Abstract

The success of pretrained cross-lingual language models relies on two essential abilities, i.e., generalization ability for learning downstream tasks in a source language, and cross-lingual transferability for transferring the task knowledge to other languages. However, current methods jointly learn the two abilities in a single-phase cross-lingual pretraining process, resulting in a trade-off between generalization and cross-lingual transfer. In this paper, we propose cross-lingual language model meta-pretraining, which learns the two abilities in different training phases. Our method introduces an additional meta-pretraining phase before cross-lingual pretraining, where the model learns generalization ability on a large-scale monolingual corpus. Then, the model focuses on learning cross-lingual transfer on a multilingual corpus. Experimental results show that our method improves both generalization and cross-lingual transfer, and produces better-aligned representations across different languages.

1 Introduction

Pretraining-finetuning language models (LMs) has become an emerging paradigm in natural language processing (NLP) (Bommasani et al., 2021). Nonetheless, most resources like training data are English-centric, making it hard to apply this paradigm to other languages. Recently, pretrained cross-lingual LMs have shown to be effective for cross-lingual transfer, and become a promising way to build multilingual NLP applications (Conneau and Lample, 2019; Conneau et al., 2020; Chi et al., 2021b).

Cross-lingual LMs are typically pretrained with the unsupervised language modeling tasks on an unlabeled multilingual text corpus (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2021). The pretrained cross-lingual LMs achieve zero-shot cross-lingual transfer, which enables the models to be finetuned on the downstream task training data of one source language but directly applied to other target languages (Wu and Dredze, 2019; K et al., 2020). Pretrained cross-lingual LMs can be further improved by utilizing cross-lingual supervisions from parallel corpora (Conneau and Lample, 2019; Chi et al., 2021b). Moreover, token-level alignments are also proven to be beneficial for cross-lingual transfer (Cao et al., 2020; Zhao et al., 2021; Chi et al., 2021c).

The success of pretrained cross-lingual LMs relies on two essential abilities. The first ability is the generalization ability, which helps the model to learn downstream tasks in a source language. The second ability is cross-lingual transferability that transfers the downstream task knowledge to target languages. However, current methods jointly learn the two abilities in a single-phase cross-lingual pretraining process, resulting in a trade-off between generalization and cross-lingual transfer. The trade-off is typically achieved by a rebalanced language distribution (Conneau and Lample, 2019) that controls the frequency of a language to be used in cross-lingual pretraining. As shown in Figure 1, sampling the source language more frequently encourages generalization

![Figure 1: Cross-lingual pretraining methods typically employ a single-phase cross-lingual pretraining process, resulting in a trade-off between generalization and cross-lingual transfer.](image-url)
to downstream tasks, which provides better results on the source language. Nonetheless, it is accompanied by the weakening of cross-lingual transferability, resulting in the low performance on target languages. On the contrary, increasing the probabilities of the target languages improves cross-lingual transfer but harms generalization.

In this paper, we propose cross-lingual language model meta-pretraining (META-XLM), a novel paradigm for learning cross-lingual LMs that learns generalization ability and cross-lingual transferability in different training phases. Our method introduces an additional meta-pretraining phase before cross-lingual pretraining, which can be understood as the pretraining of pretraining. Specifically, META-XLM first learns generalization ability on a large-scale monolingual corpus in the meta-pretraining phase, and then focuses on learning cross-lingual transfer on a multilingual corpus in the cross-lingual pretraining phase. Finally, the model is finetuned on downstream tasks for various application scenarios of pretrained cross-lingual LMs.

We conduct extensive experiments on ten downstream tasks under three different application scenarios, including cross-lingual understanding tasks for cross-lingual transfer, multilingual classification for supervised finetuning, and cross-lingual alignment tasks for feature-based applications. Experimental results show that our method provides substantial improvements over the baseline in all of the three application scenarios. Moreover, comprehensive analysis shows that our method successfully breaks the trade-off of the single-phase pretraining, which improves both generalization and cross-lingual transfer, and produces better-aligned representations at both sentence level and word level.

Our contributions can be summarized as follows:

- We propose cross-lingual language model meta-pretraining, a novel paradigm for learning cross-lingual language models.
- We conduct extensive experiments on ten downstream tasks, and demonstrate the effectiveness of our method in three application scenarios.
- We show that our method improves both generalization and cross-lingual transfer, and produces better-aligned representations across languages, by only using additional monolingual data.

2 Related Work

Pretrained cross-lingual language models

Cross-lingual language model pretraining is first proposed by Conneau and Lample (2019), which aims to build universal cross-lingual encoders. They show that learning masked language modeling (MLM; Devlin et al. 2019) on multilingual text improves the results on cross-lingual classification. Similarly, MBERT (Devlin et al., 2019) learns a multilingual version of BERT, which is also proven to be effective for cross-lingual transfer (Wu and Dredze, 2019; Pires et al., 2019; K et al., 2020; Zheng et al., 2021). Several studies demonstrate that pretraining cross-lingual language models at scale leads to significant performance gains (Conneau et al., 2020; Xue et al., 2021). Besides, Conneau and Lample (2019) introduce parallel corpora into cross-lingual pretraining, and show that the cross-lingual transferability can be improved by using parallel corpora during pretraining. Recent studies also present various objectives that leverage parallel corpora to improve cross-lingual LM pretraining (Huang et al., 2019; Chi et al., 2021b; Luo et al., 2021; Wei et al., 2021; Kale et al., 2021; Chi et al., 2021a,d). Moreover, the word alignments implied in parallel corpora are also proven to be beneficial for cross-lingual transfer (Yang et al., 2020; Cao et al., 2020; Zhao et al., 2021; Hu et al., 2021; Chi et al., 2021c). Additionally, several pretrained cross-lingual models are designed for producing cross-lingual sentence embeddings (Artetxe and Schwenk, 2019; Feng et al., 2020), cross-lingual natural language generation (Liu et al., 2020; Luo et al., 2021; Lin et al., 2020; Ren et al., 2021), or the cross-lingual transfer of natural language generation (Chi et al., 2020; Maurya et al., 2021; Ma et al., 2021).

Cross-lingual transferability of monolingual models

Artetxe et al. (2020) demonstrate the Transformer body of a pretrained monolingual LM can also serve as a LM for another language. Similarly, Reimers and Gurevych (2020) propose a transferring framework that extends a monolingual model to new languages using parallel data. Li et al. (2021) propose a knowledge distillation method, which leverages parallel corpora to extend an existing sentence embedding model to new
languages. de Souza et al. (2021) show that a pretrained monolingual LM performs well when directly finetuning it on a new language. However, these studies focus on extending an existing monolingual model to new languages, while our goal is to pretrain a cross-lingual LM that performs cross-lingual transfer and produces universal representations. We provide related analyses in Section 5.3 and 5.4.

3 Method

Figure 2 illustrates the overview of our method, which contains three phases: meta-pretraining on a monolingual corpus, cross-lingual pretraining on a multilingual corpus, and model finetuning on downstream tasks in different application scenarios.

3.1 Meta-Pretraining

Meta-pretraining can be understood as the pretraining of pretraining. A pretrained model typically provides good initialization for downstream task models. Similarly, a meta-pretrained model provides initialization for the next model pretraining phase. In the meta-pretraining phase, we pretrain a monolingual language model on a large-scale unlabeled corpus. Following Liu et al. (2019), we use a Transformer (Vaswani et al., 2017) encoder as the backbone network, and pretrain the model with the masked language modeling (MLM; Devlin et al. 2019) task. The MLM task is to predict the masked words of an input text sequence. Each text sequence contains several sentences that consist of at most 512 tokens. Note that a pretrained monolingual language model such as RoBERTa (Liu et al., 2019) can also be directly used as a meta-pretrained model.

3.2 Cross-Lingual Pretraining

With the meta-pretrained model, the cross-lingual pretraining phase focuses on learning cross-lingual transfer. We pretrain a cross-lingual language model on a multilingual unlabeled text corpus with the MLM task. For each batch, we employ the rebalanced language sampling distribution (Conneau and Lample, 2019) to increase the probabilities of low-resource languages. Formally, considering a \( n \)-language multilingual text corpus with the size of \( N_i \) for the \( i \)-th language, the rebalanced probability of the \( i \)-th language is
\[
p_i = \frac{N_i^\alpha}{\sum_{j=1}^n N_j^\alpha},
\]
where \( \alpha \) controls the rebalanced distribution.

We use the same model architecture with the meta-pretrained model so that the parameters of the Transformer body can be directly initialized. For the vocabulary, following Conneau et al. (2020), we utilize a shared vocabulary across different languages. Thus, the resulting vocabulary is different from the meta-pretrained model. To alleviate this issue, we employ a vocabulary matching strategy. For each word in the cross-lingual vocabulary, we first directly search the word in the monolingual vocabulary, and then search the normalized word if the original one is not found.

Note that only a small part of words can be initialized because most of the words in the vocabulary are in different languages with the meta-pretrained model. An extension of the mapping strategy is to initialize more word embeddings
with easily accessible bilingual dictionaries. The method is straightforward, just first searching the word translation in the dictionaries, and then applying the mapping strategy mentioned above. Using bilingual dictionaries is not strictly an unsupervised setting, but it is an intuitive way to utilize more knowledge from the meta-pretrained model.

### 3.3 Model Finetuning

The pretrained cross-lingual LM can be applied to a wide range of downstream tasks, where the word prediction layer is replaced with various task layers. Following Devlin et al. (2019), we apply the same task layer structures with BERT for the downstream tasks. Besides, we present three application scenarios of the pretrained model as follows.

**Cross-lingual transfer** is a basic application of pretrained cross-lingual LMs. In this scenario, LMs are finetuned with training data in a source language but evaluated in various target languages. When the source language is different from the target languages, the setting is also known as zero-shot cross-lingual transfer.

**Supervised finetuning** is a common application scenario for monolingual pretrained LMs. In this scenario, a pretrained cross-lingual LM is directly used as a monolingual model for a specific target language.

**Feature-based applications** directly use the produced contextualized sentence representations without model finetuning. For example, the extracted representations can be used for cross-lingual sentence retrieval and word alignment (Hu et al., 2020; Jalili Sabet et al., 2020; Chi et al., 2021c).

### 4 Experiments

In this section, we present extensive experiments on various downstream tasks. In specific, we first evaluate the models on natural language understanding (NLU) tasks for both the cross-lingual transfer and the supervised finetuning scenarios. The two scenarios evaluate cross-lingual transfer and generalization, respectively. Then, we evaluate the models for feature-based applications on the cross-lingual alignment tasks at both sentence level and word level.

#### 4.1 Setup

**Data** We use multilingual raw sentences extracted from Wikipedia dumps as the training data, including 67.4GB unlabeled text in 94 languages. The sentences are split into subword pieces by sentencepiece (Kudo and Richardson, 2018) using the vocabulary with 250K subwords, which is provided by XLM-R (Conneau et al., 2020). For bilingual dictionaries, we use the MUSE dictionaries (Lample et al., 2018) that consist of 45 English-centric bilingual lexicons. More details about the training data are in Appendix.

**Training** The backbone of MetaXLM is a base-size Transformer (Vaswani et al., 2017) encoder with (L=12, H=768, A=12, 270M parameters). For meta-pretraining, we directly use an English RoBERTa-base (Liu et al., 2019) model as the resulting meta-pretrained model. For cross-lingual pretraining, the model parameters are optimized with the Adam (Kingma and Ba, 2015) optimizer for 200K steps with a linear scheduled learning rate of 0.0001, and a batch size of 64. The MetaXLM pretraining procedure takes about one day with an Nvidia A100 GPU. See more pretraining hyperparameters in Appendix.

**Baseline** We compare our proposed method with the single-phase cross-lingual pretraining process, i.e., pretraining the language models from scratch. We use the unsupervised XLM (Conneau and Lample, 2019) model as the baseline model, which learns the MLM task on a multilingual unlabeled text corpus. We reimplement XLM under the same pretraining setup with MetaXLM for a fair comparison, i.e. pretraining with the same pretraining task, batch size, steps, learning rates, random seeds, etc. It means that MetaXLM and XLM only differ in whether to use a meta-pretraining phase. Note that our models are pretrained with significantly fewer training data and training steps than the current state-of-the-art pretrained models because of the GPU limitation.

#### 4.2 Cross-lingual Understanding

We evaluate MetaXLM on XTREME (Hu et al., 2020) cross-lingual understanding tasks under the cross-lingual transfer setting. We use seven XTREME datasets that provide training sets in English, and dev/test sets in multiple languages. The datasets are (1) the part-of-speech (POS) tagging dataset from the Universal Dependen-
Table 1: Evaluation results on XTREME cross-lingual understanding tasks under the cross-lingual transfer setting. Models are finetuned on the English training data but evaluated on all target languages. All results are averaged over five random seeds.

| Model                  | Structured (F1) | Question Answering (F1/EM) | Classification (Acc.) | Avg  |
|------------------------|-----------------|----------------------------|------------------------|------|
|                         | POS             | NER                        | XQuAD                  | MLQA | TyDiQA | XNLI | PAWS-X |      |
| XLM                    | 63.8            | 51.0                       | 47.3 / 33.2            | 39.9 / 25.3 | 26.8 / 15.2 | 57.9 | 77.8   | 49.2 |
| META XLM               | 63.7            | 54.0                       | 57.9 / 42.0            | 51.5 / 34.3 | 39.7 / 23.0 | 64.6 | 83.2   | 55.7 |
| +Dict                  | 67.4            | 55.0                       | 64.4 / 49.0            | 57.0 / 40.0 | 44.4 / 27.8 | 67.2 | 84.2   | 59.3 |

Pretraining with larger batch size and more training steps

M BERT (Hu et al., 2020) 70.3 62.2 64.5 / 49.4 61.4 / 44.2 59.7 / 43.9 65.4 81.9 63.1

Table 1 presents the evaluation results on XTREME cross-lingual understanding. For each task, the presented results are first averaged over all the test languages, and then averaged over five runs with different random seeds. The detailed results on all test languages can be found in Appendix. It can be observed that META XLM outperforms XLM on six out of seven downstream tasks, improving the average XTREME score from 49.2 to 55.7. It is worth mentioning that META XLM only uses more English training data than the baseline model, but it still improves the downstream task performance for not only English but also other languages. It demonstrates that our method successfully utilizes the pretrained monolingual LM to learn a better cross-lingual LM, which provides substantial gains on downstream cross-lingual understanding tasks. Besides, it can be found that simply using bilingual dictionaries for initialization provides further gains over META XLM, improving the average score from 55.7 to 59.3. On XNLI and PAWS-X, META XLM+Dict even outperforms MBERT, which is pretrained with significantly a larger batch size and more training steps (Devlin et al., 2019).
| Model   | English                  | German                  | French                  | Japanese                | Avg  |
|---------|--------------------------|-------------------------|-------------------------|-------------------------|------|
|         | Book DVD Music           | Book DVD Music          | Book DVD Music          | Book DVD Music          |      |
| XLM     | 84.1 81.8 84.7           | 82.2 79.8 82.5          | 83.6 83.0 84.2          | 74.3 75.2 79.3          | 81.2 |
| META XLM| 89.8 88.1 88.2           | 86.2 82.0 84.6          | 86.1 86.1 85.8          | 76.8 77.5 80.0          | 84.3 |
| +Dict   | 90.2 88.1 88.6           | 85.3 82.9 85.8          | 87.0 86.9 86.6          | 77.8 79.5 80.4          | 85.0 |

Table 2: Evaluation results on Amazon Reviews multilingual classification under the supervised finetuning setting. Models are finetuned and tested under the same domain and language. All results are averaged over five random seeds.

gual meta-pretrained model, which improves the generalization for other languages. We also see that using bilingual dictionaries also improves the results over META XLM in the supervised finetuning scenario, which is consistent with the cross-lingual transfer scenario.

4.4 Cross-Lingual Alignment

To evaluate whether our method encourages the alignment of the representations from different languages, we conduct experiments on the cross-lingual sentence retrieval task and the word alignment task for cross-lingual alignment at sentence level and word level, respectively.

Cross-lingual sentence retrieval The cross-lingual sentence retrieval task has been used for evaluating the cross-lingual sentence representations produced by the pretrained cross-lingual LMs (Hu et al., 2020). The goal of the task is to extract translation pairs from a bilingual document. Following Chi et al. (2021b), we use parallel corpora in 36 English-centric language pairs from the Tatoeba (Artetxe and Schwenk, 2019) as the test sets. The sentences representations are the average of the hidden vectors. Then, the corresponding sentence pairs can be extracted by the nearest neighbor searching method using the cosine similarity as the distance measure.

Figure 3 reports the averaged accuracy@1 scores on Tatoeba where the sentences representations are obtained from different layers. Surprisingly, META XLM greatly improves the retrieval accuracy over the baseline model, showing that our model produces more similar sentence representations for the translation pairs than the baseline model. It indicates that our method produces universal sentence representations that are better-aligned while only using additional unlabeled English text.

Word alignment We evaluate the word-level representations with the word alignment task. The goal of the word alignment task is to extract the corresponding word pairs given an input translation pair. We use the optimal-transport alignment method (Chi et al., 2021c) for the task. The method uses the hidden vectors as word representations, and produces the resulting aligned word pairs with optimal transport. Following Jalili Sabet et al. (2020), we use parallel corpora with golden alignments as test sets, which contain 1,244 translation pairs collected from EuroParl, WPT2003, and WPT2005. The resulting word pairs are evaluated by the alignment error rate (AER; Och and Ney 2003). A lower AER score indicates the word representations produces by the model are better-aligned across different languages.

Figure 4 illustrates the AER scores over different layers of the pretrained models, where layer-0 stands for the word embedding layer. It can be observed that META XLM produces the word alignments with greatly lower AER scores than XLM on all layers except the embedding layer. The results are consistent with the results of sen-

---

1 [www.i6.informatik.rwth-aachen.de/goldAlignment/](http://www.i6.informatik.rwth-aachen.de/goldAlignment/)
2 [web.eecs.umich.edu/~mihalcea/wpt/](http://web.eecs.umich.edu/~mihalcea/wpt/)
3 [web.eecs.umich.edu/~mihalcea/wpt05/](http://web.eecs.umich.edu/~mihalcea/wpt05/)
tence retrieval, demonstrating that our model produces better-aligned representations at both sentence level and word level. Furthermore, we observe that both META XLM and XLM obtain the best performance around layer-8, while the performance of META XLM is more stable than XLM.

5 Analysis and Discussion

5.1 Effects of Meta-Pretraining

To explore how meta-pretraining improves the cross-lingual pretraining, we analyze the cross-lingual effectiveness of XLM and META XLM at various pretraining steps. We evaluate the pretrained models with the pretraining steps ranging from 2K to 40K on the MLQA cross-lingual question answering. Note that the models are pretrained with different steps but finetuned under the same setup. Figure 5 illustrates the model performance curves along with the cross-lingual pretraining, including (a) the F1 scores averaged over the six non-English MLQA dev sets, (b) the F1 scores on the English MLQA dev set, and (c) the cross-lingual transfer gap (Hu et al., 2020). The cross-lingual transfer gap measures the difference between the score in English and the average score in the other languages. For a certain performance of English, the reduction of gap score indicates the growth of cross-lingual transferability.

Lagging phenomenon of cross-lingual transfer

We first analyze the performance curves of XLM to study the mechanism of the single-phase cross-lingual pretraining. From Figure 5 (a) and (b), we observe that the results on both English and the other languages all remain at a low level in the first 14K steps. Although the performance of XLM on English has shown a rapid improvement after the 14K step, the performance on non-English sets shows a similar trend till the 20K step. In Figure 5 (c), the transfer gap of XLM has a fast increase from the 14K step to the 20K step, while remaining a relatively stable transfer gap after the 20K step. The above observations suggest a lagging phenomenon of cross-lingual transfer in a single-phase pretraining process, i.e., the learning of cross-lingual transfer lags behind the learning of generalization. The observations are also consistent with (Dufter and Schütze, 2020).

META XLM vs. XLM

Comparing to XLM, the results of META XLM on non-English dev sets keep rising at the very beginning. Different from XLM starting with a low transfer gap, META XLM starts with a high transfer gap but continually reduces the gap during training. For the results on the English dev set, META XLM remains a high performance because of the meta-pretraining on English data. Combining the observations, it shows that our meta-pretraining method improves the cross-lingual transfer in two aspects. First, the meta-pretrained model has already obtained very good generalization ability, which enables the model to focuses on learning cross-lingual transfer. Besides, the meta-pretrained model provides good initialization for learning pretraining tasks on other languages, leading to better generalization for not only the meta-pretraining language, but also the other languages.

5.2 Effects of Transfer Direction

In addition to English-source cross-lingual transfer, we explore the cross-lingual transferability of
Figure 6: Effects of transfer direction. The models are finetuned on Amazon Reviews using the four languages as source languages, respectively.

(a) XLM (b) MetaXLM

Table 3: Effects of knowledge distillation from the monolingual model to the cross-lingual model.

| Model       | XQuAD | MLQA  | TyDiQA | XNLI | PAWS-X |
|-------------|-------|-------|--------|------|--------|
| XLM         | 47.3  | 39.9  | 26.8   | 57.9 | 77.8   |
| XLM+KD      | 45.2  | 41.1  | 30.0   | 53.8 | 81.7   |
| META XLM    | 57.9  | 51.5  | 39.7   | 64.6 | 83.2   |

Table 4: Effects of parameter freezing. ‘+freezing’ means freezing the Transformer body during pretraining.

| Model       | XQuAD | MLQA  | TyDiQA | XNLI | PAWS-X |
|-------------|-------|-------|--------|------|--------|
| XLM         |       |       |        |      |        |
| XLM+KD      | 45.2  | 41.1  | 30.0   | 53.8 | 81.7   |
| META XLM    | 45.2  | 41.1  | 30.0   | 53.8 | 81.7   |

5.3 Meta-Pretraining vs. Knowledge Distillation

Inspired by Li et al. (2021) that uses knowledge distillation (KD; Hinton et al. 2015) to improve resulting multilingual sentence embeddings. We explore whether KD is beneficial for cross-lingual LM pretraining. We pretrain a XLM model with an auxiliary KD task, where the English RoBERTa-base (Liu et al., 2019) model and the cross-lingual LM are set as the teacher model, and the student model, respectively. Note that the KD task is only performed on the English examples within the batches because the teacher model is monolingual. Table 3 presents the evaluation results of KD. XLM+KD only obtains slight improvements over XLM. Compared with META XLM, XLM+KD uses more computational cost, with one more network forwarding for each step, while obtaining much lower results. Thus, our method is a more practical way to utilize monolingual knowledge to improve cross-lingual pretraining.

5.4 Effects of Parameter Freezing

We investigate whether freezing the initialized meta-pretrained parameters improves the cross-lingual effectiveness. Inspired by MonoTrans (Artetxe et al., 2020), which freezes the Transformer body to extend an English BERT to a new language, we pretrain a variant of META XLM with the Transformer body parameters frozen, i.e., only updating the embedding layers and the language modeling head during the pretraining phase.

The evaluation results are shown in Table 4. It shows an interesting fact that META XLM+freezing outperforms XLM on three of the downstream tasks, while preserving a reasonable high performance on the other two tasks. It indicates that the Transformer body of the meta-pretrained monolingual LM can also serve as a Transformer body of a cross-lingual LM. This finding is also consistent with the findings in MonoTrans. Despite the cross-lingual transferability of the Transformer body, freezing the body does not perform better than the standard META XLM on the downstream tasks. Thus, it is a better choice to update all parameters during the cross-lingual pretraining phase.

5.5 Effects of Initialization

We conduct experiments to study the effects of initializing different components of the model. After the meta-pretraining phase, we pretrain several variants of META XLM where we initialize the Transformer body only, the word embeddings only, or both. Table 5 compares the evaluation
Table 5: Effects of initializing different components for pretraining. ‘Body’ and ‘Emb’ stand for the Transformer body and the word embeddings, respectively. ‘None’ means random initialization.

|        | XQuAD | MLQA | TyDiQA | XNLI | PAWS-X |
|--------|-------|------|--------|------|--------|
| None   | 47.3  | 39.9 | 26.8   | 57.9 | 77.8   |
| Body   | 43.5  | 38.0 | 28.4   | 51.3 | 76.1   |
| Emb    | 46.8  | 39.4 | 31.4   | 55.7 | 77.9   |
| Both   | 57.9  | 51.5 | 39.7   | 64.6 | 83.2   |

results on cross-lingual understanding. It shows that the model produces the best results when both the Transformer body and embedding layers are initialized. Differently, initializing only one of the components can provide slight improvements on TyDiQA and PAWS-X, but harms the results on the other three tasks. This suggests that the joint initialization of the two components encourages the model to extract language-invariant knowledge, which facilitates the learning of cross-lingual LMs. On the contrary, extracting language-invariant knowledge can be difficult when only one of the components is initialized.

6 Conclusion

In this paper, we propose cross-lingual language model meta-pretraining, which provides a novel paradigm for learning cross-lingual language models. Extensive experiments on ten downstream tasks demonstrate the effectiveness of our method in three application scenarios. Moreover, we show that our method improves both generalization and cross-lingual transfer, and produces universal representations that are better-aligned at both sentence level and word level, by only using additional monolingual data.

Nonetheless, we should point out that our method introduces an additional meta-pretraining phase, which potentially requires more computational resources. Fortunately, pretrained monolingual models are currently easily accessible, which can be directly used as meta-pretrained models. Besides, the meta-pretraining phase is independent of the cross-lingual pretraining phase. Thus, our method can also be easily applied to other pretrained cross-lingual models. For future work, we would like to apply our method to language model pretraining at a larger scale. Exploring cross-lingual language model meta-pretraining for natural language generation is also an interesting research topic.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In ACL 2020, pages 4623–4637. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. TACL, 7(0):597–610.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In ICLR 2020.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021a. mT6: Multilingual pretrained text-to-text transformer with translation pairs. arXiv preprint arXiv:2104.08692.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In AAAI 2020, pages 7570–7577. AAAI Press.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021b. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In NAACL 2021, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021c. Improving pretrained cross-lingual language models via self-labeled word alignment. In ACL 2021, pages 3418–3430, Online. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021d. Xlm-e: Cross-lingual language model pre-training via electra. arXiv preprint arXiv:2106.16138.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. TACL, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In
Philipp Dufter and Hinrich Schütze. 2020. Identifying

Alexis Conneau and Guillaume Lample. 2019. Cross-

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen

Leandro Rodrigues de Souza, Rodrigo Nogueira, and

Alexis Conneau, Ruty Rinott, Guillaume Lample, Ad-

Leandro Rodrigues de Souza, Rodrigo Nogueira, and

Philipp Dufter and Hinrich Schütze. 2020. Identifying

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova. 2019. BERT: Pre-training of
deep bidirectional transformers for language under-

Masoud Jalili Sabet, Philipp Dufter, François Yvon,

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Sidd-

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham

Haoyang Huang, Yaobo Liang, Nan Duan, Ming

Masoud Jalili Sabet, Philipp Dufter, François Yvon,

Karthikeyan K, Zihan Wang, Stephen Mayhew, and

Mihir Kale, Aditya Siddhant, Rami Al-Rfou, Linting

Xue, Noah Constant, and Melvin Johnson. 2021. nnMT5 - is parallel data still relevant for pre-training
massively multilingual language models? In ACL
2021, pages 683–691. Online. Association for Compu-
tational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A
method for stochastic optimization. In ICLR 2015,
San Diego, CA.

Taku Kudo and John Richardson. 2018. Sentence-
Piece: A simple and language independent subword
tokenizer and detokenizer for neural text processing.
In EMNLP: System Demonstrations, pages 66–71,
Brussels, Belgium. Association for Computational
Linguistics.

Guillaume Lample, Alexis Conneau, Marc’Aurelio
Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018.
Word translation without parallel data. In ICLR 2018.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian
Riedel, and Holger Schwenk. 2020. MLQA: Eval-
uating cross-lingual extractive question answering.
In ACL 2020, pages 7315–7330, Online. Association
for Computational Linguistics.

Zuchao Li, Kevin Parnow, Hai Zhao, Zhuosheng
Zhang, Rui Wang, Masao Utiyama, and Eiichiro
Sumita. 2021. Cross-lingual transferring of pre-
trained contextualized language models. arXiv preprint
arXiv:2107.12627.

Zehui Lin, Xia Pan, Mingxuan Wang, Xipeng Qiu,
Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-
training multilingual neural machine translation by
leveraging alignment information. In EMNLP 2020,
pages 2649–2663, Online. Association for Compu-
tational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man-
dar Joshi, Danqi Chen, Omer Levy, Mike Lewis,
Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining
approach. arXiv preprint arXiv:1907.11692.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin
Bi, Songfang Huang, Fei Huang, and Luo Si.
2021. VECO: Variable and flexible cross-lingual
pre-training for language understanding and gener-
ation. In ACL 2021, pages 3980–3994, Online.
Association for Computational Linguistics.
Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *ACL 2021*, pages 2804–2818, Online. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *ACL 2017*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. *How multilingual is multilingual BERT?* In *ACL 2019*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *ACL 2010*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. *SQuAD: 100,000+ questions for machine comprehension of text*. In *EMNLP 2016*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. *Making monolingual sentence embeddings multilingual using knowledge distillation*. In *EMNLP 2020*, pages 4512–4525, Online. Association for Computational Linguistics.

Shuo Ren, Long Zhou, Shujie Liu, Furu Wei, Ming Zhou, and Shuai Ma. 2021. *SemFace: Pre-training encoder and decoder with a semantic interface for neural machine translation*. In *ACL 2021*, pages 4518–4527, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. In *NeurIPS 2017*, pages 5998–6008. Curran Associates, Inc.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. *On learning universal representations across languages*. In *ICLR 2021*.  Shijie Wu and Mark Dredze. 2019. *Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT*. In *EMNLP 2019*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. *mt5: A massively multilingual pre-trained text-to-text transformer*. In *NAACL 2021*, pages 483–498, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *AAAI 2020*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. *PAWS-X: A cross-lingual adversarial dataset for paraphrase identification*. In *EMNLP 2019*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, and et al. 2019. *Universal dependencies 2.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
A Pretraining Data

Following Chi et al. (2021c), we sample raw sentences from Wikipedia dump\(^4\) using a re-balanced language sampling distribution (Conneau and Lample, 2019) with $\alpha = 0.7$. Table 6 shows the statistics of Wikipedia dump we use.

| Code | Size (GB) | Code | Size (GB) | Code | Size (GB) |
|------|-----------|------|-----------|------|-----------|
| af   | 0.12      | br  | 0.28      | pa  | 0.10      |
| am   | 0.01      | hu  | 0.80      | pl  | 1.55      |
| ar   | 1.29      | hy  | 0.60      | ps  | 0.04      |
| as   | 0.04      | id  | 0.52      | pt  | 1.50      |
| az   | 0.24      | is  | 0.05      | ro  | 0.42      |
| ba   | 0.13      | it  | 2.70      | ru  | 5.63      |
| be   | 0.31      | ja  | 2.65      | sa  | 0.04      |
| bg   | 0.62      | ka  | 0.37      | sd  | 0.02      |
| bn   | 0.41      | kk  | 0.29      | si  | 0.09      |
| ca   | 1.10      | km  | 0.12      | sk  | 0.21      |
| ckb  | 0.00      | ko  | 0.25      | sl  | 0.21      |
| cs   | 0.81      | ko  | 0.56      | sq  | 0.11      |
| cy   | 0.06      | ky  | 0.10      | sr  | 0.74      |
| da   | 0.33      | la  | 0.05      | sv  | 1.70      |
| de   | 5.43      | lo  | 0.01      | sw  | 0.03      |
| el   | 0.73      | lt  | 0.19      | ta  | 0.46      |
| en   | 12.58     | lv  | 0.12      | te  | 0.45      |
| eo   | 0.25      | mk  | 0.34      | tg  | 0.04      |
| es   | 3.38      | ml  | 0.28      | th  | 0.52      |
| et   | 0.23      | mn  | 0.05      | tl  | 0.04      |
| eu   | 0.24      | mr  | 0.10      | tr  | 0.43      |
| fa   | 0.66      | ms  | 0.20      | tt  | 0.09      |
| fi   | 0.68      | mt  | 0.01      | ug  | 0.03      |
| fr   | 4.00      | my  | 0.15      | uk  | 2.43      |
| ga   | 0.03      | ne  | 0.06      | ur  | 0.13      |
| gl   | 0.27      | nl  | 1.38      | uz  | 0.06      |
| gu   | 0.09      | nn  | 0.13      | vi  | 0.76      |
| he   | 1.11      | no  | 0.54      | yi  | 0.02      |
| hi   | 0.38      | or  | 0.04      | zh  | 1.08      |

Table 6: The statistics of Wikipedia dump.

B Hyperparameters

B.1 Cross-lingual Pretraining

Table 7 shows the hyperparameters for cross-lingual pretraining. Both the baseline model and our model are pretrained using the same hyperparameters.

| Hyperparameters | Value |
|-----------------|-------|
| Layers          | 12    |
| Hidden size     | 768   |
| FFN inner hidden size | 3,072 |
| Attention heads | 12    |
| Training steps  | 200K  |
| Batch size      | 64    |
| Adam $\epsilon$ | 1e-6  |
| Adam $\beta$    | (0.9, 0.98) |
| Learning rate   | 1e-4  |
| Learning rate schedule | Linear |
| Warmup steps    | 10K   |
| Gradient clipping | 1.0   |
| Weight decay    | 0.01  |
| Dropout rate    | 0.1   |

Table 7: Hyperparameters for cross-lingual pretraining.

B.2 Model Finetuning

Table 8 shows the hyperparameters for model fine-tuning.

C Detailed Results

Table 14-13 show the detailed results on XTREME cross-lingual understanding tasks.

\(^4\)https://dumps.wikimedia.org/
| Model       | en | es | de | el | ru | te | ar | vi | th | zh | hi | Avg |
|-------------|----|----|----|----|----|----|----|----|----|----|----|-----|
| XLM         | 73.2 | 52.1 | 38.2 | 50.7 | 37.1 | 41.9 | 26.8 | 49.0 | 33.7 | 35.4 | 21.2 | 39.3 | 24.7 | 48.3 | 31.8 | 44.0 | 33.1 | 49.5 | 33.6 | 37.1 | 23.9 | 47.3 | 33.2 |
| META+XLM    | 82.0 | 70.3 | 67.0 | 53.2 | 66.4 | 52.6 | 53.1 | 34.6 | 64.4 | 47.7 | 41.6 | 24.4 | 50.8 | 33.6 | 62.2 | 42.6 | 60.9 | 37.4 | 56.0 | 38.7 | 42.2 | 27.3 | 57.9 | 42.0 |
| +Dict       | 82.9 | 71.3 | 72.7 | 59.2 | 73.3 | 59.6 | 62.4 | 44.6 | 68.3 | 52.2 | 52.3 | 35.9 | 57.7 | 40.7 | 66.2 | 46.5 | 58.3 | 46.9 | 58.7 | 41.6 | 55.7 | 40.8 | 64.4 | 49.0 |

Table 9: Results on XQuAD question answering.

| Model       | en | fr | ar | bn | fi | id | ko | ru | sw | te | Avg |
|-------------|----|----|----|----|----|----|----|----|----|----|-----|
| XLM         | 50.1 | 37.2 | 28.9 | 13.5 | 20.4 | 13.1 | 23.2 | 11.0 | 29.8 | 16.4 | 19.4 | 12.0 | 34.0 | 16.1 | 22.7 | 11.1 | 12.6 | 6.8 | 26.8 | 15.2 |
| META+XLM    | 67.5 | 54.2 | 48.9 | 20.4 | 27.0 | 15.2 | 43.9 | 25.2 | 50.8 | 31.2 | 23.2 | 15.0 | 51.3 | 26.9 | 30.0 | 12.3 | 15.2 | 6.8 | 39.7 | 23.0 |
| +Dict       | 67.6 | 54.2 | 51.8 | 28.1 | 29.3 | 15.4 | 51.3 | 33.0 | 54.3 | 36.2 | 33.9 | 23.6 | 55.0 | 31.0 | 36.3 | 16.8 | 20.4 | 12.3 | 44.4 | 27.8 |

Table 10: Results on MLQA question answering.

| Model       | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | te | Avg |
|-------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-----|
| XLM         | 74.4 | 63.4 | 64.1 | 60.4 | 59.7 | 59.0 | 58.0 | 53.0 | 55.5 | 60.2 | 54.0 | 54.5 | 55.7 | 45.3 | 51.7 | 57.9 |
| META+XLM    | 82.8 | 72.2 | 73.3 | 68.9 | 64.6 | 68.1 | 66.2 | 58.3 | 63.0 | 68.9 | 61.0 | 46.8 | 58.3 | 44.8 | 56.5 | 64.6 |
| +Dict       | 83.1 | 74.2 | 75.4 | 71.4 | 68.3 | 71.7 | 70.6 | 62.8 | 65.3 | 70.2 | 65.4 | 64.5 | 63.1 | 45.7 | 56.9 | 67.2 |

Table 11: Results on TyDiQA question answering.

| Model       | en | fr | de | es | ja | ko | zh | Avg |
|-------------|----|----|----|----|----|----|----|-----|
| XLM         | 90.7 | 81.2 | 78.9 | 81.2 | 69.9 | 68.2 | 74.0 | 77.8 |
| META+XLM    | 93.8 | 88.4 | 86.2 | 87.6 | 75.5 | 72.5 | 78.2 | 83.2 |
| +Dict       | 94.3 | 89.6 | 87.7 | 89.2 | 75.9 | 73.3 | 79.5 | 84.2 |

Table 12: Results on XNLI natural language inference.

| Model       | af | ar | bg | de | el | en | es | et | eu | fa | fi | fr | he | hi | hu | id | it |
|-------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| XLM         | 81.9 | 53.8 | 83.7 | 85.3 | 79.7 | 94.5 | 84.9 | 71.2 | 55.5 | 57.7 | 72.5 | 81.6 | 55.4 | 57.5 | 76.5 | 66.5 | 83.1 |
| META+XLM    | 85.6 | 61.9 | 84.8 | 84.3 | 83.2 | 95.2 | 87.3 | 72.1 | 54.1 | 65.1 | 72.5 | 84.4 | 61.1 | 47.0 | 79.4 | 70.3 | 85.5 |
| +Dict       | 85.3 | 65.1 | 85.4 | 86.0 | 84.4 | 95.2 | 87.7 | 73.8 | 53.8 | 69.2 | 74.9 | 86.3 | 62.3 | 62.0 | 78.6 | 70.8 | 87.0 |

Table 13: Results on PAWS-X paraphrase adversaries.

| Model       | af | ar | bg | de | el | en | es | et | eu | fa | fi | fr | he | hi | hu | id | it |
|-------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| XLM         | 17.5 | 60.8 | 47.2 | 54.5 | 86.2 | 83.4 | 83.9 | 58.4 | 57.3 | 38.4 | 80.4 | 59.2 | 45.1 | 50.8 | 23.9 | 16.9 | 63.8 |
| META+XLM    | 10.8 | 60.2 | 43.6 | 51.2 | 87.4 | 86.4 | 87.3 | 53.5 | 49.1 | 33.7 | 80.1 | 58.0 | 38.7 | 53.0 | 22.0 | 14.7 | 63.7 |
| +Dict       | 29.4 | 60.8 | 49.3 | 55.9 | 87.8 | 86.9 | 88.1 | 58.8 | 53.1 | 46.4 | 77.7 | 62.3 | 50.7 | 53.6 | 22.6 | 34.0 | 67.4 |

Table 14: Results on part-of-speech tagging.
Table 15: Results on Wikiann named entity recognition.