# Enhancing Multilingual Language Model with Massive Multilingual Knowledge Triples

**Linlin Liu**[1,2*] **Xin Li**[1] **Ruidan He**[1] **Lidong Bing**[1†] **Shafiq Joty**[2,3] **Luo Si**[1]
[1]DAMO Academy, Alibaba Group
[2]Nanyang Technological University, Singapore
[3]Salesforce Research
[1]{linlin.liu, xinting.lx, ruidan.he, l.bing, luo.si}@alibaba-inc.com [2]srjoty@ntu.edu.sg

## Abstract

Knowledge-enhanced language representation learning has shown promising results across various knowledge-intensive NLP tasks. However, prior methods are limited in efficient utilization of multilingual knowledge graph (KG) data for language model (LM) pretraining. They often train LMs with KGs in indirect ways, relying on extra entity/relation embeddings to facilitate knowledge injection. In this work, we explore methods to make better use of the multilingual annotation and language agnostic property of KG triples, and present novel knowledge based multilingual language models (KMLMs) trained directly on the knowledge triples. We first generate a large amount of multilingual synthetic sentences using the Wikidata KG triples. Then based on the intra- and inter-sentence structures of the generated data, we design pretraining tasks to enable the LMs to not only memorize the factual knowledge but also learn useful logical patterns. Our pretrained KMLMs demonstrate significant performance improvements on a wide range of knowledge-intensive cross-lingual tasks, including named entity recognition (NER), factual knowledge retrieval, relation classification, and a newly designed logical reasoning task.[1]

## 1 Introduction

Pretrained Language Models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have achieved superior performances on a wide range of NLP tasks. Existing PLMs usually learn universal language representations from general-purpose large-scale corpora but do not concentrate on capturing world's factual knowledge. It has been shown that knowledge graphs (KGs), such as Wikidata (Vrandečić and Krötzsch, 2014) and Freebase (Bollacker et al., 2008), can provide rich factual information for better language understanding. Many studies have demonstrated the effectiveness of incorporating such factual knowledge into monolingual PLMs (Peters et al., 2019; Zhang et al., 2019; Liu et al., 2020a; Poerner et al., 2020; Wang et al., 2021a). Following this, a few recent attempts have been made to enhance multilingual PLMs with Wikipedia or KG triples (Calixto et al., 2021; Ri et al., 2022; Jiang et al., 2022). However, due to the structural difference between KG and texts, existing KG based pretraining often relies on extra relation/entity embeddings or additional KG encoders for knowledge enhancement. These extra embeddings/components may add significantly more parameters which in turn increase inference complexity, or cause inconsistency between pretrain and downstream tasks. For example, mLUKE (Ri et al., 2022) has to enumerate all possible entity spans for NER to minimize the inconsistency caused by entity and entity position embeddings. Other methods (Liu et al., 2020a; Jiang et al., 2022) also require KG triples to be combined with relevant natural sentences as model input during training or inference.

In this work, we propose KMLM, a novel **K**nowledge-based **M**ultilingual **L**anguage **M**odel pretrained on massive multilingual KG triples. Unlike prior knowledge enhanced models (Zhang et al., 2019; Peters et al., 2019; Liu et al., 2020a; Wang et al., 2021a), our model requires neither a separate encoder to encode entities/relations, nor heterogeneous information fusion to fuse multiple types of embeddings (e.g., entities from KGs and words from sentences). The key idea of our method is to convert the structured knowledge from KGs to sequential data which can be directly fed as input to the LM during pretraining. Specifically, we generate three types of training data – the *parallel knowledge data*, the *code-switched knowledge data*

[*] Linlin Liu is under the Joint PhD Program between Alibaba and Nanyang Technological University.
[†] Corresponding author.
[1] Our code, data and pretrained models are available at https://github.com/ntunlp/kmlm.git.



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6878–6890
December 7-11, 2022 ©2022 Association for Computational Linguistics

and the *reasoning-based data*. The first two are obtained by generating parallel or code-switched sentences from triples of Wikidata (Vrandečić and Krötzsch, 2014), a collaboratively edited multilingual KG. The reasoning-based data, containing rich logical patterns, is constructed by converting cycles from Wikidata into word sequences in different languages. We then design pretraining tasks that are operated on the parallel/code-switched data to memorize the factual knowledge across languages, and on the reasoning-based data to learn the logical patterns.

Compared to existing knowledge-enhanced pretraining methods (Zhang et al., 2019; Liu et al., 2020a; Peters et al., 2019; Jiang et al., 2022), KMLM has the following key advantages. (1) KMLM is explicitly trained to derive new knowledge through logical reasoning. Therefore, in addition to memorizing knowledge facts, it also learns the logical patterns from the data. (2) KMLM does not require a separate encoder for KG encoding, and eliminates relation/entity embeddings, which enables KMLM to be trained on a larger set of entities and relations without adding extra parameters. The token embeddings are enhanced directly with knowledge related training data. (3) KMLM does not rely on any entity linker to link the text to the corresponding KG entities, as done in existing methods (Zhang et al., 2019; Peters et al., 2019; Poerner et al., 2020). This ensures KMLM to utilize more KG triples even if they are not linked to any text data, and avoids noise caused by incorrect links. (4) KMLM keeps the model structure of the multilingual PLM without introducing any additional component during both training and inference stages. This makes the training much easier, and the trained model is directly applicable to downstream NLP tasks.

We evaluate KMLM on a wide range of knowledge-intensive cross-lingual tasks, including NER, factual knowledge retrieval, relation classification, and logical reasoning which is a novel task designed by us to test the reasoning capability of the models. Our KMLM achieves consistent and significant improvements on all knowledge-intensive tasks, meanwhile it does not sacrifice the performance on general NLP tasks.

## 2 Related Work

Knowledge-enhanced language modeling aims to incorporate knowledge, concepts and relations into the PLMs (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020), which proved to be beneficial to language understanding (Talmor et al., 2020a).

The existing approaches mainly focus on monolingual PLMs, which can be roughly divided into two lines: implicit knowledge modeling and explicit knowledge injection. Previous attempts on implicit knowledge modeling usually consist of entity-level masked language modeling (Sun et al., 2019; Liu et al., 2020a), entity-based replacement prediction (Xiong et al., 2020), knowledge embedding loss as regularization (Wang et al., 2021b) and universal knowledge-text prediction (Sun et al., 2021). In contrast to implicit knowledge modeling, the methods of explicit knowledge injection separately maintain a group of parameters for representing structural knowledge. Such methods (Zhang et al., 2019) usually require a heterogeneous information fusion component to fuse multiple types of embeddings obtained from the text and KGs. Zhang et al. (2019) and Poerner et al. (2020) employ external entity linker to discover the entities in the text and perform feature interaction between the token embeddings and entity embeddings during the encoding phase of a transformer model. Peters et al. (2019) borrow the pre-computed knowledge embeddings as the supporting features of training an internal entity linker. Wang et al. (2021a) insert an adapter component (Houlsby et al., 2019; He et al., 2021) in each transformer layer to store the learned factual knowledge.

Extending knowledge based pretraining methods to the multilingual setting has received increasing interest recently. Zhou et al. (2022b) propose an auto-regressive model trained on knowledge triples for multilingual KG completion. Calixto et al. (2021); Ri et al. (2022) attempt improving multilingual entity representation via Wikipedia hyperlink prediction, however, their methods add a large amount of parameters due to the reliance on extra entity embeddings. For example, the mLUKE$_{\text{BASE}}$ (Ri et al., 2022) model initialized with XLM-R$_{\text{BASE}}$ doubles the number of parameters (586M vs 270M). Similar to us, Jiang et al. (2022) also utilize KG for PLM pretraining. They employ KG and Wikipedia entity descriptions to inject knowledge into multilingual LM, but relation embeddings are also required to assist learning.

Moreover, the above methods only focus on memorizing the existing facts but ignore the reasoning over the unseen/implicit knowledge that

| ID | Language | Label | Aliases |
|---|---|---|---|
| Q1420 | English | motor car | auto, autocar, ... |
| | Spanish | automóvil | coche, carro, ... |
| | Hungarian | autó | gépkocsi, személyautó, ... |
| | ... | ... | ... |

Table 1: An example (https://www.wikidata.org/wiki/Q1420) of the Wikidata entity labels and aliases in multiple languages. Q1420 is the unique entity ID.

is derivable from the existing facts. Such reasoning capability is regarded as a crucial part of building consistent and controllable knowledge-based models (Talmor et al., 2020b). In this paper, our explored methods for multilingual knowledge-enhanced pretraining boost the capability of implicit knowledge reasoning, together with the purpose of consolidating knowledge modeling and multilingual pretraining (Mulcaire et al., 2019; Conneau et al., 2020; Liu et al., 2022).

## 3 Framework

In this section, we describe the proposed framework for knowledge based multilingual language model (KMLM) pretraining. We first describe the process to generate knowledge-intensive multilingual training data, followed by the pretraining tasks to train the language models to memorize factual knowledge and learn logical patterns from the generated data.

### 3.1 Knowledge Intensive Training Data

In addition to the large-scale plain text corpus that is commonly used for language model pretraining, we also generate a large amount of knowledge intensive training data from Wikidata (Vrandečić and Krötzsch, 2014), a publicly accessible knowledge base edited collaboratively. Wikidata is composed of massive amounts of KG triples $(h, r, t)$, where $h$ and $t$ are the head and tail entities respectively, $r$ is the relation type. As shown in Table 1, most of the entities, as well as the relations in Wikidata, are annotated in multiple languages. In each language, many aliases are also given though some of them are used infrequently.

**Code-Switched Synthetic Sentences** Training language models on high-quality code-switched sentences is one of the most intuitive ways to learn language agnostic representation (Winata et al., 2019), where the translations of words/phrases can be treated in a similar way as their aliases. The code

Original (en):
(motor car, designed to carry, passenger)

Code-Switched (en-fr):
motor car [mask] **conçu pour transporter** [mask] passenger.
motor car [mask] designed to carry [mask] **passager**.

Code-Switched (en-fr) & Alias-Replaced:
automobile [mask] **destiné au transport** [mask] passenger.
motor car [mask] intended to carry [mask] **passager**.

Parallel (en-fr) & Alias-Replaced:
autocar [mask] designed to carry [mask] passenger.
**voiture** [mask] **conçu pour transporter** [mask] **passager**.

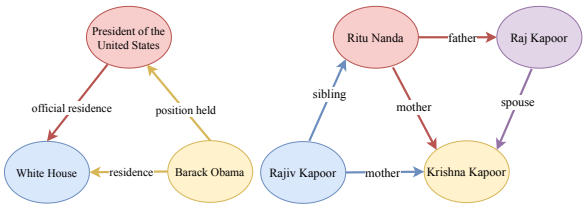
Figure 1: Examples of the en-fr code-switched and parallel synthetic sentences. The words replaced with translations or aliases are marked with bold font and underline, respectively.

mixing techniques have also proved to be helpful for improving cross-lingual transfer performance in many NLP tasks (Qin et al., 2020; Santy et al., 2021). Therefore, we propose a novel method to generate code-switched synthetic sentences using the multilingual KG triples. See Fig. 1 for some generated examples.

For each triple $(h, r, t)$ in Wikidata, we use $h_{l,0}$ to denote the default label of $h$ in language $l$. For the entity Q1420 in Table 1, $h_{en,0}$ is "motor car" and $h_{es,0}$ is "automóvi". $h_{l,i}$ denotes the aliases when the integer $i > 0$. We define $r_{l,i}$ and $t_{l,i}$ in the same way for the relation and the tail entity, respectively. Since English is resource-rich and often treated as the source language for cross-lingual transfer, we only consider language pairs of $\{(en, l')\}$ for code switching, where $l'$ is an arbitrary non-English language. With such a design, English can also work as a bridge for cross-lingual transfer between a pair of none English languages.

Specifically, the code-switched sentences for $(h, r, t)$ can be generated in 4 steps: 1) select a language pair $(en, l')$; 2) find the English default labels $(h_{en,0}, r_{en,0}, t_{en,0})$; 3) For each item in the triple, uniformly sample a value $v \in \{true, false\}$, if $v$ is $true$ and the item has a translation (i.e. default label) in $l'$, then replace the item with the translation in $l'$; 4) generate the sequence of *"h [mask] r [mask] t."* by inserting two mask tokens. The alias-replaced sentences can be generated in a similar way, except that we randomly sample aliases in the desired language to replace the default label in steps 2 and 3.

**Parallel Synthetic Sentences** Parallel data has also been widely exploited to improve cross-lingual




(a) Cycle of length 3. (b) Cycle of length 4.

Figure 2: Examples of extracted cycles of length 3 and 4.

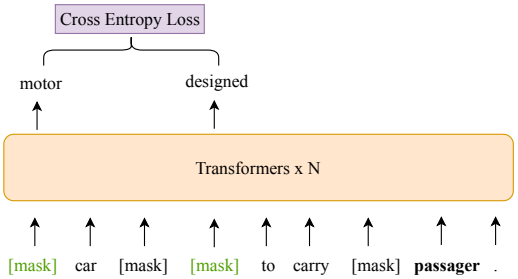



Figure 3: MLM on the code-switched synthetic sentence *"motor car [mask] designed to carry [mask]* ***passager****."*. The cross entropy loss $\mathcal{L}_K$ for **K**nowledge Oriented Pretraining is only computed over the randomly masked entity and relation tokens highlighted in lime green. For simplicity, the sub-word tokens are not shown in this example.

transfer (Aharoni et al., 2019; Conneau and Lample, 2019; Chi et al., 2021). However, it is expensive to obtain a large amount of parallel data for LM pretraining. We propose a method to generate a large amount of knowledge intensive parallel synthetic sentences, with a minor modification of the method for generating code-switched sentences described above. For each triple $(h, r, t)$ extracted from Wikidata, the corresponding synthetic sentences in different languages can be generated by first finding the default labels $(h_{l,0}, r_{l,0}, t_{l,0})$ for each language $l$, and then inserting mask tokens to generate sequences in the form *"h [mask] r [mask] t."*. Fig. 1 shows an example. More sentences can be generated by replacing the default labels with their aliases.

**Reasoning-Based Training Data** The capability of logical reasoning allows humans to solve complex problems with limited information. However, this ability did not receive much attention in the previous LM pretraining methods. In KGs, we can use nodes to represent entities, and edges between any two nodes to represent their relations. In order to train the model to learn logical patterns, we generate a large amount of reasoning-based training data by finding cycles from the Wikidata KG. As shown with an example in Fig. 2(a), the cycles of length 3 can be viewed as the basic component for more complex logical reasoning process. We train language models to learn the entity-relation co-occurrence patterns so as to infer the best candidate relations for incomplete cycles, i.e. deriving the implicit information from the given context.

Similar to the structure of the parallel/code-switched synthetic sentences described above, the cycles in Fig. 2(a) is composed of 3 triples, and hence can be converted to 3 synthetic sentences (the first example in Fig. 4). To increase the difficulty, we also extract cycles of length 4 to generate the reasoning oriented training data. However, we find that simply increasing the length of cycles makes the samples less logically coherent. Thus, we add an extra constraint that each length-4 cycle is required to have at least one additional diagonal edge. Fig. 2(b) shows such an example. It can be converted to a training sample of 5 sentences in the same way as above. For the multilingual reasoning-based data, we only generate monolingual sentences, i.e. without applying code mixing.

We treat Wikidata as an undirected graph when extracting cycles. Given an entity, the length-3 cycles containing this entity can be easily extracted by first finding all the neighbouring entities, and then iterating through the pairs of neighbouring entities to check whether they are also connected. The length-4 cycles with an additional diagonal edge connecting any two neighbours can be extracted with a few extra steps. Assuming we have identified a length-3 cycle containing entity $A$ and its two neighbouring entities $B$ and $C$, we can iterate through the neighbours of $B$ (excluding $A$ and $C$) to check whether it is also connected to $C$. We remove the duplicate cycles in data generation.

### 3.2 Pretraining Tasks

**Multilingual Knowledge Oriented Pretraining** In the generated code-switched and parallel synthetic sentences, the *"[mask]"* tokens are added between entities and relations to denote the linking words. For example, the first mask token in *"motor car [mask] designed to carry [mask]* ***passager****."* may denote *"is"*, while the second one may denote *"certains"* (French word "certains" means "some" or "certain"). Since the ground truth of such masked linking words are not known, we do not compute

the loss for those corresponding predictions. Instead, we randomly mask the remaining tokens in the parallel/code-switched synthetic sentence, and compute the cross entropy loss over these masked entity and relation tokens (Fig. 3). We use $\mathcal{L}_K$ to denote this cross entropy loss for **K**nowledge Oriented Pretraining. Note that our models are not trained on the sentence pairs like the Translation LM loss or TLM (Conneau and Lample, 2019) when utilizing the parallel or code-switched pairs. Alternatively, we shuffle the data, and feed one sentence into the model each time (as shown in Fig. 3), which makes our model inputs more consistent with those of the downstream tasks.

**Logical Reasoning Oriented Pretraining** We design tasks to train the model to learn logical reasoning patterns from the synthetic sentences generated from the length-3 and length-4 cycles. As can be seen in Fig. 4, both of the relation prediction and entity prediction problems are cast as masked language modeling. For the length-3 cycles, each entity appears exactly twice in every training sample. Formulating the task as a masked entity prediction problem may lead to shortcut learning (Geirhos et al., 2020) by simply counting the appearance numbers of the entities. Therefore, we only mask one random relation in each sample for model training, and let the model learn to predict the masked relation tokens based on the context.

Two types of tasks are designed to train the model to learn reasoning with the length-4 cycles: 1) For 80% of the time, we train the model to predict randomly masked relation and entities. We first mask one random relation. To increase the difficulty, we also mask one or two randomly selected entities at equal chance. The lower half of Fig. 4 shows an example where one relation and one entity are masked. 2) For the remaining 20% of the time, we randomly mask a whole sentence to let the model learn to derive new knowledge from the remaining context. To provide some hints on the expected new knowledge, we keep the relation of the selected sentence unmasked, i.e., only mask its two entities. The loss $\mathcal{L}_L$ for **L**ogical Reasoning Oriented Pretraining can also be computed with the cross entropy loss over the masked tokens. Note that masked entity prediction is not always non-trivial in this task. For example, when we mask exactly one entity and the entity $E$ only appears once in the masked sample, then it is easy to guess $E$ is the masked one. In Fig. 4, a concrete example is masking the first appearance of *"Raj Kapoor"* in the original sentence of the length-4 cycle. We do not deliberately avoid such cases, since they may help introduce more diversity to the training data.

**Loss Function** In addition to the pretraining tasks designed above, we also train the model on the plain text data with the original masked language modeling loss $\mathcal{L}_{MLM}$ used in previous work (Devlin et al., 2019; Conneau et al., 2020). Therefore, the final loss can be computed as:

$$\mathcal{L} = \mathcal{L}_{MLM} + \alpha(\mathcal{L}_K + \mathcal{L}_L) \quad (1)$$

where $\alpha$ is a hyper-parameter to adjust the weights of the original MLM and the losses for modeling the multilingual knowledge and logical reasoning.

## 4 Experiments

We first describe the pretraining details of our KMLMs. Then we verify its effectiveness on the knowledge-intensive tasks. Finally, we examine its performance on general cross-lingual tasks. In all of the tasks except X-FACTR (Jiang et al., 2020), the PLMs are fine-tuned on the English training set and then evaluated on the target language test sets. The evaluation results are averaged over 3 runs with different random seeds. X-FACTR does not require fine-tuning, so the PLMs are directly evaluated using the official code. The results of the baseline models are reproduced in the same environment.

### 4.1 Pretraining Details

Our proposed framework can be conveniently implemented on top of the existing transformer encoder based models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) without any modification to the model structure. Therefore, instead of pretraining the model from scratch, it is more time- and cost-efficient to initialize the model with the checkpoints of existing pretrained models. We build our knowledge intensive training data in 10 languages: English, Vietnamese, Dutch, German, French, Italian, Spanish, Japanese, Korean and Chinese. We only use the 5M entities and 822 relations filtered by Wang et al. (2021b), and generate 250M code-switched synthetic sentences, 190M parallel synthetic sentences[2] and

[2]Half of the parallel and code-switched sentences are also alias-replaced. The size of the generated parallel data is smaller than the code-switched one because some of the entities/relations do not have annotation in the target language.

**Length-3 Cycle:**
Original: President of the United States [mask] official residence [mask] White House. Barack Obama [mask] residence [mask] White House. Barack Obama [mask] position held [mask] President of the United States.

Masked: President of the United States [mask] official residence [mask] White House. Barack Obama [mask] [mask] [mask] White House. Barack Obama [mask] position held [mask] President of the United States.

**Length-4 Cycle:**
Original: Ritu Nanda [mask] father [mask] Raj Kapoor. Ritu Nanda [mask] mother [mask] Krishna Kapoor. Rajiv Kapoor [mask] mother [mask] Krishna Kapoor. Rajiv Kapoor [mask] sibling [mask] Ritu Nanda. Raj Kapoor [mask] spouse [mask] Krishna Kapoor.

Masked: Ritu Nanda [mask] father [mask] Raj Kapoor. Ritu Nanda [mask] mother [mask] [mask] [mask]. Rajiv Kapoor [mask] mother [mask] Krishna Kapoor. Rajiv Kapoor [mask] [mask] [mask] Ritu Nanda. Raj Kapoor [mask] spouse [mask] Krishna Kapoor.

Figure 4: Examples of the masked training samples for logical reasoning. The relations are highlighted in orange. The masked entity and relation tokens are highlighted in lime green.

100M reasoning-based samples following the steps in §3.1. In addition, 260M sentences are sampled from the CC100 corpus[3] (Wenzek et al., 2020) for the 10 languages. Our models KMLM-XLM-$R_{BASE}$ and KMLM-XLM-$R_{LARGE}$ are initialized with XLM-$R_{BASE}$ and XLM-$R_{LARGE}$, respectively. Then we continue to pretrain these models with the proposed tasks (§3.2). $KMLM_{CS}$, $KMLM_{Parallel}$ and $KMLM_{Mix}$ are used to differentiate the models trained on the code-switched data, parallel data and the concatenated data of these two, respectively. The reasoning-based data is used in all these three models, and ablation studies are presented in §4.5 to verify the effectiveness of logical reasoning task.

Previous studies showed that the original mBERT model outperforms XLM-R on the X-FACTR (Jiang et al., 2020) and RELX (Köksal and Özgür, 2020) tasks, so we also initialize KMLM-$mBERT_{BASE}$ with $mBERT_{BASE}$, and train it on Wikipedia corpus for a more faithful comparison[4]. We find the $KMLM_{CS}$ and $KMLM_{Mix}$ models initialized with the XLM-$R_{BASE}$ checkpoint outperform the corresponding $KMLM_{Parallel}$ model in most of the tasks, so we only train $KMLM_{CS}$ and $KMLM_{Mix}$ when comparing with XLM-$R_{LARGE}$ and $mBERT_{BASE}$. See Appendix §A.1 for more pretraining details.

## 4.2 Cross-lingual Named Entity Recognition

Named entity recognition (NER) (Lample et al., 2016; Liu et al., 2021; Zhou et al., 2022a) involves identifying and classifying named entities from unstructured text data. The elimination of entity/relation embeddings allows our models to be trained directly on a larger amount of entities without adding extra parameters or increasing computation cost. Direct training on entity-intensive synthetic sentences may also help improving entity representation more efficiently. We conduct experiments on the CoNLL02/03 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) and WikiAnn (Pan et al., 2017) NER data to verify the effectiveness of our framework. The same transformer-based NER model and hyperparameters as Hu et al. (2020) are used in our experiments.

| | en | de | nl | es | $avg_{tgt}$ | $\Delta_{avg}$ |
|---|---|---|---|---|---|---|
| $mBERT_{BASE}$ † | 90.6 | 69.2 | 77.9 | 75.4 | 74.2 | - |
| XLM-K‡ | 90.7 | 72.9 | 80.3 | 75.2 | 76.1 | - |
| XLM-$R_{BASE}$ | 91.16 | 68.87 | 79.00 | **76.70** | 74.86 | 0 |
| $KMLM_{CS}$-XLM-$R_{BASE}$ (ours) | **91.47** | 73.52 | 80.95 | 76.59 | 77.02 | +2.16 |
| $KMLM_{Parallel}$-XLM-$R_{BASE}$ (ours) | 91.44 | **73.96** | 81.06 | 75.94 | 76.99 | +2.13 |
| $KMLM_{Mix}$-XLM-$R_{BASE}$ (ours) | 91.38 | 73.95 | **81.29** | 76.17 | **77.14** | +2.28 |
| XLM-$R_{LARGE}$ | 92.98 | 73.79 | 82.00 | 79.33 | 78.37 | 0 |
| $KMLM_{CS}$-XLM-$R_{LARGE}$ (ours) | 92.81 | 76.22 | **84.12** | 78.63 | 79.66 | +1.29 |
| $KMLM_{Mix}$-XLM-$R_{LARGE}$ (ours) | **93.12** | **76.88** | 82.84 | **80.08** | **79.93** | +1.56 |

Table 2: Zero-shot cross-lingual NER F1 on the CoNLL02/03 datasets. The average results of non-English languages are reported in column $avg_{tgt}$. † The results are from (Liang et al., 2020). ‡ The results are from (Jiang et al., 2022).

The results on CoNLL02/03 data are presented in Table 2. Compared with XLM-$R_{BASE}$, all of our corresponding models improve the average F1 on target languages by more than 2.13 points. Especially on German, all of our models demonstrate at least 4.65 absolute gains. Moreover, all of our models also outperform XLM-K (Jiang et al., 2022), a knowledge-enhanced multilingual LM proposed in a recent work. Even when compared with XLM-$R_{LARGE}$, our large model still improves the average performance by 1.56. The WikiAnn dataset allows us to evaluate our models on all of the 10 languages involved in pretraining. Jiang et al. (2022) did not report XLM-K results on WikiAnn, so we evaluate their pretrained model on WikiAnn and the following knowledge intensive tasks for better comparison. As the results shown in Table 3, our best base and large models outperform the corre-

[3] http://data.statmt.org/cc-100/

[4] The original $mBERT_{BASE}$ is trained using the Wikipedia.

| | en | vi | nl | de | fr | it | es | ja | ko | zh | avg$_{tgt}$ | $\Delta_{avg}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XLM-K | 83.32 | **72.80** | 81.39 | 76.96 | 78.75 | 78.81 | 71.60 | 17.14 | **57.75** | 19.68 | 61.65 | - |
| XLM-R$_{\text{BASE}}$ | 82.59 | 68.09 | 80.08 | 74.71 | 76.50 | 77.06 | 71.05 | 20.34 | 48.46 | **26.32** | 60.29 | 0 |
| KMLM$_{\text{CS}}$-XLM-R$_{\text{BASE}}$ (ours) | 83.43 | 70.55 | 82.18 | **77.87** | 79.19 | **80.06** | 75.96 | 19.32 | 57.54 | 20.95 | 62.62 | +2.33 |
| KMLM$_{\text{Parallel}}$-XLM-R$_{\text{BASE}}$ (ours) | **83.54** | 70.93 | **82.30** | 77.79 | 78.40 | 79.83 | 76.06 | 18.16 | 57.40 | 20.44 | 62.37 | +2.08 |
| KMLM$_{\text{Mix}}$-XLM-R$_{\text{BASE}}$ (ours) | 83.42 | 70.24 | 82.22 | 77.30 | **79.93** | 80.03 | **76.72** | **20.78** | 56.70 | 22.49 | **62.93** | +2.64 |
| XLM-R$_{\text{LARGE}}$ | 84.34 | 77.61 | 83.72 | 78.92 | 79.93 | 81.24 | 73.59 | 18.94 | 59.27 | **28.35** | 64.62 | 0 |
| KMLM$_{\text{CS}}$-XLM-R$_{\text{LARGE}}$ (ours) | **85.07** | 77.89 | 84.55 | **81.32** | **83.65** | **82.57** | **78.93** | 14.95 | 60.68 | 19.43 | 64.89 | +0.27 |
| KMLM$_{\text{Mix}}$-XLM-R$_{\text{LARGE}}$ (ours) | 84.87 | **77.99** | **84.62** | 81.13 | 82.85 | 82.28 | 77.30 | **21.22** | **61.88** | 26.69 | **66.22** | +1.60 |

Table 3: Zero-shot cross-lingual NER F1 on the WikiAnn dataset.

sponding XLM-R models by 2.64 and 1.60 respectively. From both datasets we observe KMLM$_{\text{CS}}$-XLM-R$_{\text{BASE}}$ performs better than KMLM$_{\text{Parallel}}$-XLM-R$_{\text{BASE}}$, which shows the efficacy of the code-switching technique for large-scale cross-lingual pretraining. Moreover, both KMLM$_{\text{Mix}}$-XLM-R$_{\text{BASE}}$ and KMLM$_{\text{Mix}}$-XLM-R$_{\text{LARGE}}$ (i.e. the models pretrained on the mixed code-switched and parallel data) surpass all of the compared models in terms of F1, suggesting that the mixed data can help further generalize the representations across languages.

### 4.3 Factual Knowledge Retrieval

X-FACTR (Jiang et al., 2020) is a benchmark for assessing the capability of multilingual pretrained language model on capturing factual knowledge. It provides multilingual cloze-style question templates and the underlying idea is to query knowledge from the models for filling in the blank of these question templates. From (Jiang et al., 2020), we notice the performance of XLM-R$_{\text{BASE}}$ is much worse than mBERT$_{\text{BASE}}$ (see Table 4). It is probably because mBERT$_{\text{BASE}}$ has a much smaller vocabulary than XLM-R (120k vs 250k) and employs Wikipedia corpus instead of the general data crawled from the Internet. So we also pretrain KMLM$_{\text{CS}}$-mBERT$_{\text{BASE}}$ for more comprehensive comparison. As we can see from Table 4, all of the models trained with our framework demonstrate significant improvements on factual knowledge retrieval accuracy, which again indicates the benefits of our method on factual knowledge acquisition. Our model still demonstrates better performance than XLM-K, though it is also trained using Wikipedia.

### 4.4 Cross-lingual Relation Classification

RELX (Köksal and Özgür, 2020) is developed by selecting a subset of KBP-37 (Zhang and Wang, 2015), a commonly-used English relation classification dataset, and by generating human translations and annotations in French, German, Spanish, and Turkish. We evaluate the same set of models as §4.3, since mBERT$_{\text{BASE}}$ also outperforms XLM-R$_{\text{BASE}}$ on this task. The evaluation script provided by Köksal and Özgür (2020) is used to finetune the pretrained models on English training set and evaluate on the target language test sets. As the results shown in Table 5, all of our models achieves consistently higher accuracy than XLM-K and XLM-R.

Context:
(Poland, located in time zone, UTC+01:00)
(Poland, located in time zone, Central European Time)

Question:
What is the relation between UTC+01:00 and Central European Time?

Choices:
part of, said to be the same as, located in time zone, instance of, has part, followed by

Answer:
said to be the same as

Figure 5: An example (English) extracted from our cross-lingual logical reasoning (XLR) dataset.

### 4.5 Cross-lingual Logical Reasoning

**Dataset** To verify the effectiveness of our logical reasoning oriented pretraining tasks (§3.2) in an intrinsic way, we propose a cross-lingual logical reasoning (XLR) task in the form of multiple-choice questions. An example of such reasoning question is given in Fig. 5. The dataset is constructed using the cycles extracted from Wikidata. We manually annotate 1,050 samples in English and then translated them to the other 9 non-English languages (see Sec. 4.1) to build the multilingual test sets. The 3k train samples and 1k dev samples in English are also generated and cleaned automatically. The cycles used to build the test set are removed from the pretraining data, so our PLMs have never seen them beforehand. The detailed dataset construction steps can be found in Appendix §A.2.

| | en | es | fr | nl | ja | ko | vi | zh | avg |
|---|---|---|---|---|---|---|---|---|---|
| XLM-K | 7.7 | **7.3** | 3.6 | 5.0 | 0.3 | 4.0 | 5.0 | 0.9 | 4.2 |
| XLM-$R_{BASE}$ | 4.5 | 3.1 | 2.0 | 1.6 | **1.8** | 2.1 | 3.6 | 1.0 | 2.5 |
| $KMLM_{CS}$-XLM-$R_{BASE}$ (ours) | **8.6** | 4.8 | 4.2 | 5.6 | 1.6 | 4.2 | 5.8 | 3.0 | 4.7 |
| $KMLM_{Parallel}$-XLM-$R_{BASE}$ (ours) | 8.1 | 5.4 | 3.7 | **6.1** | 1.6 | 4.7 | **6.3** | 2.4 | 4.9 |
| $KMLM_{Mix}$-XLM-$R_{BASE}$ (ours) | 7.9 | 5.1 | **4.8** | **6.1** | 1.7 | **4.8** | 6.2 | **3.1** | **5.0** |
| XLM-$R_{LARGE}$ | 7.9 | 4.4 | 3.8 | 5.0 | **2.9** | 5.2 | 5.7 | 1.0 | 4.5 |
| $KMLM_{CS}$-XLM-$R_{LARGE}$ (ours) | 10.5 | 5.5 | 6.9 | 7.1 | 1.1 | 6.7 | 5.7 | 1.6 | 5.6 |
| $KMLM_{Mix}$-XLM-$R_{LARGE}$ (ours) | **11.1** | **5.8** | **7.3** | **7.7** | 1.4 | **7.1** | **6** | **3.8** | **6.3** |
| $mBERT_{BASE}$ | 8.4 | 8.7 | 5.5 | 8.6 | 1.0 | 2.0 | 4.7 | 4.5 | 5.4 |
| $KMLM_{CS}$-$mBERT_{BASE}$ (ours) | **13.0** | 10.9 | 8.5 | **11.8** | 2.0 | **3.2** | **10.1** | 10.7 | 8.8 |
| $KMLM_{Mix}$-$mBERT_{BASE}$ (ours) | 12.5 | **11.3** | **8.7** | 11.7 | **2.2** | 3 | 9.9 | **11.6** | **8.9** |

Table 4: Factual knowledge retrieval results (acc., %) on X-FACTR.

| | en | es | de | fr | $avg_{tgt}$ | $\Delta_{avg}$ |
|---|---|---|---|---|---|---|
| XLM-K | 59.1 | **58.3** | 55.9 | 56.4 | 57.4 | - |
| XLM-$R_{BASE}$ | **62.7** | 55.1 | 54.8 | 54.3 | 54.7 | 0 |
| $KMLM_{CS}$-XLM-$R_{BASE}$ (ours) | 61.2 | 57.9 | 56.6 | 55.9 | 57.9 | +3.2 |
| $KMLM_{Parallel}$-XLM-$R_{BASE}$ (ours) | 62.6 | 56.6 | **56.9** | 55.0 | 57.8 | +3.1 |
| $KMLM_{Mix}$-XLM-$R_{BASE}$ (ours) | 61.6 | 56.8 | 57 | **58.4** | **58.5** | +3.8 |
| $XLMR_{LARGE}$ | 62.8 | 62.6 | 60.4 | 59.5 | 60.8 | 0 |
| $KMLM_{CS}$-XLM-$R_{LARGE}$ (ours) | 63.5 | **63.7** | 60.1 | 60.4 | **61.4** | +0.6 |
| $KMLM_{Mix}$-XLM-$R_{LARGE}$ (ours) | **63.6** | 61.7 | **61.8** | **60.7** | **61.4** | +0.6 |
| $mBERT_{BASE}$ | **65.8** | 58.9 | 58.5 | 58.2 | 58.5 | 0 |
| $KMLM_{CS}$-$mBERT_{BASE}$ (ours) | 64.2 | 59.5 | **59.1** | **61.7** | **60.1** | +1.6 |
| $KMLM_{Mix}$-$mBERT_{BASE}$ (ours) | 60.9 | **61.3** | 57.8 | 60.3 | 59.8 | +1.3 |

Table 5: Zero-shot cross-lingual relation classification performance (acc., %) on RELX.

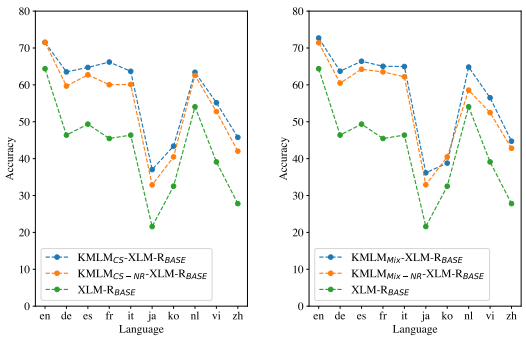



Figure 6: Comparison of the models trained with and without logical reasoning task.

**Results** We modify the multiple choice evaluation script implemented by Hugging Face[5] for this experiment. The models are finetuned on the English training set, and evaluated on the test sets in different target languages. Results are presented in Table 6. All of our models outperform the baselines significantly. Unlike on the previous tasks, where $KMLM_{Mix}$ often performs the best, $KMLM_{CS}$ shows slightly higher accuracy than $KMLM_{Mix}$. We also conduct ablation study to verify the effectiveness of our proposed logical reasoning oriented pretraining task. We pretrain the **N**one-**R**easoning models, $KMLM_{CS\text{-}NR}$-XLM-$R_{BASE}$ and $KMLM_{Mix\text{-}NR}$-XLM-$R_{BASE}$ on the same data as $KMLM_{CS}$-XLM-$R_{BASE}$ and $KMLM_{Mix}$-XLM-$R_{BASE}$, but without the logical reasoning tasks, i.e., with the MLM task only on the reasoning-based data. As presented in Fig. 6, the none-reasoning models also performs better than XLM-$R_{BASE}$, which shows the usefulness of our reasoning-based data. We also observe $KMLM_{CS}$-XLM-$R_{BASE}$ and $KMLM_{Mix}$-XLM-$R_{BASE}$, i.e., the models pretrained with logical reasoning tasks, consistently perform the best, which proves our proposed task can help models learn logical patterns more efficiently.

## 4.6 General Cross-lingual Tasks

Recall that our models are directly trained on the structured KG data. Though we attempt to minimize its difference from the natural sentences when designing the pretraining tasks, it is unknown how the difference affects cross-lingual transfer performance on the general NLP tasks. Therefore, we also evaluate our models on the part-of-speech (POS) tagging, question answering and classification tasks prepared by XTREME (Hu et al., 2020). Experimental results are shown in Table 7. Note that many of the languages covered by these tasks are not in our pretraining data, but we include all their results when computing the average performance. Overall, the performance of our models is comparable with the baselines on all of the tasks, except POS. Possibly because the POS task is more sensitive to the change of the training sentence structures. Though some of our models perform slightly better than the baselines on XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020), we find the performance gain of our models on TyDiQA[6] (Clark et al., 2020) is more obvious, which is a more challenging QA task that

[5] https://github.com/huggingface/transformers/tree/master/examples/pytorch/multiple-choice

[6] Same as XTREME, we use the gold passage version of TyDiQA.

| | en | de | es | fr | it | ja | ko | nl | vi | zh | $\text{avg}_{tgt}$ | $\Delta_{avg}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XLM-K | 58.29 | 44.19 | 44.41 | 45.14 | 41.02 | 25.94 | 34.32 | 46.92 | 38.19 | 29.81 | 38.88 | - |
| $\text{XLM-R}_{\text{BASE}}$ | 64.38 | 46.38 | 49.36 | 45.46 | 46.38 | 21.58 | 32.53 | 54.06 | 39.11 | 27.80 | 40.30 | 0 |
| $\text{KMLM}_{\text{CS}}$-$\text{XLM-R}_{\text{BASE}}$ (ours) | 71.52 | 63.52 | 64.73 | **66.19** | 63.68 | **37.04** | **43.39** | 63.42 | 55.17 | **45.77** | **55.88** | +15.58 |
| $\text{KMLM}_{\text{Parallel}}$-$\text{XLM-R}_{\text{BASE}}$ (ours) | 70.03 | 60.48 | 61.78 | 62.38 | 62.95 | 35.21 | 44.76 | 62.03 | **57.75** | 42.60 | 54.44 | +14.14 |
| $\text{KMLM}_{\text{Mix}}$-$\text{XLM-R}_{\text{BASE}}$ (ours) | **72.70** | **63.71** | **66.41** | 65.05 | **64.98** | 36.16 | 38.79 | **64.83** | 56.51 | 44.73 | 55.69 | +15.39 |
| $\text{XLM-R}_{\text{LARGE}}$ | 79.39 | 68.38 | 73.46 | 71.20 | 70.82 | 56.25 | 47.61 | 70.98 | 66.00 | 55.11 | 64.42 | 0 |
| $\text{KMLM}_{\text{CS}}$-$\text{XLM-R}_{\text{LARGE}}$ (ours) | **87.07** | **85.87** | 83.58 | **86.03** | **83.80** | 75.04 | 75.39 | 85.01 | **83.58** | **83.74** | **82.45** | +18.03 |
| $\text{KMLM}_{\text{Mix}}$-$\text{XLM-R}_{\text{LARGE}}$ (ours) | 86.67 | 84.25 | **83.75** | 85.14 | 82.89 | **76.03** | **77.30** | **85.30** | 82.86 | 82.57 | 82.23 | +17.81 |

Table 6: Zero-shot cross-lingual logical reasoning performance (acc., %).

| Metrics | POS F1 | XQuAD F1/EM | MLQA F1/EM | TyDiQA F1/EM | XNLI Acc. | PAWSX Acc. |
|---|---|---|---|---|---|---|
| $\text{XLM-R}_{\text{BASE}}$ | **72.8** | **69.6/53.9** | 64.9/47.2 | 45.2/28.5 | **73.7** | 84.8 |
| $\text{KMLM}_{\text{CS}}$-$\text{XLM-R}_{\text{BASE}}$ (ours) | 71.2 | 69.2/53.3 | 64.7/47.0 | 48.7/29.2 | 73.6 | **85.1** |
| $\text{KMLM}_{\text{Parallel}}$-$\text{XLM-R}_{\text{BASE}}$ (ours) | 71.2 | 69.3/53.4 | 64.7/46.8 | **51.4/32.8** | 73.3 | 84.1 |
| $\text{KMLM}_{\text{Mix}}$-$\text{XLM-R}_{\text{BASE}}$ (ours) | 71.4 | 69.5/53.3 | **65.4/47.3** | 49.6/32.6 | 72.9 | 84.7 |
| MMTE† | 73.5 | 64.4/46.2 | 60.3/41.4 | 58.1/43.8 | 67.4 | 81.3 |
| $\text{mBERT}^{\dagger}_{\text{LARGE}}$ | 70.3 | 64.5/49.4 | 61.4/44.2 | 59.7/43.0 | 65.4 | 81.9 |
| $\text{XLM-R}_{\text{LARGE}}$ | **74.6** | 76.8/60.9 | **72.5/54.2** | 66.6/46.6 | 79.0 | 87.8 |
| $\text{KMLM}_{\text{CS}}$-$\text{XLM-R}_{\text{LARGE}}$ (ours) | 72.4 | 76.5/60.6 | 72.0/53.7 | 66.4/47.9 | 78.6 | 87.7 |
| $\text{KMLM}_{\text{Mix}}$-$\text{XLM-R}_{\text{LARGE}}$ (ours) | 72.8 | **77.3/61.7** | 72.1/53.7 | **67.9/50.4** | **79.2** | **88.0** |

Table 7: Zero-shot cross-lingual POS, QA and classification results. Note that the performance of the languages not appearing in our prepared pretraining data are also counted. †The results are from (Hu et al., 2020).

has less lexical overlap between question-answer pairs. From these results we can see that, when our KMLMs achieve consistent improvements on the knowledge-intensive tasks, as shown by the experimental results in the previous subsections, it does not sacrifice the performance on the general NLP tasks.

## 5 Conclusions

In this paper, we have presented a novel framework for knowledge-based multilingual language pretraining. Our approach firstly creates a synthetic multilingual corpus from the existing KG and then tailor-makes two pretraining tasks, multilingual knowledge oriented pretraining and logical reasoning oriented pretraining. These multilingual pretraining tasks not only facilitate factual knowledge memorization but also boost the capability of implicit knowledge modeling. We evaluate the proposed framework on a series of knowledge-intensive cross-lingual tasks and the comparison results consistently demonstrate its effectiveness.

## Limitations

The KMLM models proposed in this work are pretrained on 10 languages in our experiments, so it is unclear whether scaling up to more languages will help further improve its performance on the downstream tasks. Due to the high computation cost, we leave it for future work. Despite the promising performance improvement on the knowledge intensive tasks, we also observe that KMLM do not perform well on the part-of-speech tagging tasks (§4.6). It is possibly caused by the large amount of synthetic sentences used in pretraining, where mask tokens are used to replace the linking words. In future, we will explore efficient ways to leverage pretrained denoising models (Liu et al., 2020b) or graph-to-sequence models (Ammanabrolu and Riedl, 2021) to convert the synthetic sentences or knowledge triples to the form more close to natural sentences.

## Ethical Impact

Neural models have achieved significant success in many NLP tasks, especially for the popular languages like English, Spanish, etc. However, neural models are data hungry, which poses challenges for applying them to the low-resource languages due to the limited NLP resources. In this work, we propose methods to inject knowledge into the multilingual pretrained language models, and enhance their logical reasoning ability. Through extensive experiments, our methods have been proven effective in a wide range of knowledge intensive multilingual NLP tasks. Therefore, our proposed method could help overcome the resource barrier, and enable the advances in NLP to benefit a wider range of population.

## Acknowledgements

This research is partly supported by the Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University. Linlin Liu would like to thank the support from Interdisciplinary Graduate School, Nanyang Technological University.

## References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

*gies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Prithviraj Ammanabrolu and Mark Riedl. 2021. Learning knowledge graph-based world models of textual environments. In *Advances in Neural Information Processing Systems*, volume 34, pages 3720–3731. Curran Associates, Inc.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901.

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting Wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3651–3661, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

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. *Transactions of the Association for Computational Linguistics*, 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*, pages 4411–4421.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. In *AAAI 2022*.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of EMNLP*, pages 5943–5959.

Abdullatif Köksal and Arzucan Özgür. 2020. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of EMNLP*, pages 340–350.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pretraining, understanding and generation. In *Proceedings of EMNLP*, pages 6008–6018.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Linlin Liu, Thien Hai Nguyen, Shafiq Joty, Lidong Bing, and Luo Si. 2022. Towards multi-sense cross-lingual alignment of contextual embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4381–4396, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020a. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of AAAI*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar 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*.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of NAACL-NLT*, pages 3912–3918.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of EMNLP*, pages 43–54.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of EMNLP*, pages 803–818.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. Bertologicomix: How does code-mixing interact with multilingual bert? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020a. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020b. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. In *Proceedings of NeurIPS*, pages 20227–20237.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of ACL*, pages 1405–1418.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

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

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280. Association for Computational Linguistics.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *Proceedings of ICLR*.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL*, pages 1441–1451.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022a. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.

Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. 2022b. Prix-LM: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5412–5424, Dublin, Ireland. Association for Computational Linguistics.

# A Appendix

## A.1 Language Model Pretraining Details

**Training Data** The statistics of the data used for pretraining are shown in Table 8.

| Description | Number |
|---|---|
| languages | 10 |
| code switched synthetic sentences | 246,783,693 |
| parallel synthetic sentences | 190,576,098 |
| unique relation combinations in length-3 cycles | 29,819 |
| unique relation combinations in length-4 cycles | 239,966 |
| reasoning based training samples from length-3 cycles | 24,142,272 |
| reasoning based training samples from length-4 cycles | 73,881,422 |
| sampled CC100 sentences (KMLM-XLM-R only) | 260,000,000 |
| sampled Wikipedia sentences (KMLM-mBERT only) | 153,011,930 |

Table 8: Statistics of the data used for pretraining.

**Hyper-Parameters** The hyper-parameters used for language model pretraining are presented in Table 9. After pretraining, we finetune the models on the plain text data with max sequence length of 512 for another 600 steps. Due to the high computation cost of LM pretraining, we do not run many experiments for hyper-parameter searching. Instead, the learning rate, batch size, mlm probability are determined according to those used in the previous LM pretraining studies. To determine the knowledge task loss weight $\alpha$ for large scale pretraining, we compare $\alpha \in \{0.5, 0.3, 0.1\}$ using the base models pretrained on a smaller dataset. Each base model takes about 30 days to train with 8 V100 GPUs.

## A.2 Cross-lingual Logical Reasoning Task

We propose a cross-lingual logical reasoning (XLR) task in the form of multiple-choice questions to verify the effectiveness of our logical reasoning oriented pretraining tasks in an intrinsic way. An example of such reasoning question is given in Fig. 5. The dataset is constructed using the length-3 and length-4 cycles extracted from Wikidata. For each cycle, we pick a triplet to create the question and answer. The question is created by asking the relation between a pair of entities in that triplet.

| Hyper-parameter | Value |
|---|---|
| learning rate | 5e-5 |
| weight decay | 0 |
| optimizer | AdamW |
| number of train epochs | 1 |
| batch size for the natural sentences | 9,600 |
| batch size for code switched knowledge data | 9,600 |
| batch size for reasoning data | 9,600 |
| mlm probability | 0.15 |
| max sequence length | 128 |
| number of warmup steps | 100 |
| knowledge task loss weight ($\alpha$ in the loss function) | 0.3 |

Table 9: Hyper-parameters used for language model pretraining.

6 choices are provided for each question (including the correct answer), which contains all of the relations appear in the cycle and some sampled relations associated with the two entities. The remaining triplets from the cycle are used as the context, which is in the form of knowledge graph (see Fig. 5). The model is required to select the most probable choice according to the given context and question. We provide correct and incorrect examples to the annotators, and manually annotate 1,050 samples in English to build the test set. The train and dev sets are automatically generated, and then cleaned by balancing the appearances of entities, relations and answers. After cleaning, we randomly select 3k train samples and 1k dev samples for the experiment. Then the multilingual test data in the other 9 non-English languages are generated by selecting the entity/relation labels in the desired languages from Wikidata. The cycles used to build the test set are removed from the pretraining data, so our PLMs have never seen them beforehand.

Statistics of the self-constructed cross-lingual logic reasoning (XLR) dataset are presented in Table 10. The multilingual test data in the 9 non-English languages are generated by selecting the entity/relation labels in the desired languages from Wikidata. So the statistics for their test sets are the same as English.

### A.3 Impact of the Logical Reasoning Tasks

As discussed in §4.5, we pretrain the **N**one-**R**easoning models, KMLM$_{\text{CS-NR}}$-XLM-R$_{\text{BASE}}$ and KMLM$_{\text{Mix-NR}}$-XLM-R$_{\text{BASE}}$ on the same data as KMLM$_{\text{CS}}$-XLM-R$_{\text{BASE}}$ and KMLM$_{\text{Mix}}$-XLM-R$_{\text{BASE}}$, but without the logical reasoning tasks. The none-reasoning models generally perform worse than the corresponding models trained with the logical reasoning tasks, which proves the usefulness of the tailored logical reasoning oriented pretraining task for logical reasoning.

| Description | Value |
|---|---|
| number of samples in the train set | 3,000 |
| number of samples in the dev set | 1,000 |
| number of samples in the test set | 1,050 |
| train set unique relation combinations | 1,419 |
| dev set unique relation combinations | 746 |
| test set unique relation combinations | 444 |

Table 10: Statistics of the self-constructed cross-lingual logic reasoning data (English).

| | en | de | nl | es | avg$_{tgt}$ |
|---|---|---|---|---|---|
| KMLM$_{\text{CS}}$-XLM-R$_{\text{BASE}}$ | 91.47 | 73.52 | 80.95 | 76.59 | 77.02 |
| KMLM$_{\text{CS-NR}}$-XLM-R$_{\text{BASE}}$ | 91.38 | 73.76 | 81.55 | 76.03 | 77.11 |

Table 11: Zero-shot cross-lingual NER F1 on the CoNLL02/03 datasets.

| | en | vi | nl | de | fr | it |
|---|---|---|---|---|---|---|
| KMLM$_{\text{CS}}$-XLM-R$_{\text{BASE}}$ | 83.43 | 70.55 | 82.18 | 77.87 | 79.19 | 80.06 |
| KMLM$_{\text{CS-NR}}$-XLM-R$_{\text{BASE}}$ | 83.75 | 70.73 | 82.44 | 78.03 | 78.88 | 80.20 |
| | es | ja | ko | zh | avg$_{tgt}$ | |
| KMLM$_{\text{CS}}$-XLM-R$_{\text{BASE}}$ | 75.96 | 19.32 | 57.54 | 20.95 | 62.62 | |
| KMLM$_{\text{CS-NR}}$-XLM-R$_{\text{BASE}}$ | 74.97 | 18.39 | 58.16 | 20.58 | 62.49 | |

Table 12: Zero-shot cross-lingual NER F1 on the WikiAnn dataset.

In order to explore the impact of the logical reasoning oriented pretraining tasks on the none-logical reasoning tasks, we also conduct ablation studies to compare the performance of KMLM$_{\text{CS-NR}}$-XLM-R$_{\text{BASE}}$ and KMLM$_{\text{CS}}$-XLM-R$_{\text{BASE}}$ on the CoNLL02/03 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) and WikiAnn (Pan et al., 2017) NER data. From the results presented in Table 11 and 12 we can see that the average performance on the target languages are very close, which shows the logical reasoning oriented pretraining tasks do not have obvious impact on zero-shot cross-lingual NER.