

We compare our reimplementation with reported results in Table 9.

Table 10: Links of source.

Item	Links
WMT 2016	http://www.statmt.org/wmt16/translation-task.html
WMT 2018	http://www.statmt.org/wmt18/translation-task.html
FLoRes	https://github.com/facebookresearch/flores
Indic-NLP Library	https://github.com/anoopkunchukuttan/indic_nlp_library
XLM	https://github.com/facebookresearch/XLM
multi-BLEU.perl	https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-BLEU.perl
Moses tokenizer	https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
Kytea	http://www.phontron.com/kytea/
XTREME	https://github.com/google-research/xtreme
fastBPE	https://github.com/glample/fastBPE
MUSE	https://github.com/facebookresearch/MUSE
Cambridge Dictionary	https://dictionary.cambridge.org/
SemEval'17	https://alt.qcri.org/semeval2017/task2/
WikiExtractor	https://github.com/attardi/wikiextractor
PyThaiNLP	https://github.com/PyThaiNLP/pythainlp
Stanford Word Segmenter Chang et al. (2008)	https://nlp.stanford.edu/software/segmenter.html
Tensor2Tensor	https://github.com/tensorflow
HuggingFace	https://huggingface.co
ORPUS, Wikipedia v1.0	https://opus.nlpl.eu

G SOURCE

We list all the links of dataset, tools, and other sources in Table 10. Note that for multilingual tasks, datasets can be downloaded from the XTREME link except for UNMT and cross-embeddings.