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).

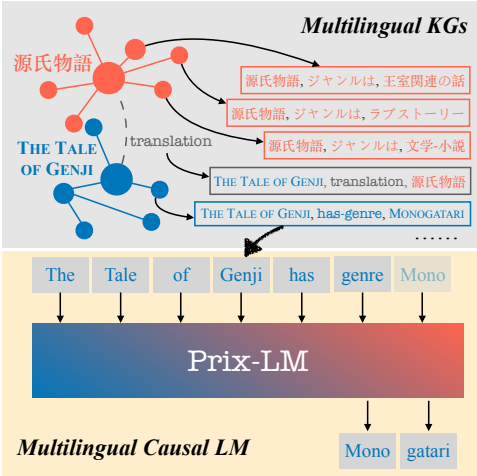

Figure 1: An illustration of the main idea supporting Prix-LM: it infuses complementary multilingual knowledge from KGs into a multilingual causal LM; e.g., Japanese KG stores more comprehensive genre information of THE TALE OF GENJI than KGs in other languages. Through cross-lingual links (translations), such knowledge is then propagated across languages.

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.

[1]This 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*.

2 `Prix-LM`

We now describe `Prix-LM`, first outlining the data structure and pretraining task, and then describing its pretraining procedure in full (§2.1), and efficient inference approaches with `Prix-LM` (§2.2).

Pretraining Task. We rely on multilingual DBpedia, but note that `Prix-LM` is also applicable to other KBs. DBpedia contains two types of structured knowledge: monolingual knowledge triples, and cross-lingual links between entities. The monolingual triples represent (relational) facts expressed in a structured manner. Each triple is denoted as $\{e_1, r, e_2\}$: the elements of a triple are identified as the subject entity e_1, relation (or predicate) r, and object entity e_2, respectively (see also Figure 1 for examples). For instance, the fact "*The capital of England is London*" can be represented as $\{\textsc{England}, \texttt{capital}, \textsc{London}\}$. The cross-lingual links, denoted as $\{e_a, e_b\}$, represent the correspondence of 'meaning-identical' entities e_a and e_b in two different languages: e.g., the English entity London is mapped to Londres in Spanish.

We treat both types of knowledge using the same input format $\{s, p, o\}$, where $s = e_1, p = r, o = e_2$ for monolingual knowledge triples, and $s = e_a, p = null, o = e_b$ for cross-lingual entity links. The pretraining task is then generating o given s and p. This objective is consistent with the link prediction task and also benefits other entity-related downstream tasks, as empirically validated later.

2.1 Pretraining Language Models

`Prix-LM` is initialized by a multilingual PLM such as XLM-R (Conneau et al., 2020): starting from XLM-R's pretrained weights, we train on the structured knowledge from a multilingual KB.

Input Representation. We represent knowledge from the KB as sequences of tokens. In particular, given some knowledge fact $\{s, p, o\}$, where each element is the surface name of an entity or a relation, we tokenize[2] the elements to sequences of subtokens X_s, X_p, and X_o. We treat each element in the knowledge fact as a different text segment and concatenate them to form a single sequence. We further introduce special tokens to represent different types of knowledge:

(1) Monolingual Triples. We use special tokens to indicate the role of each element in the triple, which

[2]XLM-R's dedicated multilingual tokenizer is used to processes entity and relation names in each language.

converts the sequence to the following format:

<s> [S] $\boldsymbol{X_s}$ </s> </s> [P] $\boldsymbol{X_p}$ </s> </s> [O] $\boldsymbol{X_o}$ [EOS]</s>.

<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> [S] $\boldsymbol{X_s}$ </s> </s> [P] [*S-LAN*] [*O-LAN*] </s> </s> [O] $\boldsymbol{X_o}$ [EOS]</s>.

<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 London to a Spanish entity Londres, 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 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:

$$\mathcal{L}_{\text{LM}} = -\sum_{x_t \in X_o \cup \{\text{[EOS]}\}} \log \text{P}(x_t \mid x_{<t}),$$

where $\text{P}(x_t \mid x_{<t})$ is the conditional probability of generating x_t given previous subtokens. The probability of generating token x_t is calculated from the hidden state of its previous token $\boldsymbol{h}_{t-1}$ in the final layer of Transformer as follows:

$$\text{P}(x_t \mid x_{<t}) = \text{softmax}(\boldsymbol{W}\boldsymbol{h}_{t-1}),$$

where $\boldsymbol{W}$ 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 $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$, we adapt them to a causal LM:

$$\text{ATT}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\intercal}{\sqrt{d}} + \boldsymbol{M}\right)\boldsymbol{V},$$

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

$$\boldsymbol{M}_{ij} = \begin{cases} 0 & x_i \notin X_o \cup \{\text{[EOS]}\} \\ 0 & x_i \in X_o \cup \{\text{[EOS]}\}, j \leq i \\ -\infty & x_i \in X_o \cup \{\text{[EOS]}\}, j > i \end{cases}$$

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 $\mathcal{O}$. A simple way to perform inference is to construct candidate tuples $\{s, p, o'\}$ using each $o' \in \mathcal{O}$ and return the one with the minimum LM loss. This straightforward approach requires encoding $|\mathcal{O}|$ sequences. However, as $|\mathcal{O}|$ 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 $-\log \mathrm{P}(w_1|X_0)$ 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 $o \in \mathcal{O}$, 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 ∞. 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 $L \cdot K \ll |\mathcal{O}|$, 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 ≥ 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.

	lang.→	EN	IT	DE	FR	FI	ET	TR	HU	JA	avg.
	# entities (K)	2175	525	304	671	187	32	159	151	422	-
	# triples (K)	7256	1543	618	1912	634	66	528	535	1159	-
Hits@1	TransE	11.3	4.1	4.8	3.0	2.4	2.6	6.1	11.4	1.9	5.3
	ComplEx	15.3	12.8	11.6	16.3	18.8	16.3	16.3	15.0	12.7	15.0
	RotatE	19.7	17.3	17.5	23.0	19.8	21.5	26.2	29.8	15.8	21.2
	Prix-LM (Single)	25.5	17.9	17.8	23.8	19.0	16.1	37.6	32.6	19.7	23.3
	Prix-LM (All)	**27.3**	**22.7**	**20.8**	**25.0**	**22.4**	**25.8**	**41.8**	**35.1**	**20.6**	**26.8**
Hits@3	TransE	28.0	25.0	24.0	27.2	26.0	20.0	31.0	36.1	20.6	26.4
	ComplEx	22.3	22.2	20.7	24.0	30.1	24.8	26.9	29.0	22.9	24.8
	RotatE	29.6	28.4	26.8	30.1	**32.8**	34.6	37.4	42.6	26.7	32.1
	Prix-LM (Single)	34.1	27.7	24.8	29.6	27.6	25.6	46.1	44.1	**29.4**	32.1
	Prix-LM (All)	**35.6**	**32.2**	**29.7**	**32.4**	31.8	**36.7**	**49.8**	**47.5**	**29.4**	**36.1**
Hits@10	TransE	41.4	42.3	38.8	43.5	**47.9**	38.3	50.3	51.0	37.9	43.5
	ComplEx	32.2	34.7	32.7	35.7	44.4	35.6	41.7	45.0	35.5	37.5
	RotatE	39.1	42.2	40.0	**44.9**	47.7	46.4	52.3	55.2	**40.0**	45.3
	Prix-LM (Single)	42.5	38.2	33.3	37.6	39.2	34.8	54.3	55.4	36.7	41.3
	Prix-LM (All)	**44.3**	**42.5**	**40.1**	40.3	44.0	**47.5**	**58.7**	**56.8**	38.0	**45.8**

Table 1: Link prediction statistics and results. The languages (see Appendix for the language codes) are ordered based on their proximity to English (e.g., IT, DE and FR being close to EN and HU and JA are distant to EN; Chiswick and Miller 2005). FI, ET, TR and HU have less than 1M Wikipedia articles and are relatively low-resource.

lang.→	TE	LO	MR	avg.
XLM-R + Mirror	2.1	4.0	0.1	2.1
mBERT + Mirror	3.2	**8.0**	0.1	3.8
Prix-LM + Mirror	**13.09**	7.6	**21.0**	**13.9**

Table 2: XEL accuracy on the LR-XEL task for low-resource languages.

are reported in Tab. 1. The evaluation metrics are standard *Hits@1*, *Hits@3*, and *Hits@10*.[7]

Models in Comparison. We refer to our model as Prix-LM (All) and compare it to the following groups of baselines. First, we compare to three representative and widely used KG embedding models[8]: 1) TransE (Bordes et al., 2013) interprets relations as translations from source to target entities, 2) ComplEx (Trouillon et al., 2016) uses complex-valued embedding to handle binary relations, while 3) RotatE (Sun et al., 2019) interprets relations as rotations from source to target entities in the complex space. In fact, RotatE additionally uses a self-adversarial sampling strategy in training, and offers state-of-the-art performance on several KG completion benchmarks (Rossi et al., 2021). Second, Prix-LM (Single) is the ablated monolingual version of Prix-LM, which uses an identical model structure to Prix-LM (All), but is trained only on monolingual knowledge triples of the test language. Training adopts the same strategy from prior work on pretraining monolingual LMs for KG completion (Bosselut et al., 2019; Yao et al., 2019). We train the Prix-LM (Single) for the same number 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.

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

[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.

lang.→	EN	ES	DE	FI	RU	TR	KO	ZH	JA	TH	avg.
XLM-R + Mirror	**75.4**	34.0	13.7	4.2	7.4	19.5	1.8	1.4	2.7	3.2	16.3
mBERT + Mirror	73.1	40.1	16.6	4.4	5.0	22.0	1.9	1.1	2.3	2.4	16.9
Prix-LM (Single) + Mirror	**75.4**	39.5	16.9	8.4	12.4	27.4	2.1	3.5	4.1	6.9	19.7
Prix-LM (All) + Mirror	71.9	**49.2**	**25.7**	**15.2**	**24.5**	**34.1**	**9.3**	**6.9**	**13.7**	**14.5**	**26.5**

Table 3: XEL Accuracy on XL-BEL.

lang.→	EN-IT		EN-TR		EN-RU		EN-FI		FI-RU		FI-TR	
model↓	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR
XLM-R + Mirror	**12.0**	16.6	**6.9**	8.6	2.9	5.9	5.9	7.4	2.0	3.3	5.7	7.0
Prix-LM + Mirror	11.5	**20.4**	6.7	**11.1**	**3.7**	**11.4**	**6.9**	**11.5**	**4.2**	**9.0**	**7.7**	**11.0**

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

lang.→	EN	IT	DE	FR	FI	ET	TR	HU	avg.
XLM-R	21.0	19.3	13.9	7.6	5.6	6.1	20.5	6.1	12.5
Prix-LM	**23.8**	**21.8**	**20.7**	**17.8**	**16.1**	**7.4**	**23.9**	**13.1**	**18.1**

Table 5: Accuracy on mLAMA.

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 counterpart 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

[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`.[13] 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 FI, 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

[13]This is a fair comparison as XLM-R and `Prix-LM` share the same tokenizer and their prediction candidate spaces are thus the same. We exclude multi-token answers as they require multi-token decoding modules, which will be different for causal LMs like `Prix-LM` versus MLMs such as XLM-R.

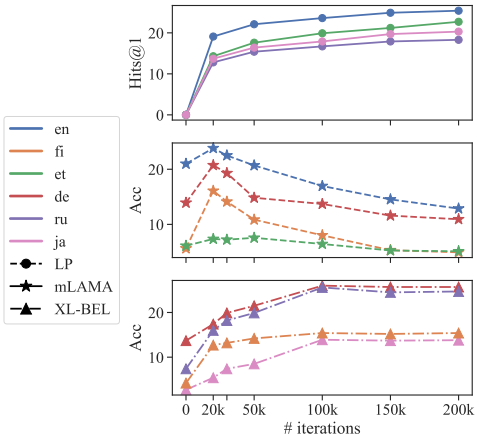

Figure 2: `Prix-LM` performance on LP, mLAMA, and XL-BEL over different checkpoints. Results of a sample of languages are shown for clarity.

lang.→	EN	IT	DE	FR	FI	ET	TR	HU	JA	avg.
Hits@1	17.2	22.9	17.0	16.0	18.3	31.3	19.2	28.5	12.4	20.3
Hits@3	24.7	30.1	24.0	22.3	23.5	37.7	24.7	38.5	19.0	27.1
Hits@10	31.0	34.9	28.9	27.8	31.9	42.3	30.8	44.2	23.6	32.8

Table 6: LP scores of `Prix-LM` (All) on unseen entities.

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. KnowBERT (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 sequence and predicting plausibility of the sequence. Wang et al. (2021a) improve KG-BERT by splitting a subject-relation-object knowledge triple into a subject-relation pair representation and an object entity representation, then modeling their similarities with a dual/Siamese neural network.[14] While prior studies have focused on incorporating monolingual (English) structured knowledge into PLMs, our work focuses on connecting knowledge in many languages, allowing knowledge in each language to be transferred and collectively enriched.

Multilingual LMs pretrained via MLM, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau 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 augmenting multilingual PLMs with structured knowledge. 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 information 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 transferring complementary structured knowledge across languages. This can substantially enrich (limited) prior knowledge also in monolingual KBs.

In another line of work, multilingual KG embeddings (Chen et al., 2017, 2021; Sun et al., 2020a, 2021) were developed to support cross-KG knowledge 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). Research 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 leverage 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 objective, utilizing monolingual knowledge triples and cross-lingual links. It embeds knowledge from the KB in different languages into a shared representation space, which benefits transferring complementary 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 effectiveness 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]`.

[14] Other 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.

[15] We 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

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *EMNLP 2017*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL 2018*, pages 789–798.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *EMNLP 2021*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS 2013*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *ACL 2019*.

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting Wikipedia hyperlinks. In *NAACL 2021*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI 2010*.

Muhao Chen, Weijia Shi, Ben Zhou, and Dan Roth. 2021. Cross-lingual entity alignment with incidental supervision. In *EACL 2021*.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI 2017*.

Xuelu Chen, Muhao Chen, Changjun Fan, Ankith Uppunda, Yizhou Sun, and Carlo Zaniolo. 2020. Multilingual knowledge graph completion via ensemble knowledge transfer. In *EMNLP 2020 (Findings)*.

Barry R Chiswick and Paul W Miller. 2005. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26(1):1–11.

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 *ACL 2020*.

Zihang Dai, Lei Li, and Wei Xu. 2016. CFO: Conditional focused neural question answering with large-scale knowledge bases. In *ACL 2016*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *AAAI 2018*.

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

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *KDD 2014*.

MS Fabian, Kasneci Gjergji, WEIKUM Gerhard, et al. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *WWW 2007*.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *ACL 2019*.

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

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

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

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *EACL 2021*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015*.

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

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021a. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *EMNLP 2021*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *ACL-IJCNLP 2021*.

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 *AAAI 2020*.

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. *TACL*, 8:726–742.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *EMNLP-IJCNLP 2019*.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *ACL 2018*.

Heiko Paulheim. 2018. How much is a triple? estimating the cost of knowledge graph creation. In *ISWC 2018*.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *EMNLP-IJCNLP 2019*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP 2019*.

Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–49.

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *WWW 2007*.

Zequn Sun, Muhao Chen, and Wei Hu. 2021. Knowing the no-match: Entity alignment with dangling cases. In *ACL-IJCNLP 2021*.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020a. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *AAAI 2020*.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020b. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 13(11):2326–2340.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge graph embedding by relational rotation in complex space. In *ICLR 2019*.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP 2015*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR.

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

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *EMNLP 2020*.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021a. Structure-augmented text representation learning for efficient knowledge graph completion. In *WWW 2021*.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW 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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP 2020: System Demonstrations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL 2021*.

Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. GraphDialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *EMNLP 2020*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *ArXiv*, abs/1909.03193.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *KDD 2016*.

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

Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *TACL*, 8:109–124.

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 $\mathcal{V}$.
Output: Predicted entity.
Create the initial sequence X_0 by concatenating s and p.
Create a set of sequences $\mathcal{X} = \emptyset$.
$\mathcal{X}_0 = \{(X_0, 0)\}$.
for $t = 1, ..., L$ **do**
  $\mathcal{X}_t = \emptyset$.
  for $X, l \in \mathcal{X}_{t-1}$ **do**
    for $w \in \mathcal{V}$ **do**
      Add $(\{X, w\}, l - \log \mathrm{P}(w_t|X))$ to $\mathcal{X}$ and $\mathcal{X}_t$.
  Remove the sequences in $\mathcal{X}_t$ that cannot expand to entities in O.
  Keep at most K sequences in $\mathcal{X}_t$ with the smallest loss.
For object entities that appear in $\mathcal{X}$, return the one with the smallest loss.

A Language Codes

EN	English
ES	Spanish
IT	Italian
DE	German
FR	French
FI	Finnish
ET	Estonian
HU	Hungarian
RU	Russian
TR	Turkish
KO	Korean
JA	Japanese
ZH	Chinese
TH	Thai
TE	Telugu
LO	Lao
MR	Marathi

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.