USC University of Southern California

UNIVERSITY OF CAMBRIDGE

Association for Computational Linguistics

ACL 2022

Prix-LM: Pretraining for Multilingual Knowledge Base Construction

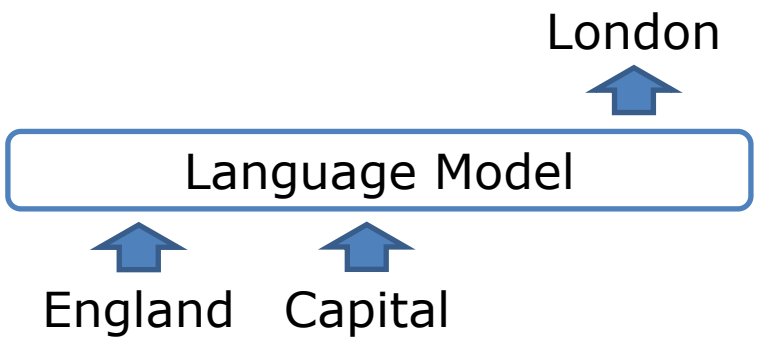
Wenxuan Zhou[1], Fangyu Liu[2], Ivan Vulić[2], Nigel Collier[2], Muhao Chen[1]

University of Southern California[1], University of Cambridge[2]

LM for Knowledge Base (KB) Construction

Knowledge triple: (England, Capital, London)

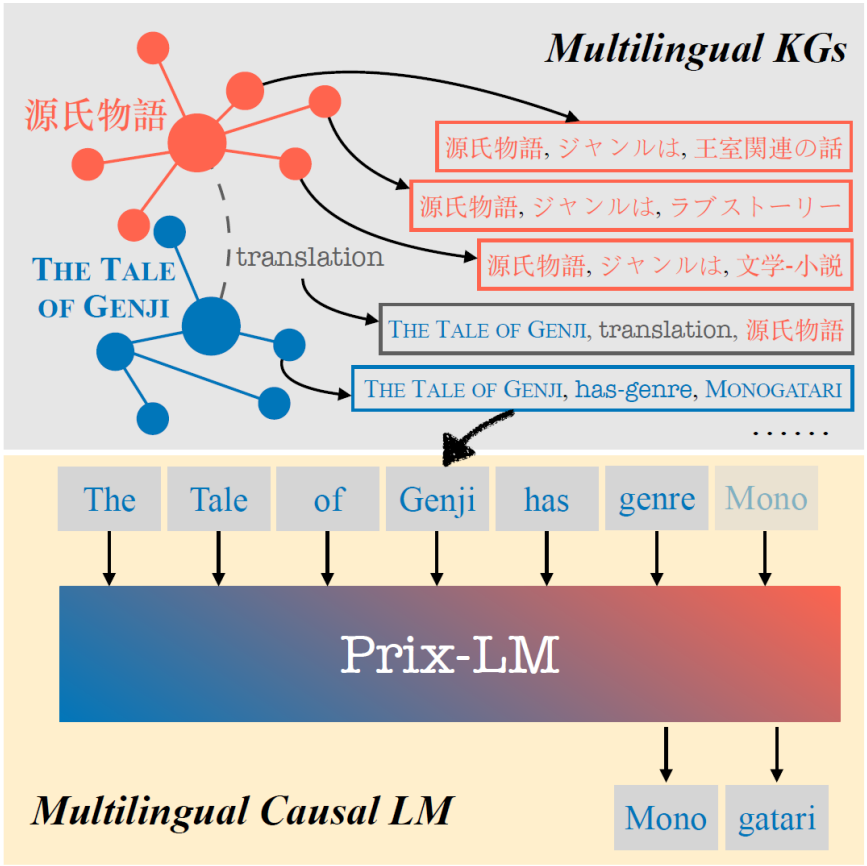
Subject entity | Predicate | Object entity



Pros of LM for KB construction:

1. A **scalable** way to represent and infer structural knowledge.
2. Can **generalize** to novel entities/relations.

Multilingual KB Construction



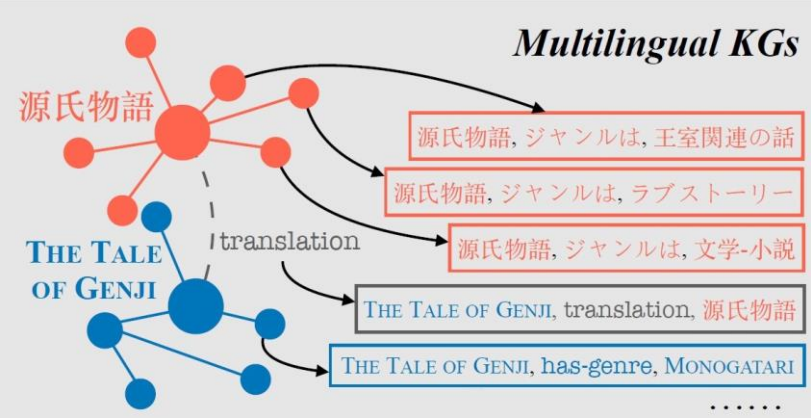
Motivation

1. KBs in different languages may contain **complementary knowledge**.

2. **Low-resource** languages may suffer severely from **missing entities/relations**.

Prix-LM: LM for multilingual KB construction and completion

Pretraining – Two Types of Knowledge



Monolingual triples: describing a fact in a single language

Cross-lingual links: identical entities/relations in two different languages

Pretraining – Input Representation

Convert structured inputs to text for language modeling.

Monolingual triples:

[CLS] [S] **subject** [SEP] [P] **predicate** [SEP] [O] **object** [EOS]

[S], [P], [O], [EOS] are **special tokens** indicating subject, predicate, object, end of sequence. [CLS], [SEP] are classification token/separators in LM.

E.g. [CLS] [S] **England** [SEP] [P] **capital** [SEP] [O] **London** [EOS]

Cross-lingual links:

[CLS] [S] **subject** [SEP] [P] **[S-LAN] [O-LAN]** [SEP] [O] **object** [EOS]

[S-LAN], [O-LAN] are **special tokens** indicating the languages of entities.

E.g. [CLS] [S] **London** [SEP] [P] **[EN] [ES]** [SEP] [O] **Londres** [EOS]

Pretraining - Model

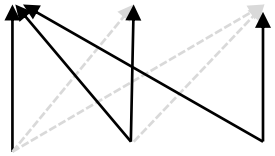
Models: pretrained masked language models (XLM-R)
LM objective: given subject and predicate, generate object.

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

Attention masks: prevent leakage of ground truth

Outputs

Los **Angeles** [EOS]

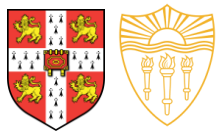


Attentions in object: should not look future tokens.

[CLS] [S] **USC** [SEP] [P] **location** [SEP] [O] **Los** **Angeles**

Inputs

Pretraining - Setup



Corpus: DBpedia in 87 languages (supported by XLM-R [Lample et al. 2019])

- 52M monolingual triples.
- 142M cross-lingual links.

Model: finetuned from XLM-R-base using our training objective.

Hyperparameters: same as XLM-R.

Inference – Autoregressive

Goal: given subject entity e_s and predicate p, determine the most probable object entity o from a (large) collection of entities.

Naïve way:
- Compute LM loss for all triples $(e_s, p, o'), o' \in E$.
- Time complexity: $|E|$

Constrained beam search:
- (High-level) idea:
  - Generate one token at a time, select from tokens that constitute entities.
  - Only keep K sequences with the smallest LM loss in the expansion set for beam search.
  - Repeat until [EOS] is generated or hit the maximum length.
- Time complexity: max length $\times K$

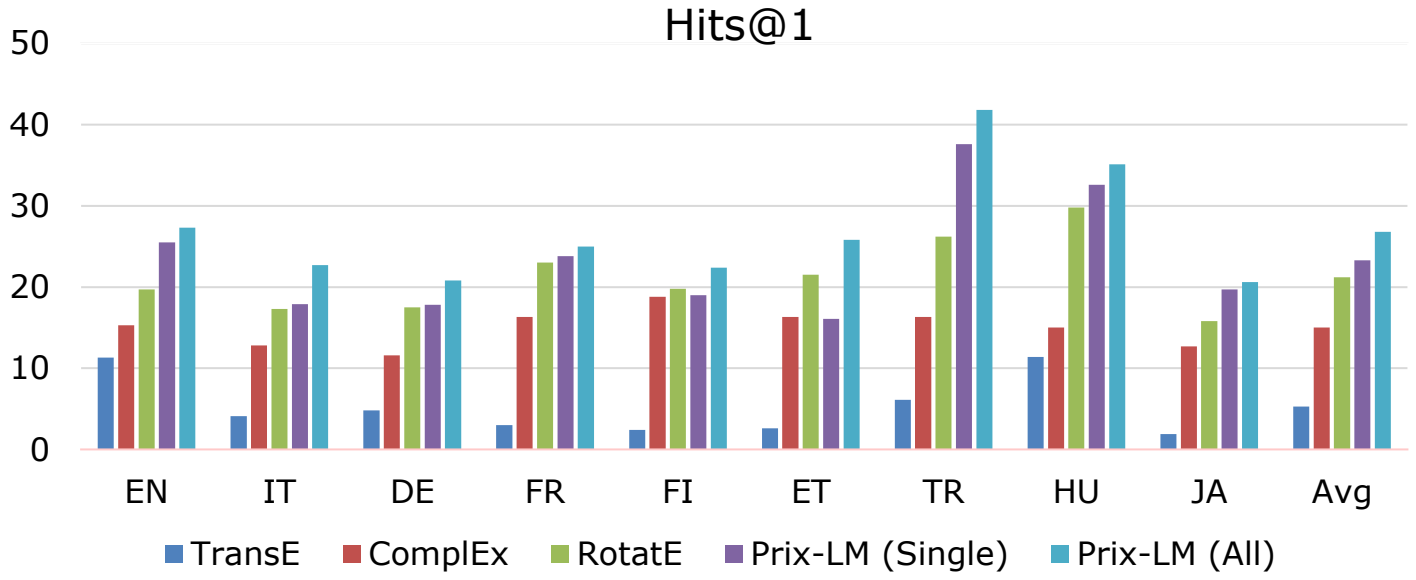
Inference – Similarity-based

Goal: retrieve the nearest neighbors using the embedding similarity.

Method: refine the [CLS] embedding using Mirror-BERT.

Experiments - Link prediction

Task: Given subject s, predicate p, and a large collection of entities E, determine the object entity $o \in E$.
Data: DBpedia, randomly reserve 10% triples as test set.

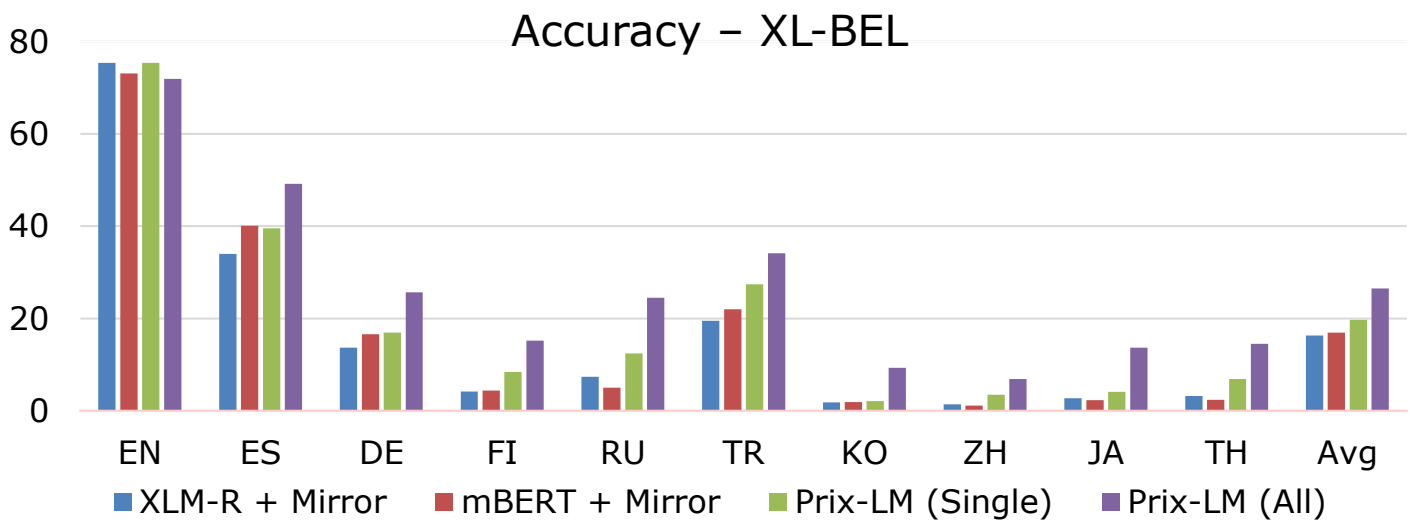


Observations:

1. Prix-LM (all) consistently outperforms other baselines.
2. Multilingual Prix-LM outperforms the monolingual one.
3. Results of Hits@3 and Hits@10 follow similar trends.

Experiments – Cross-lingual Entity Linking

Task: Link entity mentions in different languages.
Data: XL-BEL, LR-BEL



Prix-LM (all) consistently outperforms other baselines.

We also did experiments on tasks including **Bilingual Lexicon Induction**, **Prompt-based Knowledge Probing**, and **Link Prediction on Unseen Entities**. Please refer to our paper for details.

Conclusion

1. We propose Prix-LM, a **unified multilingual representation model** that can capture, propagate and enrich knowledge in and from multilingual KBs.

2. Prix-LM embeds knowledge from the KB in different languages into a shared representation space, which benefits transferring **complementary knowledge** between languages.

3. Experiments on 4 tasks demonstrate the **effectiveness and robustness** of Prix-LM for automatic KB construction in multilingual setups.

Paper & Code