Prix-LM: Pretraining for Multilingual Knowledge Base Construction

Anonymous ACL submission

Abstract

Knowledge bases (KBs) contain plenty of structured world and commonsense knowledge. As such, they often complement distributional text-based information and facilitate various downstream tasks. Since their manual construction is resource- and time-intensive, recent efforts have tried leveraging large pretrained language models (PLMs) to generate additional monolingual knowledge facts for KBs. However, such methods have not been attempted for building and enriching multilingual KBs. Besides wider application, such multilingual KBs can provide richer combined knowledge than monolingual (e.g., English) KBs. Knowledge expressed in different languages may be complementary and unequally distributed: this implies that the knowledge available in high-resource languages can be transferred to low-resource ones. To achieve this, it is crucial to represent multilingual knowledge in a shared/unified space. To this end, we propose a unified representation model, Prix-LM, for multilingual KB construction and completion. We leverage two types of knowledge, monolingual triples and cross-lingual links, extracted from existing multilingual KBs, and tune a multilingual language encoder XLM-R via a causal language modeling objective. Prix-LM integrates useful multilingual and KB-based factual knowledge into a single model. Experiments on standard entity-related tasks, such as link prediction in multiple languages, cross-lingual entity linking and bilingual lexicon induction, demonstrate its effectiveness, with gains reported over strong task-specialised baselines.

1 Introduction

Multilingual knowledge bases (KBs), such as DBpedia (Lehmann et al., 2015), Wikidata (Vrandečić and Krötzsch, 2014), and YAGO (Suchanek et al., 2007), provide structured knowledge expressed in multiple languages. Those KBs are modeled as knowledge graphs (KGs) that possess two types of knowledge: monolingual triples which describe relations of entities, and cross-lingual links which match entities across languages. The knowledge stored in such KGs facilitates various downstream applications such as question answering (Dai et al., 2016; Bauer et al., 2018; Wang et al., 2021b), recommendation (Zhang et al., 2016; Wang et al., 2018, 2021c), and dialogue systems (Madotto et al., 2018; Liu et al., 2019; Yang et al., 2020).

Manually constructing large-scale knowledge bases has been labor-intensive and expensive (Paulheim, 2018), leading to a surge of interest in automatic knowledge base construction (Ji et al., 2021). Recent research (Bosselut et al., 2019; Yao et al., 2019; Wang et al., 2020, inter alia) proposes to generate structured knowledge using pretrained language models (PLMs; Devlin et al. 2019), where missing elements in KB facts (i.e., triples) can be completed (i.e., filled in) by the PLM.

While these methods arguably perform well for
English, such automatic KB construction has not yet been tried for multilingual KBs – improving the knowledge in multilingual KBs would have a positive impact on applications in other languages beyond English. Moreover, KBs in multiple languages may possess complementary knowledge, and knowledge bases in low-resource languages often suffer severely from missing entities and facts. This issue could be mitigated by propagating knowledge from multiple well-populated high-resource languages’ KBs (e.g., English and French KBs) to the KBs of low-resource languages, this way ‘collectively’ improving the content stored in the full multilingual KB.1

However, training LMs to capture structural knowledge independently for each language will fall short of utilizing complementary and transferable knowledge available in other languages. Therefore, a unified representation model is required, which can capture, propagate and enrich knowledge in multilingual KBs. In this work, we thus propose to train a language model for constructing multilingual KBs. Starting from XLM-R (Conneau et al., 2020) as our base model, we then pretrain it on the multilingual DBpedia, which stores both monolingual triples and cross-lingual links (see Figure 1). We transform both types of knowledge into sequences of tokens and pretrain the language model with a causal LM objective on such transformed sequences. The monolingual triples infuse structured knowledge into the language model, while the cross-lingual links help align knowledge between different languages. This way, the proposed model Prix-LM (Pre-trained Knowledge-incorporated Cross-lingual Language Model) is capable of mapping knowledge of different languages into a unified/shared space.

We evaluate our model on four different tasks essential for automatic KB construction, covering both high-resource and low-resource languages: link prediction, cross-lingual entity linking, bilingual lexicon induction, and prompt-based LM knowledge probing. The main results across all tasks indicate that Prix-LM brings consistent and substantial gains over various state-of-the-art methods, demonstrating its effectiveness.

1This intuition is illustrated by the example in Figure 1. Consider the prediction of facts (e.g., genre) about the oldest Japanese novel The Tale of Genji. English DBpedia records its genre only as Monogatari (story), whereas complementary knowledge can be propagated from the Japanese KB, which provides finer-grained genre information, including Love Story, Royal Family Related Story, and Monogatari.

2XLM-R’s dedicated multilingual tokenizer is used to process entity and relation names in each language.
converts the sequence to the following format:

\[
\]

\(<s>\) is the special token denoting beginning of sequence; \(<s>\) is the separator token, both adopted from XLM-R. Additional special tokens \([S]\), \([P]\) and \([O]\) denote the respective roles of subject, predicate, and object of the input knowledge fact. \([EOS]\) is the end-of-sequence token.

(2) Cross-Lingual Links. As the same surface form of an entity can be associated with more than language, we use special language tokens to indicate the actual language of each entity. These extra tokens can also be interpreted as the relation between entities. The processed sequence obtains the following format:

\[
\]

\(<s>\) and \(<s>\) are the same as for monolingual triples. \([S-LAN]\) and \([O-LAN]\) denote two placeholders for language tokens, where they get replaced by the two-character ISO 639-1 codes of the source and target language, respectively. For example, if the cross-lingual connects an English entity \(L_{ond\text{os}}\) to a Spanish entity \(L_{ond\text{res}}\), the two language tokens \([EN]\) \([ES]\) will be appended to the token \([P]\). The new special tokens are randomly initialized, and optimized during training. The original special tokens are kept and also optimized.

Training Objective. The main training objective of Prix-LM is to perform completion of both monolingual knowledge triples and cross-lingual entity links (see §2). In particular, given \(X_s\) and \(X_p\), the model must predict 1) \(X_o\) from monolingual triples (i.e., \(X_p\) is a proper relation), or \(X_o\) as the cross-lingual counterpart of \(X_s\) for cross-lingual pairs (i.e., \(X_p\) is a pair of language tokens). This task can be formulated into an autoregressive language modeling training objective:

\[
L_{LM} = - \sum_{s \in \text{seen}(\text{[EOS]})} \log P(x_s \mid x_o),
\]

where \(P(x_s \mid x_o)\) is the conditional probability of generating \(x_s\) given previous subtokens. The probability of generating token \(x_s\) is calculated from the hidden state of its previous token \(h_{t-1}\) in the final layer of Transformer as follows:

\[
P(x_s \mid x_o) = \text{softmax}(W h_{t-1}),
\]

where \(W\) is a trainable parameter initialized from PLMs for subtoken prediction. Note that this training objective is applied to both monolingual knowledge triples and cross-lingual links as they can both be encoded in the same \([s, p, o]\) format.

Since models like mBERT or XLM-R rely on masked language modeling which also looks ‘into the future’, subtokens can be leaked by attention.

Therefore, we create adaptations to support causal autoregressive training using attention masks (Yang et al., 2019), so that the \(X_o\) subtokens can only access their previous subtokens. In particular, in the Transformer blocks, given the query \(Q\), key \(K\), and value \(V\), we adapt them to a causal LM:

\[
\text{att}(Q, K, V) = \text{softmax}\left(\frac{Q K^\top}{\sqrt{d}} + M\right) V,
\]

where \(Q, K, V \in \mathbb{R}^{b \times d}\); \(l\) is the length of the input sequence, \(d\) is the hidden size, \(M \in \mathbb{R}^{b \times l}\) is an attention mask, which is set as follows:

\[
M_{ij} = \begin{cases} 0 & x_i \notin X_s \cup \{\text{[EOS]}\} \\ x_i \in X_s \cup \{\text{[EOS]}\}, j \leq i \\ -\infty & x_i \in X_s \cup \{\text{[EOS]}\}, j > i \end{cases}
\]

2.2 Inference

Different downstream tasks might require different types of inference: e.g., while link prediction tasks should rely on autoregressive inference, similarity-based tasks such as cross-lingual entity linking rely on similarity-based inference, that is, finding nearest neighbors in the multilingual space. In what follows, we outline both inference types.

Autoregressive Inference. For link prediction tasks test input is in the format of \([s, p, ?]\), where the model is supposed to generate the missing \(o\) given \(s\) and \(p\). For such tasks, \(o\) comes from a known set of candidate entities \(O\). A simple way to perform inference is to construct candidate tuples \([s, p, o']\) using each \(o' \in O\) and return the one with the minimum LM loss. This straightforward approach requires encoding \(|O|\) sequences. However, as \(|O|\) can be large for high-resource languages (e.g., 2M items for English), this might yield a prohibitively expensive inference procedure. We thus propose to speed up inference by applying and adapting the constrained beam search (Anderson et al., 2017). In a nutshell, instead of calculating loss on the whole sequence, we generate one subtoken at a time and only keep several most promising sequences in the expansion set for beam search. The generation process ends when we exceed the maximum length of entities.

More precisely, given \(s\) and \(p\) (or only \(s\) when dealing with cross-lingual links), we concatenate
them as the initial sequence \( X_0 \) and initialize the sequence loss to 0. We then extend the sequence using subtokens from the PLM’s vocabulary \( \mathcal{V} \). For each subtoken \( w_1 \in \mathcal{V} \), we create a new sequence \( \{X_0, w_1\} \) and add \(- \log P \left( w_1 | X_0 \right)\) to the sequence loss. For the next round, we only keep the sequences that can be expanded to an entity in the expansion set, and retain at most \( K \) sequences with the smallest sequence loss, where \( K \) is a hyperparameter. This process is repeated until there are no more candidate sequences to be added to the expansion set. Finally, for any candidate entity \( o \in O \), if it has been generated from a corresponding candidate sequence, we set its loss to the total LM loss (sum of sequence losses), otherwise we set its loss to \( \infty \). Finally, we return the entity with the smallest loss. A more formal description of this procedure is summarized in Alg. 1 in the Appendix.

This inference variant only requires encoding at most \( L \cdot K \) sequences, where \( L \) is the maximum number of subtokens in an entity. It is much more efficient when \( L \cdot K < |O| \), which generally holds for tasks such as link prediction.

**Similarity-Based Inference.** For some tasks it is crucial to retrieve nearest neighbors (NN) via embedding similarity in the multilingual space. Based on prior findings concerning multilingual PLMs (Liu et al., 2021b) and our own preliminary experiments, out-of-the-box Prix-LM produces entity embeddings of insufficient quality. However, we can transform them into entity encoders via a simple and efficient unsupervised Mirror-BERT procedure (Liu et al., 2021a). In short, Mirror-BERT is a contrastive learning method that calibrates PLMs and converts them into strong universal lexical or sentence encoders. The NN search is then performed with the transformed “Mirror-BERT” Prix-LM variant.\(^3\)

### 3 Experiments and Results

In this section, we evaluate Prix-LM in both high-resource and low-resource languages. The focus is on four tasks that are directly or indirectly related to KB construction. 1) Link prediction (LP) is the core task for automatic KB construction since it discovers missing links given incomplete KBs. 2) Knowledge probing from LMs (LM-KP) can also be seen as a type of KB completion task as it performs entity retrieval given a subject entity and a relation. 3) Cross-lingual entity linking (XEL) and 4) Bilingual lexicon induction (BLI) can be very useful for multilingual KB construction as they help to find cross-lingual entity links.

#### 3.1 Experimental Setup

**Training Configuration.** We train our model on knowledge facts for 87 languages which are represented both in DBpedia and in XLM-R (Base). The training set comprises 52M monolingual knowledge triples and 142M cross-lingual links.

We implement our model using Huggingface’s Transformers library (Wolf et al., 2020), and primarily follow the optimization hyperparameters of XLM-R.\(^4\) For LP we use the final checkpoint; for LM-LP, results are reported using the checkpoint at 20k steps; for BLI and XEL, the checkpoint at 150k steps is used. We discuss the rationales of checkpoint selection in §3.6.

**Inference Configuration.** For similarity-based inference, as in previous work (Liu et al., 2021a) the Mirror-BERT procedure relies on the 10k most frequent English words for contrastive learning.\(^5\) For constrained beam search, used with the LP task, we set the hyperparameter \( K \) to 50.

#### 3.2 Link Prediction

**(Short) Task Description.** Following relevant prior work (Bosselut et al., 2019; Yao et al., 2019), given a subject entity \( e_1 \) and relation \( r \), the aim of the LP task is to determine the object entity \( e_2 \).

**Task Setup.** We evaluate all models on DBpedia. We randomly sample 10\% of the monolingual triples as the test set for 9 languages and use remaining data to train the model.\(^6\) The data statistics

\(^3\)For a fair comparison, we also apply the same transformation on baseline PLMs.

\(^4\)In summary: The model is trained for 5 epochs with the Adam optimizer (Kingma and Ba, 2015) using \( \beta_1 = 0.9, \beta_2 = 0.98 \) and a batch size of 1,024. The learning rate is 5e–5, with a warmup for the first 6\% steps followed by a linear learning rate decay to 0. We use dropout (Srivastava et al., 2014) with a rate of 0.1 on all layers and attention weights. For efficiency, we drop all triples with sequence lengths \( \geq 30 \), which only constitutes less than 1.3\% of all triples. The full training takes about 5 days with one Nvidia RTX 8000 GPU.

\(^5\)We use English words only for simplicity and direct comparisons. According to Liu et al. (2021a), Mirror-BERT tuning which uses words from the actual test language pair might yield even better performance. Our training config is identical to the original Mirror-BERT work, except the use of a smaller batch size (128 instead of 200) due to hardware constraints.

\(^6\)Following Bordes et al. (2013), we use the filtered setting, removing corrupted triples appearing in the training or test set. Moreover, following existing LP tasks (Toutanova et al., 2015; Dettmers et al., 2018) we remove redundant triples \((e_1, r_1, e_2)\) from the test set if \((e_2, r_2, e_1)\) appears in the training set.
### Results and Discussion

The results in Tab. 1 show that the Prix-LM (All) achieves the best Hits@1 on average, outperforming TransE, ComplEx, and RotatE by 21.5%, 11.8%, and 5.6%, respectively. It also outperforms the baselines on Hits@3 and Hits@10. Moreover, Prix-LM (All) outperforms in almost all languages its monolingual counterpart Prix-LM (Single): the average improvements are > 3% across all metrics, demonstrating that the model can effectively leverage complementary knowledge captured and transferred through massive pretraining on multiple languages. Interestingly, the advantages of Prix-LM (both Single and All models) over baselines are not restricted to low resource languages but are observed across the board. This hints that, beyond integrating multilingual knowledge, Prix-LM is essentially a well-suited framework for KB completion in general.

### 3.3 Cross-lingual Entity Linking

#### (Short) Task Description

In XEL\(^9\), a model is asked to link an entity mention in any language to a corresponding entity in an English KB or in a language-agnostic KB.\(^10\) XEL can contribute to multilingual KB construction in two ways. First, since XEL links mentions extracted from free text to KBs, it can be leveraged to enrich KBs with textual attributes. Second, it also provides a way of epochs as Prix-LM (All): this means that the embeddings of subtokens in the test language are updated for the same number of times.

---

\(^7\)We do not calculate mean rank and mean reciprocal rank as constrained beam search does not yield full ranked lists.

\(^8\)The KG embedding baselines are implemented based on OpenKE (Han et al., 2018) and trained using the default hyper-parameters in the library.

\(^9\)XEL in our work refers only to entity mention disambiguation; it does not cover the mention detection subtask.

\(^10\)A language-agnostic KB has universal interlingual concepts without being restricted to a specific language.
Table 3: XEL Accuracy on XL-BEL.

<table>
<thead>
<tr>
<th>lang.→</th>
<th>EN</th>
<th>IT</th>
<th>DE</th>
<th>FR</th>
<th>FI</th>
<th>ET</th>
<th>TR</th>
<th>KO</th>
<th>ZH</th>
<th>JA</th>
<th>TH</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLM-R + Mirror</td>
<td>75.4</td>
<td>34.0</td>
<td>13.7</td>
<td>4.2</td>
<td>7.4</td>
<td>19.5</td>
<td>1.8</td>
<td>1.4</td>
<td>4.2</td>
<td>7.0</td>
<td>12.5</td>
<td>16.3</td>
</tr>
<tr>
<td>mBERT + Mirror</td>
<td>73.1</td>
<td>40.1</td>
<td>16.6</td>
<td>4.4</td>
<td>5.0</td>
<td>22.0</td>
<td>1.9</td>
<td>1.1</td>
<td>2.3</td>
<td>2.4</td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td>Prix-LM (Single) + Mirror</td>
<td>75.4</td>
<td>39.5</td>
<td>16.9</td>
<td>8.4</td>
<td>12.4</td>
<td>27.4</td>
<td>2.1</td>
<td>3.5</td>
<td>4.1</td>
<td>6.9</td>
<td>19.7</td>
<td></td>
</tr>
<tr>
<td>Prix-LM (All) + Mirror</td>
<td>71.9</td>
<td>49.2</td>
<td>25.7</td>
<td>15.2</td>
<td>24.5</td>
<td>34.1</td>
<td>9.3</td>
<td>6.9</td>
<td>13.7</td>
<td>14.5</td>
<td>26.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Accuracy and MRR for BLI. mBERT results are omitted since it performs much worse than XLM-R.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>XLM-R + Mirror</td>
<td>12.0</td>
<td>16.6</td>
<td>6.9</td>
<td>8.6</td>
<td>2.9</td>
<td>5.9</td>
<td>5.9</td>
<td>7.4</td>
<td>2.0</td>
<td>3.3</td>
<td>5.7</td>
</tr>
<tr>
<td>Prix-LM + Mirror</td>
<td>11.5</td>
<td>20.4</td>
<td>6.7</td>
<td>11.1</td>
<td>3.7</td>
<td>11.4</td>
<td>6.9</td>
<td>11.5</td>
<td>4.2</td>
<td>9.0</td>
<td>7.7</td>
</tr>
</tbody>
</table>

Table 5: Accuracy on mLAMA.

<table>
<thead>
<tr>
<th>lang.→</th>
<th>EN</th>
<th>IT</th>
<th>DE</th>
<th>FR</th>
<th>FI</th>
<th>ET</th>
<th>TR</th>
<th>KO</th>
<th>ZH</th>
<th>JA</th>
<th>TH</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLM-R</td>
<td>21.0</td>
<td>19.3</td>
<td>13.9</td>
<td>7.6</td>
<td>5.6</td>
<td>6.1</td>
<td>20.5</td>
<td>6.1</td>
<td>12.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prix-LM</td>
<td>23.8</td>
<td>21.8</td>
<td>20.7</td>
<td>17.8</td>
<td>16.1</td>
<td>7.4</td>
<td>23.9</td>
<td>13.1</td>
<td>18.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To disambiguate knowledge with similar surface forms but different grounded contexts.

**Task Setup.** We evaluate Prix-LM on two XEL benchmarks: (i) the Low-resource XEL benchmark (LR-XEL; Zhou et al. 2020) and (ii) cross-lingual biomedical entity linking (XL-BEL; Liu et al. 2021b). LR-XEL covers three low-resource languages te, lo, and mr\(^{11}\) where the model needs to associate mentions in those languages to the English Wikipedia pages. XL-BEL covers ten typologically diverse languages (see Tab. 3 for the full list). It requires the model to link an entity mention to entries in UMLS (Bodenreider, 2004), a language-agnostic medical knowledge base.

**Models in Comparison.** For XEL and all following tasks, we use multilingual MLMs (i.e. mBERT and XLM-R) as our baselines as they are the canonical models frequently used in prior work and have shown promising results in cross-lingual entity-centric tasks (Vulić et al., 2020; Liu et al., 2021b; Kassner et al., 2021). We remind the reader that the ‘Mirror-BERT’ fine-tuning step is always applied, yielding an increase in performance.

**Results and Discussion.** On LR-XEL, Prix-LM achieves gains for all three languages over its base model XLM-R. Especially on mr, where XLM-R and mBERT are almost fully ineffective, Prix-LM leads to over 20% of absolute accuracy gain, again showing the effectiveness of incorporating multilingual structural knowledge. On lo, mBERT is slightly better than Prix-LM, but Prix-LM again yields gains over its base model: XLM-R. On XL-BEL, a large increase is again observed for almost all target languages (see Prix-LM (All) + Mirror). The only exception is English, where the model performance drops by 3.5%. This is likely to be a consequence of trading-off some of the extensive English knowledge when learning on multilingual triples. Beyond English, substantial improvements are obtained in other Indo-European languages including Spanish, German and Russian (+10-20%), stressing the necessity of knowledge injection even for high-resource languages. Like LP, we also experimented with Prix-LM trained with only monolingual data (see Prix-LM (Single) + Mirror). Except for English, very large boosts are obtained on all other languages when comparing All and Single models, confirming that multilingual training has provided substantial complementary knowledge.

**3.4 Bilingual Lexicon Induction**

**Short Task Description.** BLI aims to find a counterpoint word or phrase in a target language. Similar to XEL, BLI can also evaluate how well a model can align a cross-lingual (entity) space.

**Task Setup.** We adopt the standard supervised embedding alignment setting (Glavaš et al., 2019) of VecMap (Artetxe et al., 2018) with 5k translation pairs reserved for training (i.e., for learning linear alignment maps) and additional 2k pairs for testing. The similarity metric is the standard cross-domain similarity local scaling (CSLS; Lample et al. 2018).\(^{12}\) We experiment with six language pairs and report accuracy (i.e., Hits@1) and mean reciprocal rank (MRR).

**Results and Discussion.** The results are provided in Tab. 4. There are accuracy gains observed on 4/6 language pairs, while MRR improves for all pairs. These findings further confirm that Prix-LM (All) + Mirror)

---

\(^{11}\)Marathi (mr), an Indo-Aryan language spoken in Western India, written in Devanagari script), Lao (lo), a Kra-Dai language written in Lao script) and Telugu (te), a Dravidian language spoken in southeastern India written in Telugu script).

\(^{12}\)Note that the models are not fine-tuned but only their embeddings are used. Further, note that the word translation pairs in the BLI test sets have < 0.001% overlap with the cross-lingual links used in Prix-LM training.
in general learns better entity representations and improved cross-lingual entity space alignments.

3.5 Prompt-based Knowledge Probing

(Short) Task Description. LM-KP (Petroni et al., 2019) queries a PLM with (typically human-designed) prompts/templates such as Dante was born in __ (the answer should be Florence). It can be viewed as a type of KB completion since the queries and answers are converted from/into KB triples: in this case, {Dante, born-in, Florence}.

Task Setup. We probe how much knowledge a PLM contains in multiple languages relying on the multilingual Language Model Analysis (mLAMA) benchmark (Kassner et al., 2021). To ensure a strictly fair comparison, we only compare XLM-R and Prix-LM. For both Prix-LM and XLM-R, we take the word with highest probability at the [Mask] token as the model’s prediction. Punctuation, stop words, and incomplete WordPieces are filtered out from the vocabulary during prediction.

Results and Discussion. Tab. 5 indicates that Prix-LM achieves better performance than XLM-R on mLAMA across all languages. We suspect that the benefits of Prix-LM training are twofold. First, multilingual knowledge is captured in the unified LM representation, which improves LM-KP as a knowledge-intensive task. The effect of this is particularly pronounced on low-resource languages such as ri, et, and hu, showing that transferring knowledge from other languages is effective. Second, the Prix-LM training on knowledge triples is essentially an adaptive fine-tuning step (Ruder, 2021) that exposes knowledge from the existing PLMs’ weights. We will discuss this conjecture, among other analyses, in what follows.

3.6 Additional Analysis

Inconsistency of the Optimal Checkpoint across Tasks (Fig. 2). How many steps should we pretrain Prix-LM on knowledge triples? The plots in Fig. 2 reveal that the trend is different on tasks that require language understanding (mLAMA) versus tasks that require only entity representations (LP and XL-BEL). On mLAMA, Prix-LM’s performance increases initially and outperforms the base model (XLM-R, at step 0). However, after around 20k steps it starts to deteriorate. We speculate that this might occur due to catastrophic forgetting, as mLAMA requires NLU capability to process queries formatted as natural language. Training on knowledge triples may expose the PLMs’ capability of generating knowledge at the earlier training stages: this explains the steep increase from 0-20k iterations. However, training on knowledge triples for (too) long degrades the model’s language understanding capability. On the other hand, longer training seems almost always beneficial for LP and XL-BEL: these tasks require only high-quality entity embeddings instead of understanding complete sentences. A nuanced difference between LP and XL-BEL is that Prix-LM’s performance on XL-BEL saturates after 100k-150k steps, while on LP the Hits@1 score still increases at 200k steps.

Link Prediction on Unseen Entities (Tab. 6). KG embedding models such as RotatE require that entities in inference must be seen in training. However, the Prix-LM is able to derive (non-random) representations also for unseen entities. We evaluate this ability of Prix-LM on triples (s, r, o) where the subject entity s or object entity o is unseen during training. The results indicate that Prix-LM can generalize well also to unseen entities.

4 Related Work

Injecting Structured Knowledge into LMs. Conceptually, our work is most related to recent
work on knowledge injection into PLMs. Know-
BERT (Peters et al., 2019) connects entities in
text and KGs via an entity linker and then re-
contextualizes BERT representations conditioned
on the KG embeddings. KG-BERT (Yao et al.,
2019) trains BERT directly on knowledge triples
by linearizing their entities and relations into a se-
quence and predicting plausibility of the sequence.
Wang et al. (2021a) improve KG-BERT by split-
ing a subject-relation-object knowledge triple into
a subject-relation pair representation and an object
entity representation, then modeling their similari-
ties with a dual/Siamese neural network.14 While
prior studies have focused on incorporating mono-
lingual (English) structured knowledge into PLMs,
our work focuses on connecting knowledge in
many languages, allowing knowledge in each lan-
guage to be transferred and collectively enriched.

Multilingual LMs pretrained via MLM, such as
mBERT (Devlin et al., 2019) and XLM-R (Con-
neau et al., 2020), cover 100+languages and are
the starting point (i.e. initialization) of Prix-LM.15
With the notable exception of Calixto et al. (2021)
who rely on the prediction of Wikipedia hyperlinks
as an auxiliary/intermediate task to improve XLM-
R’s multilingual representation space for cross-
lingual transfer, there has not been any work on aug-
menting multilingual PLMs with structured knowl-
edge. Previous work has indicated that off-the-shelf
mBERT and XLM-R fail on knowledge-intensive
multilingual NLP tasks such as entity linking and
KG completion, and especially so for low-resource
languages (Liu et al., 2021b). These are the crucial
challenges addressed in this work.

KB Completion and Construction. Before
PLMs, rule-based systems and multi-staged infor-
mation extraction pipelines were typically used
for automatic KB construction (Auer et al., 2007;
Fabian et al., 2007; Hoffart et al., 2013; Dong et al.,
2014). However, such methods require expensive
human effort for rule or feature creation (Carlson
et al., 2010; Vrandečić and Krötzsch, 2014), or they rely on (semi-)structured corpora with easy-to-
consume formats (Lehmann et al., 2015). Petroni
et al. (2019) showed that modern PLMs such as
BERT could also be used as KBs: querying PLMs
with fill-in-the-blank-style queries, a substantial
amount of factual knowledge can be extracted. This
in turn provides an efficient way to address the
challenges of traditional KB methods. Jiang et al.
(2020) and Kassner et al. (2021) extended the idea
to extracting knowledge from multilingual PLMs.

Work in monolingual settings closest to ours is
COMET (Bosselut et al., 2019): Prix-LM can be seen
as an extension of this idea to multilingual and
cross-lingual setups. Prix-LM’s crucial property is
that it enables knowledge population by transferr-
ing complementary structured knowledge across
languages. This can substantially enrich (limited)
prior knowledge also in monolingual KBs.

In another line of work, multilingual KG embed-
dings (Chen et al., 2017, 2021; Sun et al., 2020a,
2020b) were developed to support cross-KG knowl-
edge alignment and link prediction. Such methods
produce a unified embedding space that allows link
prediction in a target KG based on the aligned prior
knowledge in other KGs (Chen et al., 2020). Re-
search on multilingual KG embeddings has made
rapid progress recently, e.g., see the survey of Sun
et al. (2020b). However, these methods focus on a
closed-world scenario and are unable to lever-
age open-world knowledge from natural language
texts. Prix-LM combines the best of both worlds
and is able to capture and combine knowledge from
(multilingual) KGs and multilingual texts.

5 Conclusion
We have proposed Prix-LM, a unified multilingual
representation model that can capture, propagate
and enrich knowledge in and from multilingual
KBs. Prix-LM is trained via a casual LM object-
ive, utilizing monolingual knowledge triples and
cross-lingual links. It embeds knowledge from the
KB in different languages into a shared representa-
tion space, which benefits transferring complement-
ary knowledge between languages. We have run
comprehensive experiments on 4 tasks relevant to
KB construction, and 17 diverse languages, with
performance gains that demonstrate the effective-
ness and robustness of Prix-LM for automatic KB
construction in multilingual setups. The code and
the pretrained models will be available online at:
[URL_PLACEHOLDER].

---

14Other work on knowledge injection such as K-BERT (Liu
et al., 2020a) and ERNIE (Zhang et al., 2019) mainly aims
to leverage external knowledge to improve on downstream
NLU tasks instead of performing KG completion.

15We will explore autoregressive multilingual PLMs such as
mBART (Liu et al., 2020b) and mT5 (Xue et al., 2021) in the
future. While they adopt autoregressive training objectives at
pretraining, it is non-trivial to extract high-quality embeddings
from such encoder-decoder architectures, which is crucial for
some tasks in automatic KB completion (e.g. XEL and BLI).
References


Xu Han, Shulin Cao, Lx Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. OpenKE: An open toolkit for knowledge embedding. In EMNLP 2018.


Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021b. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *ACL-IJCNLP 2021 (findings)*.

Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat-Seng Chua. 2021c. Learning intents behind interactions with knowledge graph for recommendation. In *WWW 2021*.


Algorithm 1: Constrained Beam Search

Input: Subject entity $s$, relation $p$, set of object entities $O$, maximum entity length $L$, size of expansion set $K$, PLM vocabulary set $V$.

Output: Predicted entity.

Create the initial sequence $X_0$ by concatenating $s$ and $p$.

Create a set of sequences $X = \emptyset$.

$X_0 = \{(X_0, 0)\}$.

for $t = 1, ..., L$ do

$X_t = \emptyset$.

for $X, l \in X_{t-1}$ do

for $w \in V$ do

Add $\{(X, w), l - \log P(w|X)\}$ to $X$ and $X_t$.

Remove the sequences in $X_t$ that cannot expand to entities in $O$.

Keep at most $K$ sequences in $X_t$ with the smallest loss.

For object entities that appear in $X$, return the one with the smallest loss.

A Language Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>English</td>
</tr>
<tr>
<td>es</td>
<td>Spanish</td>
</tr>
<tr>
<td>it</td>
<td>Italian</td>
</tr>
<tr>
<td>de</td>
<td>German</td>
</tr>
<tr>
<td>fr</td>
<td>French</td>
</tr>
<tr>
<td>fi</td>
<td>Finnish</td>
</tr>
<tr>
<td>et</td>
<td>Estonian</td>
</tr>
<tr>
<td>hu</td>
<td>Hungarian</td>
</tr>
<tr>
<td>ru</td>
<td>Russian</td>
</tr>
<tr>
<td>tr</td>
<td>Turkish</td>
</tr>
<tr>
<td>ko</td>
<td>Korean</td>
</tr>
<tr>
<td>ja</td>
<td>Japanese</td>
</tr>
<tr>
<td>zh</td>
<td>Chinese</td>
</tr>
<tr>
<td>th</td>
<td>Thai</td>
</tr>
<tr>
<td>te</td>
<td>Telugu</td>
</tr>
<tr>
<td>lo</td>
<td>Lao</td>
</tr>
<tr>
<td>mr</td>
<td>Marathi</td>
</tr>
</tbody>
</table>

Table 7: Language abbreviations used in the paper.

B Constrained Beam Search Algorithm

The detailed algorithm of constrained beam search is described in Alg. 1.