XLM-K: Improving Cross-Lingual Language Model Pre-training with Multilingual Knowledge

Xiaoze Jiang[1*], **Yaobo Liang**[2], **Weizhu Chen**[3], **Nan Duan**[2]
[1]Beihang University, Beijing, China
[2]Microsoft Research Asia, Beijing, China
[3]Microsoft Azure AI, Redmond, WA, USA
xzjiang@buaa.edu.cn, {yalia, wzchen, nanduan}@microsoft.com

Abstract

Cross-lingual pre-training has achieved great successes using monolingual and bilingual plain text corpora. However, most pre-trained models neglect multilingual knowledge, which is language agnostic but comprises abundant cross-lingual structure alignment. In this paper, we propose XLM-K, a cross-lingual language model incorporating multilingual knowledge in pre-training. XLM-K augments existing multilingual pre-training with two knowledge tasks, namely Masked Entity Prediction Task and Object Entailment Task. We evaluate XLM-K on MLQA, NER and XNLI. Experimental results clearly demonstrate significant improvements over existing multilingual language models. The results on MLQA and NER exhibit the superiority of XLM-K in knowledge related tasks. The success in XNLI shows a better cross-lingual transferability obtained in XLM-K. What is more, we provide a detailed probing analysis to confirm the desired knowledge captured in our pre-training regimen. The code is available at https://github.com/microsoft/Unicoder/tree/master/pretraining/xlmk.

Introduction

Recent development of pre-trained language model (Devlin et al. 2019; Liu et al. 2019) has inspired a new surge of interest in the cross-lingual scenario, such as Multilingual BERT (Devlin et al. 2019) and XLM-R (Conneau et al. 2020). Existing models are usually optimized for masked language modeling (MLM) tasks (Devlin et al. 2019) and translation tasks (Conneau and Lample 2019) using multilingual data. However, they neglect the knowledge across languages, such as entity resolution and relation reasoning. In fact, the knowledge conveys similar semantic concepts and similar meanings across languages (Vulić and Moens 2013; Chen et al. 2021), which is essential to achieve cross-lingual transferability. Therefore, how to equip pre-trained models with knowledge has become an underexplored but critical challenge for multilingual language models.

Contextual linguistic representations in language models are ordinarily trained using unlabeled and unstructured corpus, without the consideration of explicit grounding to knowledge (Févry et al. 2020; Xiong et al. 2020; Fan et al. 2021), such as entity and relation. On one side, the structural knowledge data is abundant and could be a great complement to the unstructured corpus for building a better language model. Many works have demonstrated its importance via incorporating basic knowledge into monolingual pre-trained models (Zhang et al. 2019; Staliūnaitė and Iacobacci 2020; Zhang et al. 2020; Wang et al. 2021b). On another side, knowledge is often language agnostic, e.g. different languages share the same entity via different surface forms. This can introduce a huge amount of alignment data to learn a better cross-lingual representation (Cao et al. 2018a). However, there are few existing works on exploring the multilingual entity linking and relation in the cross-lingual setting for pre-training (Huang et al. 2019; Yang et al. 2020). For example, the de-facto cross-lingual pre-training standard, i.e. MLM (Devlin et al. 2019) plus TLM (Conneau and Lample 2019), learns the correspondences between the words or sentences across the languages, neglecting the diverse background cross-lingual information behind each entity.

To address this limitation, we propose XLM-K, a cross-lingual language model incorporating multilingual knowledge in pre-training. The knowledge is injected into the XLM-K via two additional pre-trained tasks, i.e. *masked entity prediction* task and *object entailment* task. These two tasks are designed to capture the knowledge from two aspects: description semantics and structured semantics. Description semantics encourage the contextualized entity embedding in a sequence to be linked to the long entity description in the multilingual knowledge base (KB). Structured semantics, based on the triplet knowledge <*subject, relation, object*>, connect cross-lingual subject and object based on their relation and descriptions, in which the *object* is entailed by the joint of the *subject* and the *relation*. The *object* and *subject* are both represented by their description from the KB. To facilitate the cross-lingual transfer ability, on one hand, the entity and its description are from different languages. On the other hand, the textual contents of the subject and object also come from a distinct language source. We employ the contrastive learning (He et al. 2020) during pre-training to make XLM-K distinguish a positive knowledge example from a list of negative knowledge examples.

There are three main contributions in our work:

- As the first attempt, we achieve the combination be-

[*]Contribution during internship at Microsoft Research Asia.

tween the textual information and knowledge base in cross-lingual pre-training by proposing two knowledge related and cross-lingual pre-training tasks. The knowledge, connected via different languages, introduces additional information for learning a better multilingual representation.

- We evaluate XLM-K on the entity-knowledge related downstream tasks, i.e. MLQA and NER, as well as the standard multilingual benchmark XNLI. Experimental results show that XLM-K achieves new state-of-the-art results in the setting without bilingual data resource. The improvements in MLQA and NER show its superiority on knowledge related scenarios. The results on XNLI demonstrate the better cross-lingual transferability in XLM-K.
- We further perform a probing analysis (Petroni et al. 2019) on XLM-K, clearly reflecting the desired knowledge in the pre-trained models.

Related Work

Cross-Lingual Pre-training Works on cross-lingual pre-training have achieved a great success in multilingual tasks. Multilingual BERT (Devlin et al. 2019) trains a BERT model based on multilingual masked language modeling task on the monolingual corpus. XLM-R (Conneau et al. 2020) further extends the methods on a large scale corpus. These models only use monolingual data from different languages. To achieve cross-lingual token alignment, XLM (Conneau and Lample 2019) proposes translation language modeling task on parallel corpora. Unicoder (Huang et al. 2019) presents several pre-training tasks upon parallel corpora and InfoXLM (Chi et al. 2021) encourages bilingual sentence pair to be encoded more similar than the negative examples, while ERNIE-M (Ouyang et al. 2021) learns semantic alignment among multiple languages on monolingual corpora. These models leverage bilingual data to achieve better cross-lingual capability between different languages. Our method explores cross-lingual knowledge base as a new cross-lingual supervision.

Knowledge-Aware Pre-training Recent monolingual works, incorporating basic knowledge into monolingual pre-trained models, result to a better performance in downstream tasks (Rosset et al. 2020). For example, some works introduce entity information via adding a knowledge specific model structure (Broscheit 2019; Zhang et al. 2019; Févry et al. 2020). Others consider the relation information captured in the knowledge graph triples (Hayashi et al. 2020; Zhang et al. 2020; Wang et al. 2021a; Liu et al. 2020). Meanwhile, Xiong et al. (2020); Févry et al. (2020) equip monolingual language model with diverse knowledge without extra parameters. These works are almost in monolingual domain without the consideration of cross-lingual information, while the cross-lingual knowledge is learned by our model via the proposed tasks. Moreover, the standard operation of aforementioned works are mostly based on the entity names. The entity names are masked and then predicted by the model, namely the MLM task is conducted on the masked entity names. While we predict the entity description for the purpose of disambiguation of different entities with the same entity name (detailed in above). It can help our model learn more fine-grained knowledge.

Word Embedding and KB Joint Learning Many works leverage word embedding from text corpus to generate better KB embedding. Wang et al. 2014; Yamada et al. 2016; Cao et al. 2017 utilize the texts with entity mention and entity name to align word embedding and entity embedding. Toutanova et al. 2015; Han, Liu, and Sun 2016; Wu et al. 2016; Wang and Li 2016 utilize the sentences with two entity mentions as relation representation to generate better entity and relation embedding. These works mainly target to generate better graph embedding with English corpus and each entity will have a trainable embedding. Our methods focus on training a better contextualized representation for multiple languages. Meanwhile, the entity representation is generated by Transformer (Vaswani et al. 2017) model, which could further align the textual and KB embedding, as well as achieving the less trainable parameters. For cross-lingual word embeddings, most of works rely on aligned words or sentences (Ruder, Vulić, and Søgaard 2019). Cao et al. 2018b; Pan et al. 2019; Chen et al. 2021 replace the entity mention to a special entity tag and regularize one entity's different mentions in different languages to have similar embedding. Vulić and Moens 2013 use topic tag of Wikipedia pages to improve the cross-lingual ability. We also utilize entity mention in different languages as cross-lingual alignment supervision. Different from these works, we further exploit relation information to enhance the entity representation. What's more, we generate entity representation by Transformer model instead of training the separate embedding for special entity tag.

Methodology

We first present the knowledge construction strategy. Then, we introduce our two knowledge-based pre-training tasks and the training objective.

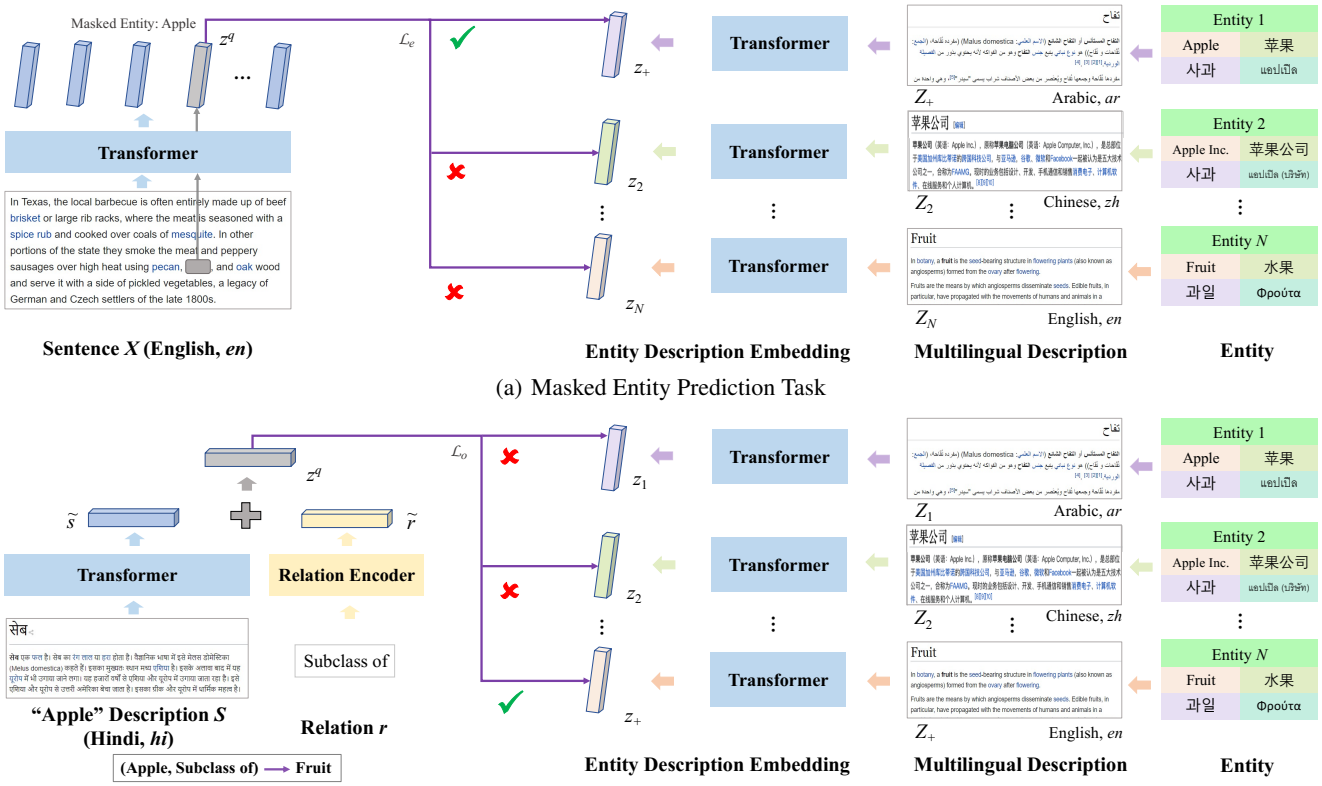
Knowledge Construction

We use *Wikipedia* and *Wikidata* (Vrandečić and Krötzsch 2014) as the data source.

Knowledge Graph A knowledge graph is a set of triplets in form <*subject, relation, object*>. We use Wikidata as our knowledge base. The triplets of Wikidata are extracted from Wikipedia and each Wikipedia page corresponding to an entity in WikiData. WikiData contains 85 million entities and 1304 relations. They formed 280 million triplets.

Entity Mention For a sentence with l words, $X = (x_1, x_2, ..., x_l)$, a mention (s, t, e) means the sub-sequence $(x_s, x_{(s+1)}, ..., x_t)$ corresponding to entity e. In our work, we use Wikipedia as data source. For each anchor in Wikipedia, it provides the link to the Wikipedia page of this entity, which can be further mapped to a unique entity in Wikidata. Wikipedia pages are from 298 languages, and every around 64 tokens contains an anchor.

Multilingual Entity Description We treat a Wikipedia page as the description of its corresponding entity in WikiData. Since Wikipedia contains multiple languages, an entity may have multiple descriptions and they come from different languages. For each page, we only keep its first 256 tokens as



(a) Masked Entity Prediction Task

(b) Object Entailment Task

Figure 1: XLM-K mainly consists of two cross-lingual pre-training tasks: (a) Masked Entity Prediction recognizes the masked entity with its knowledge description (the entity *Apple* is masked in sentence X); (b) Object Entailment predicts the textual contents of *object* with the combination of *subject* and *relation*. All the Transformers are with shared parameters.

its description. As shown in Figure 1, N multilingual entity descriptions form the candidate list $\mathcal{Z} = \{Z_1, Z_2, ..., Z_N\}$.

Masked Entity Prediction Task

Masked entity prediction task is to encourage the contextualized entity embedding in a sequence to predict the long entity description in the multilingual knowledge base (KB), rather than the prediction of entity name. It can help the disambiguation of the different entity with the same entity name. For example, as shown in Figure 1.a, the entity name of *Apple* and *Apple Inc.* are the same in Korean. It helps XLM-K learn the diverse implicit knowledge behind the mentioned words.

Given a sentence $X = (x_1, x_2, ..., x_l)$ from the cross-lingual corpus, where X is a sentence with l words from language u^{lg} (e.g. u^{lg} is *en*, shown in Figure 1.a), and a masked mention (s, t, e) (replaced by `[MASK]`), the task is to recognize the positive example Z_+ from a candidate list $\mathcal{Z}$, which contains distracting pages from multiple languages but associated with other entities. $Z_+ = (z_1, z_2, ..., z_m)$ is the description of entity e with m words from language t^{lg} (e.g. t^{lg} is *ar*, shown in Figure 1.a). Note that the description Z_+ (with a maximum 256 tokens) is extracted from the related Wikipedia page of entity e. After X being fed into the Transformer encoder, the final hidden state of x_s, denotes x_s^t, and the `[CLS]` from Z_+, denotes $\widetilde{z}$, are further fed into a non-linear projection layer (Chen et al. 2020), respectively:

$$z^q = W_2 ReLU(W_1 x_s^t) \quad (1)$$

$$z_+ = W_4 ReLU(W_3 \widetilde{z}) \quad (2)$$

where $W_1, W_3 \in \mathbb{R}^{d_w \times d_p}$ and $W_2, W_4 \in \mathbb{R}^{d_p \times d_w}$. Then the masked entity prediction loss $\mathcal{L}_e$ can be calculated by Eq. 5.

Object Entailment Task

The masked entity prediction task enriches XLM-K with sentence-level semantic knowledge, while object entailment task is designed to enhance the structured relation knowledge. As shown in Figure 1.b, given the *subject* and *relation*, the model is forced to select the *object* from the candidate list. For the purpose of entity disambiguation, the representations of *subject* and *object* are also from the long entity description.

Formally, given the subject entity's description sentence $S = (s_1, s_2, ..., s_l)$ with l words from language u^{lg} (e.g. u^{lg} is *hi*, shown in Figure 1.b), the object entity's description sentence $Z_+ = (z_1, z_2, ..., z_m)$ with m words from language t^{lg} (e.g. t^{lg} is *en*, shown in Figure 1.b) and their relation r (language agnostic), the task is to predict the object Z_+ from a cross-lingual candidate list $\mathcal{Z}$, based on S and

r. Firstly, the relation r is fed into the *Relation Encoder* (a look-up layer to output the relation embedding), and subject entity description sentence S and object entity description sentence Z_+ is fed into a separate Transformer encoder. We can get the encoded relation $\widetilde{r}$, the whole representation of subject entity description sentence $\widetilde{s}$ and object entity description sentence $\widetilde{z}$, based on their `[CLS]` in the last layer. The joint embedding of $\widetilde{s}$ and $\widetilde{r}$ is constructed as follows:

$$z^q = W_6 ReLU(W_5(\widetilde{s} + \widetilde{r})) \tag{3}$$

where $W_5 \in \mathbb{R}^{d_w \times d_p}$ and $W_6 \in \mathbb{R}^{d_p \times d_w}$ are trainable weights. The object $\widetilde{z}$ is also encoded by a non-linear projection layer:

$$z_+ = W_8 ReLU(W_7 \widetilde{z}) \tag{4}$$

where $W_7 \in \mathbb{R}^{d_w \times d_p}$ and $W_8 \in \mathbb{R}^{d_p \times d_w}$. The object entailment loss $\mathcal{L}_o$ is calculated by Eq. 5.

Joint Pre-training Objective

Although we can have different loss functions to optimise XLM-K, we choose *contrastive learning* due to its promising results in both visual representations (He et al. 2020; Chen et al. 2020) and cross-lingual pre-training (Chi et al. 2021; Pan et al. 2021). Intuitively, by distinguishing the positive sample from the negative samples using the contrastive loss, the model stores expressive knowledge acquired from the structure data. Formally, the loss can be calculated as:

$$\mathcal{L}_e(\text{and } \mathcal{L}_o) = -\log \frac{exp(z^q z_+)}{\sum_{k=1}^{N} exp(z^q z_k)} \tag{5}$$

where z_+ is the positive sample, z_k is the k-th candidate sample (encoded by the same way like z_+) and N is the size of the candidate list $\mathcal{Z}$. To avoid catastrophic forgetting of the learned knowledge from the previous training stage, we preserve the multilingual masked language modeling objective (MLM) (Devlin et al. 2019), denotes $\mathcal{L}_{\text{MLM}}$. As a result, the optimization objective of XLM-K is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_e + \mathcal{L}_o \tag{6}$$

Experiments

In this section, we will introduce implementation details of XLM-K, then, evaluate the performance of XLM-K on the downstream tasks. Lastly, we conduct probing experiments on the pre-trained models to verify the knowledge can be stored via the proposed tasks.

Implementation Details

Data and Model Structure For the multilingual masked language modeling task, we use Common Crawl dataset (Wenzek et al. 2020). The Common Crawl dataset is crawled from the whole web without restriction, which contains all the corpus from the Wikipedia. For the proposed two tasks, we use the corpus for the top 100 languages with the largest Wikipedias. The settings to balance the instances from different languages are the same as XLM-R$_{base}$ (Conneau et al. 2020). The architecture of XLM-K is set as follows: 768 hidden units, 12 heads, 12 layers, GELU activation, a dropout rate of 0.1, with a maximal input length of 256 for the proposed knowledge tasks, and 512 for MLM task.

Details of Pre-training We initialize the model with XLM-R$_{base}$ (Conneau et al. 2020) (was trained on Common Crawl), and conduct continual pre-training with the gradient accumulation of 8,192 batch size. We utilize Adam (Kingma and Ba 2015) as our optimizer. The learning rate starts with 10k warm-up steps and the peak learning rate is set to 3e-5. The size of candidate list size $N = 32$k. The candidate list is implemented as a queue, randomly initialized at the beginning of the training stage and updated by the newly encoded entities. The pre-training experiments are conducted using 16 V100 GPUs.

Details of Fine-Tuning We follow Liang et al. 2020 in these fine-tuning settings. In detail, we use Adam optimizer with warm-up and only fine-tune XLM-K on the English training set. **F**or MLQA, we fine-tune 2 epochs, with the learning rate set as 3e-5 and batch size of 12. **F**or NER, we fine-tune 20 epochs, with the learning rate set as 5e-6 and batch size of 32. **F**or XNLI, we fine-tune 10 epochs and the other settings are the same as for NER. **W**e test all the fine-tuned models on dev split of all languages for each fine-tuning epoch and select the model based on the best average performance on the dev split of all languages. To achieve a convincing comparison, we run the fine-tuning experiments with 4 random seeds and report both the average and maximum results on all downstream tasks. We also run our baseline XLM-R$_{base}$ with the same 4 seeds and report average results.

Details of Probing Following Petroni et al. (2019), we conduct probing analysis directly on the pre-trained models without any fine-tuning. The probing corpus are from four sources: Google-RE[1], T-REx (Elsahar et al. 2018), ConceptNet (Speer and Havasi 2012) and SQuAD (Rajpurkar et al. 2016). Except that ConceptNet tests for commonsense knowledge, others are all designed to probe Wiki-related knowledge.

Downstream Task Evaluation

To evaluate the performance of our model using downstream tasks, we conduct experiments on MLQA, NER and XNLI. MLQA and NER are entity-related tasks, and XNLI is a widely-used cross-lingual benchmark. Without using bilingual data in pre-training, we achieve new state-of-the-art results on these three tasks. For the convenience of reference, we display the results of the bilingual data relevant methods, namely the recently released models InfoXLM (Chi et al. 2021) and ERNIE-M (Ouyang et al. 2021), in Table 1 and Table 3 and omit the analysis. Applying bilingual data resources to XLM-K is left as future work. In the following section, **MEP** means the ablation model of Masked Entity Prediction + MLM and **OE** means Object Entailment + MLM.

MLQA MLQA (Lewis et al. 2020) is a multilingual question answering dataset, which covers 7 languages including *English*, *Spanish*, *German*, *Arabic*, *Hindi*, *Vietnamese* and *Chinese*. As a big portion of questions in MLQA are factual

[1]https://code.google.com/archive/p/relation-extraction-corpus/

Model	en	es	de	ar	hi	vi	zh	Avg
mBERT	77.7/65.2	64.3/46.6	57.9/44.3	45.7/29.8	43.8/29.7	57.1/38.6	57.5/37.3	57.7/41.6
XLM	74.9/62.4	68.0/49.8	62.2/47.6	54.8/36.3	48.8/27.3	61.4/41.8	61.1/39.6	61.6/43.5
mBERT + PPA †	79.8/ -	67.7/ -	62.3/ -	53.8/ -	57.9/ -	- / -	61.5/ -	63.8/ -
Unicoder	80.6/ -	68.6/ -	62.7/ -	57.8/ -	62.7/ -	67.5/ -	62.1/ -	66.0/ -
XLM-R$_{base}$	80.1/67.0	67.9/49.9	62.1/47.7	56.4/37.2	60.5/44.0	67.1/46.3	61.4/38.5	65.1/47.2
MEP (*avg*)	80.6/67.5	68.7/50.9	62.8/48.2	59.0/39.9	63.1/46.1	68.2/47.5	62.1/38.1	66.4/48.3
OE (*avg*)	80.8/67.8	69.1/51.2	63.2/48.6	59.0/39.6	63.7/46.3	68.5/47.3	63.0/39.5	66.7/48.6
XLM-K (*avg*)	80.8/67.7	69.3/51.6	63.2/48.9	59.8/40.5	64.3/46.9	69.0/48.0	63.1/38.8	67.1/48.9
XLM-K (*max*)	80.8/67.9	69.2/52.1	63.8/49.2	60.0/41.1	65.3/47.6	70.1/48.6	63.8/39.7	**67.7/49.5**
with Bilingual Data								
InfoXLM	81.3/68.2	69.9/51.9	64.2/49.6	60.1/40.9	65.0/47.5	70.0/48.6	64.7/41.2	67.9/49.7
ERNIE-M	81.6/68.5	70.9/52.6	65.8/50.7	61.8/41.9	65.4/47.5	70.0/49.2	65.6/41.0	**68.7/50.2**

Table 1: The results of MLQA F1/EM (exact match) scores on each language († means Post-Pre-training Alignment). The models in the second block are our ablation models MEP and OE. We run our model and ablation models four times with different seeds, where *avg* means the average results and *max* means the maximum results selected by the Avg metrics.

Model	en	es	de	nl	Avg
mBERT †	90.6	75.4	69.2	77.9	78.2
XLM-R$_{base}$ †	90.9	75.2	70.4	79.5	79.0
MEP (*avg*)	90.6	75.6	72.3	80.2	79.6
OE (*avg*)	90.9	76.0	72.7	80.1	79.9
XLM-K (*avg*)	90.7	75.2	72.9	80.3	79.8
XLM-K (*max*)	90.7	76.6	73.3	80.0	**80.1**

Table 2: The results of NER F1 scores on each language, where † means the results from (Liang et al. 2020). The models in the second block are our ablation models. The meaning of *avg* and *max* are the same as Table 1.

ones, we use it to evaluate XLM-K that is pre-trained using the multilingual knowledge.

The results on MLQA are shown in Table 1, we compare our model with mBERT (Lewis et al. 2020), XLM (Lewis et al. 2020), mBERT + PPA (Pan et al. 2021), Unicoder (Huang et al. 2019) and XLM-R$_{base}$ (Conneau et al. 2020). Since F1 and EM scores have similar observations, we take F1 scores for analysis:

(1) *The Effectiveness of XLM-K.* For the *avg* report, XLM-K achieves 67.1 averaged accuracy on F1 score, outperforming the baseline model XLM-R$_{base}$ by 2.0. For the *max* report, the model can further obtain 0.6 additional gain over the *avg* report. This clearly illustrate the superiority of XLM-K on MLQA. In addition, the model MEP and OE provide 1.3 and 1.6 improvements over XLM-R$_{base}$, respectively, which reveals that each task can capture MLQA's task-specific knowledge successfully.

(2) *The Ablation Analysis of XLM-K.* The models in the second block are the ablation models. Compared with the ablation models, XLM-K outperforms each model by 0.7 and 0.4 on Avg metrics. It indicates that the masked entity prediction and object entailment has complementary advantages on MLQA task, and the best result is achieved when using them together.

NER The cross-lingual NER (Liang et al. 2020) dataset covers 4 languages, including *English*, *Spanish*, *German* and *Dutch*, and 4 types of named entities, namely *Person*, *Location*, *Organization* and *Miscellaneous*. As shown in Table 2, compared with baseline model XLM-R$_{base}$, XLM-K improves the Avg score to 79.8 on average and 80.1 on maximum. It verifies the effectiveness of XLM-K when solving NER task. Meanwhile, the results of MEP and OE are also increased by 0.6 and 0.9 on Avg F1 score. It displays that the entity-related pre-training task has significant improvements on the entity-related downstream tasks.

XNLI The XNLI (Conneau et al. 2018) is a popular evaluation dataset for cross-lingual NLI which contains 15 languages. It's a textual inference tasks and not rerely relied on knowledge base. We present the results, comparing with mBERT (Conneau et al. 2020), XLM (Conneau et al. 2020), Unicoder (Huang et al. 2019), AMBER (Hu et al. 2021) and XLM-R$_{base}$ (Conneau et al. 2020), in Table 3 with following observations:

(1) *The Effectiveness of XLM-K.* Although XNLI is not an entity or relation -aware multilingual task, our model obtains a 0.6 gain comparing to the baseline model XLM-R$_{base}$. Each ablation model of MEP and OE improve by 0.4. These gains are marginal compared to MLQA and NER. This shows that our model is mainly works on knowledge-aware tasks. On other tasks, it won't harm the performance and even could marginally help.

(2) *The Ablation Analysis of XLM-K.* The ablation models of XLM-K on XNLI have similar results on XNLI, which increasing by 0.4 compared to the XLM-R$_{base}$ baseline 74.2. It proves each task has its own contribution to the overall improvements. Meanwhile, the ablation models still have 0.2 gap to the XLM-K, which implies the advantages towards the combination of these two tasks.

Ablation Study

The ablation analysis mentioned above demonstrates the superiority of the combination scheme of the proposed two pre-training tasks. In this section, we investigate the effectiveness of our key components.

The Effectiveness of Knowledge Tasks Our baseline model XLM-R$_{base}$ (Conneau et al. 2020) was trained on Common Crawl dataset (Wenzek et al. 2020), which covers our training data Wikipedia. As shown in Table 1, 2, 3 and 5, our

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
mBERT	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Unicoder	82.9	74.7	75.0	71.6	71.6	73.2	70.6	68.7	68.5	71.2	67.0	69.7	66.0	64.1	62.5	70.5
AMBER	84.7	76.6	76.9	74.2	72.5	74.3	73.3	73.2	70.2	73.4	65.7	71.6	66.2	59.9	61.0	71.6
XLM-R$_{base}$	84.6	78.2	79.2	77.0	75.9	77.5	75.5	72.9	72.1	74.8	71.6	73.7	69.8	64.7	65.1	74.2
MEP †	84.9	78.5	78.8	77.0	76.2	78.1	76.1	73.4	72.0	75.2	72.4	74.7	69.8	65.7	66.0	74.6
OE †	84.4	78.1	78.8	77.1	75.9	78.0	75.9	73.1	72.5	75.3	73.0	74.5	70.1	65.4	67.3	74.6
XLM-K †	84.5	78.2	78.8	77.1	76.2	78.2	76.1	73.3	72.5	75.7	72.8	74.9	70.3	65.7	67.4	74.8
XLM-K *	84.9	79.1	79.2	77.9	77.2	78.8	77.4	73.7	73.3	76.8	73.1	75.6	72.0	65.8	68.0	**75.5**
with Bilingual Data																
InfoXLM	86.4	80.6	80.8	78.9	77.8	78.9	77.6	75.6	74.0	77.0	73.7	76.7	72.0	66.4	67.1	76.2
ERNIE-M	85.5	80.1	81.2	79.2	79.1	80.4	78.1	76.8	76.3	78.3	75.8	77.4	72.9	69.5	68.8	**77.3**

Table 3: The results of XNLI test accuracy on 15 languages. The models in the second block are our ablation models, where † means *avg* results and $*$ means *max* results. The meaning of *avg* and *max* are the same as Table 1.

Model	MLQA	NER	XNLI
XLM-R$_{base}$	65.1	79.0	74.2
XLM-K w/o knowledge tasks	65.6	79.0	74.5
MEP w/o multilingual description	65.9	79.3	74.5
MEP	66.4	79.6	74.6
OE w/o contrastive loss	65.7	79.6	74.5
OE	66.7	**79.9**	74.6
XLM-K	**67.1**	79.8	**74.8**

Table 4: The ablation results on MLQA (F1 scores), NER (F1 scores) and XNLI (test accuracy) upon Avg. All the results are the *avg* mentioned in Table 1.

model XLM-K outperforms XLM-R$_{base}$ consistently. Moreover, we replace $\mathcal{L}_e$ and $\mathcal{L}_o$ in Eq. 6 with the MLM loss on multilingual Wikipedia entity descriptions. The results are shown in the second block of Table 4. Without the knowledge tasks, the performance of XLM-K w/o knowledge tasks drops by 1.5, 0.8 and 0.3 on MLQA, NER and XNLI respectively. It proves that the improvements are from the designed knowledge tasks, rather than the domain adaptation to Wikipedia. We will design more knowledge-related tasks in the future.

The Effectiveness of Multilingual Entity Description As mentioned in above, the entity knowledge, i.e. the entity-related Wikipedia page, is converted to different language resources compared to the given entity. The same operation is conducted on the *subject* and *object* in triplets, leading to the multilingual resources between the *subject* and the *object*. To study how this operation affects model performance, we report the results on the third block of Table 4. Without multilingual entity description operation, the performance of MEP w/o multilingual description drops by 0.5, 0.3 and 0.1 on MLQA, NER and XNLI respectively. It illustrates that the effectiveness of multilingual entity description. On the other hand, compared to baseline XLM-R$_{base}$, the model MEP w/o multilingual description still achieves 0.8, 0.3 and 0.3 improvements on MLQA, NER and XNLI, respectively, which reflects that applying entity description expansion without cross-lingual information in pre-training is still consistently effective for all downstream tasks.

The Effectiveness of Optimization Strategy A natural idea to introducing structural knowledge into the pre-trained language model is to classify the *relation* by the *subject* and *object* from the triplets. Motivated by this opinion, we display the results on the forth block in Table 4. The model of OE w/o contrastive loss classifies the *relation* by the concatenation of *subject* and *object* with Cross Entropy loss. Without contrastive loss, the performance drops by 1.0, 0.3 and 0.1 on MLQA, NER and XNLI respectively. This indicates the advantages of utilizing the contrastive learning towards a better cross-lingual model. We conjecture contrastive loss introduces a more challenging task than classification task. On the other hand, OE w/o contrastive loss improves the performance of baseline model XLM-R$_{base}$ from 65.1 to 65.7, 79.0 to 79.6 and 74.2 to 74.5 in MLQA, NER and XNLI, respectively. This observation certifies the importance of the structure knowledge in cross-lingual pre-training, though via an ordinary optimization strategy.

Probing Analysis

We conduct a knowledge-aware probing task based on LAMA (Petroni et al. 2019). Note that the Probing is an **analysis experiment** to evaluate how well the pre-trained language model can store the desired (Wiki) knowledge, and to explain the reason for the improvements on downstream tasks by the proposed tasks. It means that the probing is not the SOTA comparison experiment. We leave the analysis on recent multilingual LAMA (Jiang et al. 2020; Kassner, Dufter, and Schütze 2021) as our future work.

In LAMA, factual knowledge, such as <*Jack, born-in, Canada*>, is firstly converted into cloze test question, such as "*Jack was born in* ___ ". Then, a pre-trained language model is asked to predict the answer by filling in the blank of the question. There are 4 sub-tasks in the LAMA dataset. The first is **Google-RE**, which contains questions generated based on around 60k facts extracted from Wikidata and covers 3 relations. The second is **T-REx**, which contains questions generated based on a subset of Wikidata triples as well, but covers more relations (i.e. 41 relations in total). The third is **ConceptNet**, which contains questions generated based on a commonsense knowledge base (Speer, Chin,

Corpus	Relation	Statistics		XLM-R$_{base}$	XLM-K w/o K	MEP	OE	XLM-K
		# Facts	# Rel	P@1	P@1	P@1	P@1	P@1
Google-RE	birth-place	2937	1	9.3	9.8	10.1	15.0	**15.6**
	birth-date	1825	1	0.6	0.7	0.7	0.9	**1.0**
	death-place	765	1	8.0	9.1	13.4	13.8	**17.0**
	Total	5527	3	7.4	7.8	8.0	9.9	**11.2**
T-REx	1-1	937	2	48.4	49.9	52.5	50.5	**62.0**
	N-1	20006	23	22.0	25.1	27.3	21.9	**29.4**
	N-M	13096	16	17.9	21.5	25.6	22.1	**26.1**
	Total	34039	41	21.7	22.8	27.9	23.4	**29.7**
ConceptNet	Total	11458	16	**18.8**	14.2	12.0	17.6	15.7
SQuAD	Total	305	-	5.5	6.4	10.1	9.7	**11.5**

Table 5: The results of LAMA probing mean precision at one (P@1) for the baseline XLM-R$_{base}$, XLM-K w/o K (XLM-K w/o knowledge tasks), MEP, OE and XLM-K. We also reported the statistics of the facts number and relation types involved by the referred corpus.

and Havasi 2017). The last is a popular open-domain question answering dataset **SQuAD**. The number of facts and relation types covered by the each sub-task are shown in the Table 5 column *Statistics*.

Evaluation on LAMA Probing Task The LAMA probing task is conducted on the baseline model XLM-R$_{base}$, our two ablation models MEP (Masked Entity Prediction + MLM) and OE (Object Entailment + MLM), XLM-K w/o knowledge tasks, and our full model XLM-K. The results are shown in Table 5.

• **Comparison Results**

The XLM-K w/o knowledge tasks improves the performance slightly (in Google-RE, T-REx and SQuAD). It proves the improvements are from the designed tasks, rather than the domain adaptation to Wikipedia. We will detail the observations of the results on each corpus.

Google-RE XLM-K outperforms all the other models by a substantial margin, especially the baseline model XLM-R$_{base}$. It is worth noting that the two ablation models, namely MEP and OE in Table 5, realizes 0.6 and 2.5 gain respectively, which proves each knowledge-aware pre-training task can independently help pre-trained models to embed factual knowledge in a better way.

T-REx This task contains more facts and relations compared to Google-RE. XLM-K boosts the Total metrics from 21.7 to 29.7. The model MEP and model OE improves the scores by 6.2 and 1.7, respectively. These results further demonstrate the effectiveness of XLM-K on knowledge-aware tasks.

ConceptNet The ConceptNet corpus calls for the commonsense knowledge, which is a different knowledge source from Wikipedia. In this work, we mainly take Wikipedia knowledge into consideration, which can explain the worse performance on ConceptNet. Extending our model to capture more knowledge resources, such as commonsense knowledge, is our future work. Meanwhile, we notice that the performance of model OE decreases slightly compared to model MEP and XLM-K. The reason for this phenomenon may lie in that the ConceptNet is collected as the triplets-style and the relation prediction task has a great skill to handle the relation structure knowledge.

Cloze Statement	XLM-R$_{base}$	XLM-K
Phones may be made of __.	metal	*plastic*
Gnocchi is a kind of __.	beer	*food*
Tasila Mwale (born __).	in	*1984*

Table 6: Case study of LAMA probing, where the object label is the ground truth of the given statement. We compare the prediciton from our baseline XLM-R$_{base}$ and our full model XLM-K. The **subject** is highlighted with bold and the *object* is highlighted with italic.

SQuAD To investigate the performance of our model on open-domain cloze-style question answering corpus, we further evaluate the results on SQuAD. Again, our model achieves a great success on SQuAD. In detail, XLM-K achieves 11.5, which has a 6.0 gain over XLM-R$_{base}$.

• **Case Study**

To make the analysis more explicit, as shown in Table 6, we study three cases. Take the last two cases for example, to fill in the blank of "*Gnoccchi is a kind of __.*", XLM-R$_{base}$ fails to answer the question, while XLM-K successfully accomplishes the blank with "*food*". In the last case "*Tasila Mwale (born __).*", XLM-R$_{base}$ has no idea towards this fact and only predicts the answer with "*in*" to complete the phrase "*born in*". XLM-K answers this question excellently via the prediction of "*1984*". It confirms that the XLM-K is indeed equipped with more specific knowledge.

Conclusion

In this work, we present a new cross-lingual language model XLM-K to associate pre-training language model with more specific knowledge across multiple languages. Specifically, the knowledge is obtained via two knowledge-related tasks: maksed entity prediction and object entailment. Experimental results on three benchmark datasets clearly demonstrate the superiority of XLM-K. Our systematic analysis of the XLM-K advocates that XLM-K has great advantages in knowledge intensive tasks. Incorporating more diverse multilingual knowledge and jointing more advanced pre-training schemes will be addressed in future work.

References

Broscheit, S. 2019. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. In *CoNLL*, 677–685.

Cao, Y.; Hou, L.; Li, J.; Liu, Z.; Li, C.; Chen, X.; and Dong, T. 2018a. Joint Representation Learning of Cross-lingual Words and Entities via Attentive Distant Supervision. In *EMNLP*, 227–237.

Cao, Y.; Hou, L.; Li, J.; Liu, Z.; Li, C.; Chen, X.; and Dong, T. 2018b. Joint Representation Learning of Cross-lingual Words and Entities via Attentive Distant Supervision. In *EMNLP*, 227–237.

Cao, Y.; Huang, L.; Ji, H.; Chen, X.; and Li, J. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *ACL*, 1623–1633.

Chen, M.; Shi, W.; Zhou, B.; and Roth, D. 2021. Cross-lingual Entity Alignment with Incidental Supervision. In *EACL*, 645–658.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chi, Z.; Dong, L.; Wei, F.; Yang, N.; Singhal, S.; Wang, W.; Song, X.; Mao, X.-L.; Huang, H.; and Zhou, M. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *NAACL*, 3576–3588.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, 8440–8451.

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. In *NeurIPS*, 7059–7069.

Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *EMNLP*, 2475–2485.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Elsahar, H.; Vougiouklis, P.; Remaci, A.; Gravier, C.; Hare, J.; Laforest, F.; and Simperl, E. 2018. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In *LREC*, 3448–3452.

Fan, Y.; Liang, Y.; Muzio, A.; Hassan, H.; Li, H.; Zhou, M.; and Duan, N. 2021. Discovering Representation Sprachbund For Multilingual Pre-Training. In *Findings of EMNLP*, 881–894.

Févry, T.; Soares, L. B.; FitzGerald, N.; Choi, E.; and Kwiatkowski, T. 2020. Entities as Experts: Sparse Memory Access with Entity Supervision. In *EMNLP*, 4937–4951.

Han, X.; Liu, Z.; and Sun, M. 2016. Joint representation learning of text and knowledge for knowledge graph completion. In *CORR*.

Hayashi, H.; Hu, Z.; Xiong, C.; and Neubig, G. 2020. Latent relation language models. In *AAAI*, 7911–7918.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Hu, J.; Johnson, M.; Firat, O.; Siddhant, A.; and Neubig, G. 2021. Explicit Alignment Objectives for Multilingual Bidirectional Encoders. In *NAACL*, 3633–3643.

Huang, H.; Liang, Y.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; and Zhou, M. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. In *EMNLP*, 2485–2494.

Jiang, Z.; Anastasopoulos, A.; Araki, J.; Ding, H.; and Neubig, G. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. In *EMNLP*, 5943–5959.

Kassner, N.; Dufter, P.; and Schütze, H. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. In *EACL*, 3250–3258.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Lewis, P.; Oguz, B.; Rinott, R.; Riedel, S.; and Schwenk, H. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *ACL*, 7315–7330.

Liang, Y.; Duan, N.; Gong, Y.; Wu, N.; Guo, F.; Qi, W.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; Fan, X.; Zhang, R.; Agrawal, R.; Cui, E.; Wei, S.; Bharti, T.; Qiao, Y.; Chen, J.-H.; Wu, W.; Liu, S.; Yang, F.; Campos, D.; Majumder, R.; and Zhou, M. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *EMNLP*, 6008–6018.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI*, 2901–2908.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ouyang, X.; Wang, S.; Pang, C.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora. In *EMNLP*, 27–38.

Pan, L.; Hang, C.-W.; Qi, H.; Shah, A.; Yu, M.; and Potdar, S. 2021. Multilingual BERT Post-Pretraining Alignment. In *NAACL*, 210–219.

Pan, X.; Gowda, T.; Ji, H.; May, J.; and Miller, S. 2019. Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Parallel Sentence Mining. In *Workshop on DeepLo*, 56–66.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language Models as Knowledge Bases? In *EMNLP*, 2463–2473.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*, 2383–2392.

Rosset, C.; Xiong, C.; Phan, M.; Song, X.; Bennett, P.; and Tiwary, S. 2020. Knowledge-Aware Language Model Pretraining. *arXiv preprint arXiv:2007.00655*.

Ruder, S.; Vulić, I.; and Søgaard, A. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65: 569–631.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 4444–4451.

Speer, R.; and Havasi, C. 2012. Representing General Relational Knowledge in ConceptNet 5. In *LREC*, 3679–3686.

Staliūnaitė, I.; and Iacobacci, I. 2020. Compositional and Lexical Semantics in RoBERTa, BERT and DistilBERT: A Case Study on CoQA. In *EMNLP*, 7046–7056.

Toutanova, K.; Chen, D.; Pantel, P.; Poon, H.; Choudhury, P.; and Gamon, M. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*, 1499–1509.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Vulić, I.; and Moens, M.-F. 2013. Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses. In *NAACL*, 106–116.

Wang, R.; Tang, D.; Duan, N.; Wei, Z.; Huang, X.; Ji, J.; Cao, G.; Jiang, D.; and Zhou, M. 2021a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of ACL*, 1405–1418.

Wang, X.; Gao, T.; Zhu, Z.; Zhang, Z.; Liu, Z.; Li, J.; and Tang, J. 2021b. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9: 176–194.

Wang, Z.; and Li, J.-Z. 2016. Text-enhanced representation learning for knowledge graph. In *IJCAI*, 4–17.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph and text jointly embedding. In *EMNLP*, 1591–1601.

Wenzek, G.; Lachaux, M.-A.; Conneau, A.; Chaudhary, V.; Guzman, F.; Joulin, A.; and Grave, E. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *LREC*, 4003–4012.

Wu, J.; Xie, R.; Liu, Z.; and Sun, M. 2016. Knowledge representation via joint learning of sequential text and knowledge graphs. In *CORR*.

Xiong, W.; Du, J.; Wang, W. Y.; and Stoyanov, V. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *ICLR*.

Yamada, I.; Shindo, H.; Takeda, H.; and Takefuji, Y. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *CoNLL*, 250–259.

Yang, J.; Ma, S.; Zhang, D.; Wu, S.; Li, Z.; and Zhou, M. 2020. Alternating Language Modeling for Cross-Lingual Pre-Training. In *AAAI*, 9386–9393.

Zhang, H.; Liu, Z.; Xiong, C.; and Liu, Z. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In *ACL*, 2031–2043.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL*, 1441–1451.