

DKPLM: Decomposable Knowledge-enhanced Pre-trained Language Model for Natural Language Understanding

Taolin Zhang^{1,3*}, Chengyu Wang^{2*}, Nan Hu⁵, Minghui Qiu^{2†}, Chengguang Tang²
Xiaofeng He^{4,5†}, Jun Huang²

¹ School of Software Engineering, East China Normal University ² Alibaba Group

³ Shanghai Key Laboratory of Trustworthy Computing

⁴ NPPA Key Laboratory of Publishing Integration Development, ECNUP

⁵ School of Computer Science and Technology, East China Normal University
zhangtl0519@gmail.com, hunan.vinny1997@gmail.com, hexf@cs.ecnu.edu.cn
{chengyu.wcy, minghui.qmh, chengguang.tcg, huangjun.hj}@alibaba-inc.com

Abstract

Knowledge-Enhanced Pre-trained Language Models (KE-PLMs) are pre-trained models with relation triples injecting from knowledge graphs to improve language understanding abilities. To guarantee effective knowledge injection, previous studies integrate models with knowledge encoders for representing knowledge retrieved from knowledge graphs. The operations for knowledge retrieval and encoding bring significant computational burdens, restricting the usage of such models in real-world applications that require high inference speed. In this paper, we propose a novel KEPLM named DKPLM that **D**ecomposes **K**nowledge injection process of the **P**re-trained **L**anguage **M**odels in pre-training, fine-tuning and inference stages, which facilitates the applications of KEPLMs in real-world scenarios. Specifically, we first detect knowledge-aware long-tail entities as the target for knowledge injection, enhancing the KEPLMs' semantic understanding abilities and avoiding injecting redundant information. The embeddings of long-tail entities are replaced by "pseudo token representations" formed by relevant knowledge triples. We further design the relational knowledge decoding task for pre-training to force the models to truly understand the injected knowledge by relation triple reconstruction. Experiments show that our model outperforms other KEPLMs significantly over zero-shot knowledge probing tasks and multiple knowledge-aware language understanding tasks. We further show that DKPLM has a higher inference speed than other competing models due to the decomposing mechanism.

Introduction

Recently, Pre-trained Language Models (PLMs) improve various downstream NLP tasks significantly (He et al. 2020; Xu et al. 2021; Chang et al. 2021). In PLMs, the two-stage strategy (i.e., pre-training and fine-tuning) (Devlin et al. 2019) inherits the knowledge learned during pre-training and applies it to downstream tasks. Although PLMs have stored a lot of internal knowledge, it can hardly understand external background knowledge such as factual and commonsense

knowledge (Colon-Hernandez et al. 2021; Cui et al. 2021). Hence, the performance of PLMs can be improved by injecting external knowledge triples, which are referred to as Knowledge-Enhanced PLMs (KEPLMs).

In the literature, the approaches of injecting knowledge can be divided into two categories, including knowledge embedding and joint learning. (1) Knowledge embedding based approaches inject triple representations in Knowledge Graphs (KGs) trained by knowledge embedding algorithms (e.g., TransE (Bordes et al. 2013)) into contextual representations via well-designed feature fusion modules, which may contain a large number of parameters (Zhang et al. 2019; Peters et al. 2019; Su et al. 2020). As reported in (Wang et al. 2019b), different knowledge representation algorithms significantly impact the performance of PLMs. (2) Joint learning based approaches learn knowledge embeddings from KGs jointly during pre-training (Wang et al. 2019b; Sun et al. 2020; Liu et al. 2020), which are two significantly different tasks. We also observe that there are two potential drawbacks of previous methods. (1) These models inject knowledge indiscriminately into all entities in pre-training sentences, which introduces redundant and irrelevant information to PLMs (Zhang et al. 2021a). (2) Large-scale KGs are required during both fine-tuning and inference for obtaining outputs of knowledge encoders. This incurs additional computation burden that limits their usage for real-world applications that require high inference speed (Zhang et al. 2020; Malik et al. 2021).

To overcome the above problems, we present a novel KE-PLM named DKPLM that decomposes knowledge injection process of three stages for KEPLMs. A comparison between our model and other models is shown in Figure 1. Clearly, for DKPLM, knowledge injection is only applied during pre-training, without using additional knowledge encoders. Hence, during the fine-tuning and inference stages, our model can be utilized in the same way as that of BERT (Devlin et al. 2019) and other plain PLMs, which facilitates the applications of our KEPLM in real-world scenarios. Specifically, we introduce three novel techniques for pre-training DKPLM:

- *Knowledge-aware Long-tail Entity Detection*: our model detects long-tail entities for knowledge injection based

* T. Zhang and C. Wang contributed equally to this work.

† Co-corresponding authors.

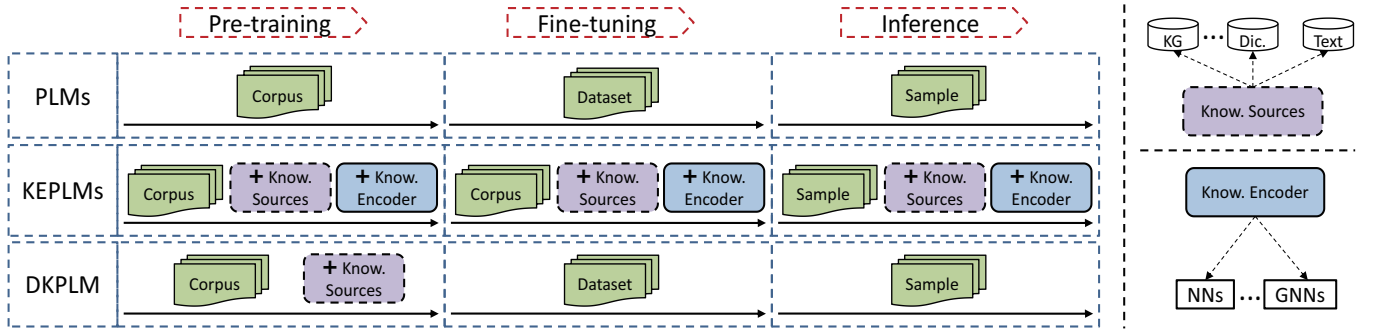


Figure 1: Comparison between DKPLM and other models. (1) Plain PLMs do not utilize external knowledge in all the three stages. (2) Existing KEPLMs utilize various knowledge sources (e.g., KGs and dictionaries) to enhance understanding abilities by using knowledge encoders in all the three stages. (3) During pre-training, DKPLM utilizes the same data sources as KEPLMs, with no knowledge encoder (e.g., Neural Networks and Graph Neural Networks) required. During fine-tuning and inference, our model does not require KGs and is highly flexible and efficient. (Best viewed in color.)

on the frequencies in the corpus, the number of adjacent entities in the KG and the semantic importance. In this way, we avoid learning too much redundant and irrelevant information (Zhang et al. 2021a).

- *Pseudo Token Representation Injection*: we replace the embeddings of detected long-tail entities with the representations of the corresponding knowledge triples generated by the shared PLM encoder, referred to “pseudo token representations”. Hence, the knowledge is injected without introducing any extra parameters to the model.
- *Relational Knowledge Decoding*: for a relation triple, we use the representations of one entity and the relation predicate to decode each token of another entity. This pre-training task acts as a supervised signal for the KEPLM, forcing the model to understand what knowledge is injected to the KEPLM.

In the experiments, we evaluate our model against strong baseline KEPLMs pre-trained using the same data sources over various knowledge-related tasks, including knowledge probing (LAMA) (Petroni et al. 2019), relation extraction and entity typing. For knowledge probing, the top-1 accuracy of four datasets is increased by +1.57% on average, compared with state-of-the-art. Meanwhile, in other tasks, our model also achieves consistent improvement. In summary, we make the following contributions in this paper:

- We present a novel KEPLM named DKPLM to inject the knowledge into PLMs, which specifically focuses on long-tail entities, decomposing the knowledge injection process of three PLMs’ stages.
- A dual knowledge injection process including encoding and decoding for long-tail entities is proposed to pre-train DKPLM, consisting of three modules: Knowledge-aware Long-tail Entity Detection, Pseudo Token Embedding Injection, and Relational Knowledge Decoding.
- In the experiments, we evaluate DKPLM over multiple public benchmark datasets, including knowledge probing (LAMA) and knowledge-aware tasks. Experimental results show that DKPLM consistently outperforms state-

of-the-art methods. An analysis on inference speed of KEPLMs is also provided.

Related Work

In this section, we briefly summarize the related work on the following two aspects: PLMs and KEPLMs.

PLMs. Recently, a variety of PLMs have been proposed to learn contextual representations. BERT (Devlin et al. 2019) (as well as its robustly optimized version RoBERTa (Liu et al. 2019b)) is the most representative work. Following BERT, many PLMs have been proposed to further improve performance in various NLP tasks. We summarize the recent studies, specifically focusing on three techniques, including self-supervised pre-training, model architectures and multi-task learning. To improve the model’s semantic understanding, several approaches extend BERT by employing novel token-level and sentence-level pre-training tasks. Notable PLMs include Baidu-ERNIE (Sun et al. 2019), StructBERT (Wang et al. 2020) and spanBERT (Joshi et al. 2020). Other models boost the performance by changing the internal encoder architectures. For example, XLNet (Yang et al. 2019) utilizes Transformer-XL (Dai et al. 2019) to encode long sequences by permutation in language tokens. Sparse self-attention (Cui et al. 2019) replaces the self-attention mechanism with more interpretable attention units. Yet other PLMs such as MT-DNN (Liu et al. 2019a) combine self-supervised pre-training with supervised learning to improve the performance of various GLUE tasks (Wang et al. 2019a).

KEPLMs. As plain PLMs are only pre-trained on large-scale unstructured corpora, they lack the language understanding abilities of important entities. Hence, KEPLMs use structured knowledge to enhance the language understanding abilities of PLMs. We summarize recent KEPLMs grouped into the following three types:

- (1) Knowledge-enhancement by entity embeddings. For example, ERNIE-THU (Zhang et al. 2019) injects entity embeddings into contextual representations via knowledge-encoders stacked by the information fusion module. KnowBERT (Peters et al. 2019) introduces the knowledge attention and recon-

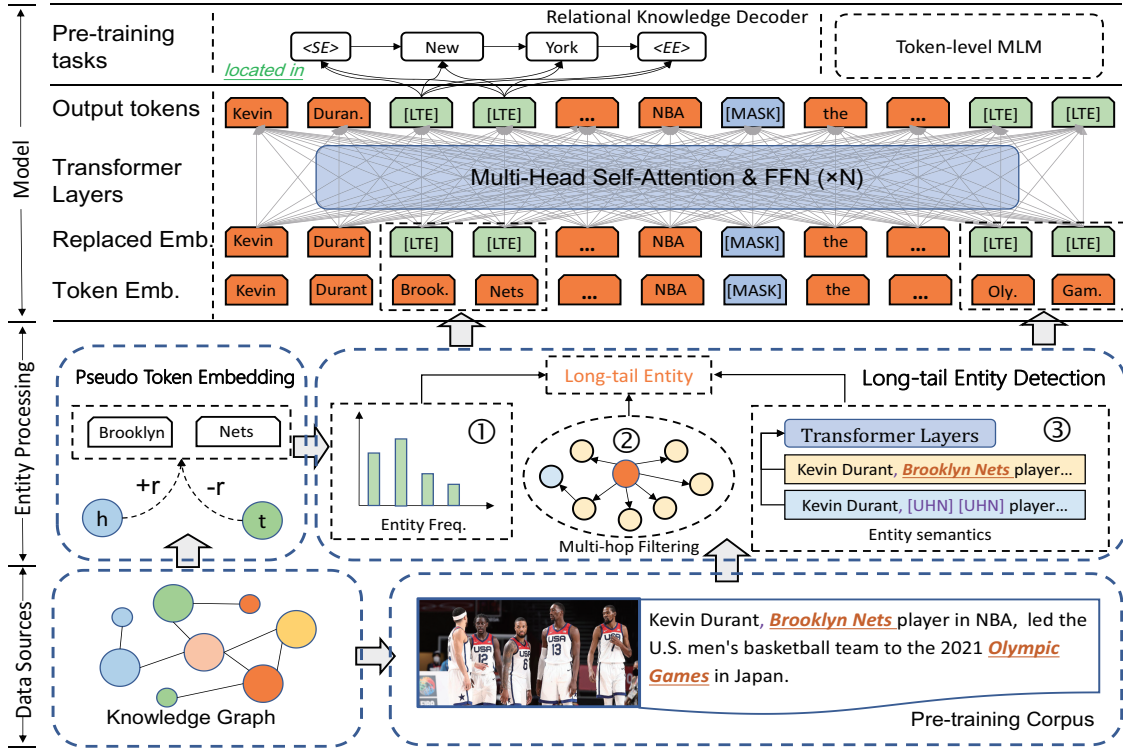


Figure 2: Overview of DKPLM. (1) **Data Sources**: large-scale pre-training corpora and relation triplets extracted from KGs. (2) **Input**: DKPLM detects long-tail entities and retrieves relation triples for learning “pseudo token embeddings”. (3) **Model**: Plain PLMs can be used as model backbones. We also propose the relational knowledge decoder for better knowledge injection. (Best viewed in color.)

textualization (KAR) and entity linking to inject knowledge embeddings to PLMs.

(2) Knowledge-enhancement by entity descriptions. These studies learn entity embeddings by knowledge descriptions. For example, pre-training corpora and entity descriptions in KEPLER (Wang et al. 2019b) are encoded into a unified semantic space within the same PLM.

(3) Knowledge-enhancement by converted triplet’s texts. K-BERT (Liu et al. 2020) and CoLAKE (Sun et al. 2020) convert relation triplets into texts and insert them into training samples without using pre-trained embeddings.

These KEPLMs require knowledge encoder modules with additional parameters to inject the knowledge into context-aware hidden representations generated by PLMs. Previous studies (Petroni et al. 2019; Broscheit 2019; Wang, Liu, and Song 2020; Cao et al. 2021) have also shown that the semantics of high-frequency and general knowledge triples are already captured by plain PLMs, and express redundant knowledge. In this paper, we argue that enhancing the understanding ability of long-tail entities can further benefit the context-aware representations of PLMs, which is one of the major focus of this work.

DKPLM: The Proposed Model

We first state some basic notations. Denote an input sequence of tokens as $\{w_1, w_2, \dots, w_n\}$, where n is the length of

the input sequence. The hidden representation of input tokens obtained by PLMs is denoted as $\{h_1, h_2, \dots, h_n\}$ and $h_i \in \mathbb{R}^{d_1}$, where d_1 is the dimension of the PLM’s output. Furthermore, we denote the knowledge graph as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ where \mathcal{E} and \mathcal{R} are the collections of entities and relation triples, respectively. In the KG, a relational knowledge triple is denoted as (e_h, r, e_t) , where e_h and e_t refer to the head entity and the tail entity, respectively. r is the specific relation predicate between e_h and e_t .

Our model DKPLM can be regarded as a knowledge-enhanced extension to a variety of PLMs, such as the encoder-based PLMs (e.g, BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b)) and auto-regressive PLMs (e.g, GPT (Radford et al. 2018)). In our work, we implement DKPLM based on the model backbone of RoBERTa (Liu et al. 2019b). Nonetheless, DKPLM can be applied to other similar PLM backbones as well, which will be left as future work.

As we wish to decompose the process of knowledge injection during the three PLMs’ stages, without using any additional knowledge encoders for knowledge representation learning, our pre-training process specifically focuses on the knowledge injection and decoding on certain tokens in the detected long-tail entities. Specifically, we aim to solve three research questions:

- **RQ1**: What types of tokens in the pre-training corpus should be detected for knowledge injection ?

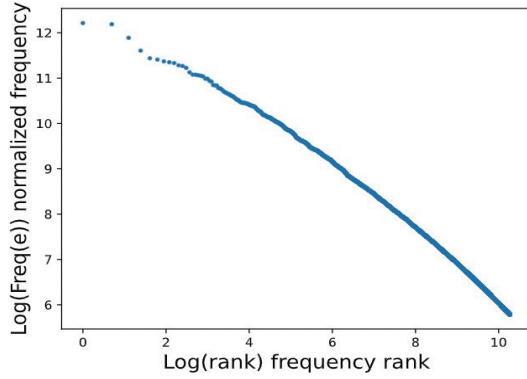


Figure 3: The distribution of entity frequencies in the English Wikipedia corpus.

- **RQ2:** How can we inject knowledge to selected tokens without additional knowledge encoders ?
- **RQ3:** How can we verify the effectiveness of injected relation triples during pre-training ?

The overall framework of DKPLM is presented in Figure 2. In the following, we introduce the techniques of DKPLM and discuss how we can address the three research questions.

Knowledge-aware Long-tail Entity Detection

Motivation and Analysis. We first extract structured knowledge triples from large-scale KGs, and link entities in KGs to the target mentions in the pre-training samples by entity linking tools (e.g., TAGME (Ferragina and Scaiella 2010)). For a better understanding of how entities are distributed in the corpus, we plot the distribution of the entity frequencies in the entire Wikipedia corpus, shown in Figure 3. As seen, it closely follows the *power-law distribution* with the formula as follows:

$$Freq(e) = \frac{C}{rank^\alpha}, \quad (1)$$

where C and α are hyper-parameters, $rank$ is the entity frequency rank and $Freq(e)$ is the frequency of the entity e . We can see that, while a few entities frequently appear, most of the entities seldom occur in the pre-training corpus, making it difficult for PLMs to learn better contextual representations.

As reported by (Zhang et al. 2021a), the high-frequency relational triples are injected into PLMs is NOT always beneficial for downstream tasks. This practice is more likely to trigger negative knowledge infusion. Hence, knowledge injection for long-tail entities instead of all entities that occur in the corpus may further improve the understanding abilities of PLMs. It should be further noted that the above analysis of entities only considers the frequencies in the pre-training corpus, ignoring the information of each entity in KGs and the importance of such entities in a sentence. In the following, we present the *Knowledge-aware Long-tail Entity Detection* mechanism to select target entities for knowledge injection.

Our Approach. In our work, we consider three neighboring information of entities to characterize the “long-tailness” property of entities, namely the entire pre-training corpus,

current input sentence and the KG. For a specific entity e , we consider the three following factors:

- **Entity Frequency:** the entity frequency w.r.t. the entire pre-training corpus, denoted as $Freq(e)$;
- **Semantic Importance:** the importance of the entity e in the sentence, denoted as $SI(e)$;
- **Knowledge Connectivity:** the number of multi-hops neighboring nodes w.r.t. the entity e in the KG, denoted as $KC(e)$.

While the computation of $Freq(e)$ is quite straightforward, it is necessary to elaborate the computation of $SI(e)$ and $KC(e)$. $SI(e)$ refers to the semantic similarity between the representation of the sentence containing the entity e and the representation of the sentence with e being replaced. The greater the similarity between sentences is, the smaller the influence on the sentence semantics is when the entity is replaced. Denote h_o and h_{rep} as the representations of the original sentence and the sentence after entity replacing. For simplicity, we use the reciprocal of cosine similarity to measure $SI(e)$:

$$SI(e) = \frac{\|h_o^T\| \cdot \|h_{rep}\|}{h_o^T \cdot h_{rep}} \quad (2)$$

In the implementation, we use the special token “[UHN]” to replace the entity e in the sentence.

$KC(e)$ represents the importance of entity e in triples’ neighboring structure and we use the multi-hops neighboring entities’ number to calculate $KC(e)$:

$$KC(e) = [\mathcal{N}(e)]_{R_{min}}^{R_{max}} \quad (3)$$

$$\mathcal{N}(e) \triangleq \{e' \mid \text{Hop}(e', e) < R_{hop} \wedge e' \in \mathcal{E}\} \quad (4)$$

where R_{min} and R_{max} are pre-defined thresholds. Specifically, we constrain the number of hops for computing $KC(e)$ is in the range of R_{min} to R_{max} . $|\cdot|$ means the neighboring entity number in the set. The Hop function denotes the number of multi-hops between entity e and entity e' in KGs’ structure. The degree of the “knowledge-aware long-tailness” $KLT(e)$ of the entity e is then calculated as:

$$KLT(e) = \mathbb{I}_{\{Freq(e) < R_{freq}\}} \cdot SI(e) \cdot KC(e) \quad (5)$$

where the $\mathbb{I}_{\{x\}}$ is the indicator function with x to be a Boolean expression. R_{freq} is a pre-defined threshold. Given a sentence, we detect all the entities and regard the entities with the $KLT(e)$ score lower than the average as knowledge-aware long-tail entities.

Pseudo Token Embedding Injection

In order to enhance the PLMs’ understanding abilities of long-tail entities, we inject knowledge triples into the positions of such entities without introducing any other parameters. Inspired by the KG embedding algorithms (Bordes et al. 2013), if an entity in the pre-training sentence is a head entity e_h of knowledge triples, the representation of e_h is modeled by the following function:

$$h_{e_h} = h_{e_t} - h_r \quad (6)$$

where h_{e_h} , h_{e_t} and h_r are the representations of the head entity e_h , the tail entity e_t and the relation predicate r , respectively. Similarly, if an entity is a tail entity e_t in the KG, we have: $h_{e_t} = h_{e_h} + h_r$.

Specifically, we use the underlying PLM as the shared encoder to acquire the knowledge representations. Consider the situation where the entity is a head entity e_h in the KG. We concatenate the tokens of the tail entity e_t and the relation predicate r , and feed them to the PLM. The token representations of the last layer of the PLM are denoted as $\mathcal{F}(e_t)$ and $\mathcal{F}(r)$, respectively.

The pseudo token representations h_{e_t} and h_r are then computed as follows:

$$h_{e_t} = \mathcal{LN}(\sigma(f_{sp}(\mathcal{F}(e_t)) W_{e_t})) \quad (7)$$

$$h_r = \mathcal{LN}(\sigma(f_{sp}(\mathcal{F}(r)) W_r)) \quad (8)$$

where \mathcal{LN} is the LayerNorm function (Ba, Kiros, and Hinton 2016) and f_{sp} is the self-attentive pooling operator (Lin et al. 2017) to generate the span representations. W_{e_t} and W_r are trainable parameters.

Since the lengths of the entity and the relation predicate are usually short, the generated representations by the PLM may be not expressive. We further consider the description text of the target entity, denoted as e_h^{des} . Let $\mathcal{F}(e_h^{des})$ be the token sequence representations of e_h^{des} , generated by the PLM. We denote the pseudo token embedding h_{e_h} of the head entity e_h as follows:

$$h_{e_h} = \tanh((h_{e_t} - h_r) \oplus \mathcal{F}(e_h^{des})) W_{e_h}, \quad (9)$$

where \oplus refers to the concatenation of two representations, and W_{e_h} is the trainable parameter.

Finally, we replace the representations of detected long-tail entities with the pseudo token representations in the PLM's embedding layer (either h_{e_h} or h_{e_t} , depending on whether the target entity is the head or the tail entity in the KG). This follows the successive multiple transformer encoder layers to incorporate the knowledge into the contextual representations without introducing any other new parameters for knowledge encoding.

Relational Knowledge Decoding

After the information of the relation triples has been injected into the model, it is not clear whether the model has understood the injected knowledge. We design a relational knowledge decoder, forcing our model to understand the injected knowledge explicitly. Specifically, we employ a self-attention pooling mechanism to obtain the masked entity span representations in the last layer:

$$h_{e_h}^o = \mathcal{LN}(\sigma(f_{sp}(\mathcal{F}(h_{e_h})) W_{e_h}^o)) \quad (10)$$

where $W_{e_h}^o$ is the learnable parameter. Given the output representation of the head entity $h_{e_h}^o$ and the relation predicate h_r , we aim to decode the tail entity.¹ Let h_d^i be the representation of the i -th token of the predicted tail entity. We have:

$$h_d^i = \tanh(\delta_d h_{e_h} h_r h_d^{i-1} \cdot W_d), \quad (11)$$

¹If the target entity is the tail entity, we can also decode the head entity in a similar fashion.

where δ_d is a scaling factor, and h_d^0 equals to $h_{e_h}^o$ as the initialization heuristics.

Because the vocabulary size is relatively large, we use the Sampled SoftMax function (Jean et al. 2015) to compare the prediction results against the ground truth. The token-level loss function \mathcal{L}_{d_i} is defined as follows:

$$\mathcal{L}_{d_i} = \frac{\exp(f_s(h_d^i, y_i))}{\exp(f_s(h_d^i, y_i)) + N \mathbb{E}_{t_n \sim Q(y_n | y_i)} [\exp(f_s(h_d^i, y_n))]} \quad (12)$$

$$f_s(h_d^i, y_i) = (h_d^i)^T \cdot y_i - \log(Q(t | t_i)) \quad (13)$$

where y_i is the ground-truth token and y_n is the negative token sampled in $Q(t_n | t_i)$. $Q(\cdot | \cdot)$ is the negative sampling function (to be described in the experiments). N is the number of negative samples. In our DKPLM model, the training objectives include two pre-training tasks: (1) relational knowledge decoding and (2) token-level Masked Language Modeling (MLM), as proposed in (Devlin et al. 2019). Hence, the total loss function of DKPLM can be denoted as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{MLM}} + (1 - \lambda_1) \mathcal{L}_{\text{De}} \quad (14)$$

where the λ_1 is the hyper-parameter and \mathcal{L}_{De} is the total decoder loss of the target entity consisting of multiple tokens.

Experiments

Pre-training Data and Model Settings

In this paper, we use English Wikipedia (2020/03/01)² as our pre-training data source, and WikiExtractor³ to process the downloaded Wikipedia dump, similar to CoLAKE (Sun et al. 2020) and ERNIE-THU (Zhang et al. 2019). We use Wikipedia anchors to align the entities in the pre-training texts recognized by entity linking tools (e.g., TAGME (Ferragina and Scialla 2010)) to WikiData5M (Wang et al. 2019b), which is a large-scale proposed KG data source including relation triples and entity description texts. The additional pre-processing and filtration steps are kept the same as ERNIE-THU (Zhang et al. 2019). In total, we have 3,085,345 entities and 822 relation types in the KG and 26 million training samples in our pre-training corpus.

We use RoBERTa-base (Liu et al. 2019b) as our model backbone. Due to the large number of entities in the KG, we extract negative entities by first extracting all entities from the specific relation and employing the PEPR algorithm (Zhang et al. 2021b) to assign scores and select the top- N entities as the results. In this paper, we set the N is 20. We set the λ_1 to $\{0.4, 0.5, 0.6\}$ and find the 0.5 is the best. The other hyper-parameters are the same as CoLAKE (Sun et al. 2020) and RoBERTa-base. All the models are implemented in PyTorch and trained using 8 V100-16G GPUs for 12 hours.

Baselines

We consider the following models as strong baselines:

ERNIE-THU (Zhang et al. 2019): The model integrates a denoising entity auto-encoder pre-training task to inject knowledge embeddings into language representations.

²<https://dumps.wikimedia.org/enwiki/>

³<https://github.com/attardi/wikiextractor>

Datasets	PLMs			KEPLMs				
	ELMo	BERT	RoBERTa	CoLAKE	K-Adapter *	KEPLER	DKPLM	Δ
Google-RE	2.2%	11.4%	5.3%	9.5%	7.0%	7.3%	10.8%	+1.3%
UHN-Google-RE	2.3%	5.7%	2.2%	4.9%	3.7%	4.1%	5.4%	+0.5%
T-REx	0.2%	32.5%	24.7%	28.8%	29.1%	24.6%	32.0%	+2.9%
UHN-T-REx	0.2%	23.3%	17.0%	20.4%	23.0%	17.1%	22.9%	-0.1%

Table 1: The performance on knowledge probing datasets. Δ represents an improvement over the best results of existing KEPLMs compared to our model. Besides, K-Adapter * is based on RoBERTa-large and uses a subset of T-REx as its training data, which may contribute to its superiority over the other KEPLMs and is unfair for DKPLM to be compared against.

KnowBERT (Peters et al. 2019): It injects rich structured knowledge representations via knowledge attention and re-contextualization.

KEPLER (Wang et al. 2019b): The model encodes texts and entities into a unified semantic space with the same PLM as the shared encoder.

CoLAKE (Sun et al. 2020): It considers a unified heterogeneous KG as the knowledge source and employs adjacency matrices to control the information flow.

K-Adapter (Wang et al. 2021): It uses different representations for different types of knowledge via neural adapters.

Knowledge Probing

The knowledge probing tasks, called LAMA (LAnguage Model Analysis) (Petroni et al. 2019), aim to measure whether the factual knowledge is stored in PLMs via cloze-style tasks. The LAMA-UHN tasks (Pörner, Waltinger, and Schütze 2019) are proposed to alleviate the problem of overly relying on the surface form of entity names, and are constructed by filtering out the easy-to-answer samples. These two tasks are evaluated under the zero-shot setting without fine-tuning, which is a fair comparison of the knowledge understanding abilities of KEPLMs. We report the macro-averaged mean precision (P@1) of DKPLM.

The performance of LAMA and LAMA-UHN tasks is summarized in Table 1. Compared to the results of other baselines, we can draw the following conclusions. (1) BERT outperforms RoBERTa by a large gap (+5.93% on average) because its vocabulary size is much smaller than RoBERTa. (2) Although our model is trained on RoBERTa-base, it achieves state-of-the-art results over three datasets (+1.57% on average). The result of our model is only 0.1% lower than K-Adapter, without using any T-REx training data and large PLM backbone. From the overall results, we can see that our learning process based on long-tail entities can effectively store and understand factual knowledge from KGs.

Knowledge-Aware Tasks

We evaluate our DKPLM model over the knowledge-aware tasks, including relation extraction and entity typing.

Entity Typing: Unlike Named Entity Recognition (Jiang et al. 2021; Shen et al. 2021), entity typing requires the model to predict fine-grained entity types in given contexts. We fine-tune our DKPLM model over Open Entity (Choi et al. 2018). Table 2 shows the performance of various models including

Model	Precision	Recall	F1
UFET (Choi et al. 2018)	77.4	60.6	68.0
BERT	76.4	71.0	73.6
RoBERTa	77.4	73.6	75.4
ERNIE _{BERT}	78.4	72.9	75.6
ERNIE _{RoBERTa}	80.3	70.2	74.9
KnowBERT _{BERT}	77.9	71.2	74.4
KnowBERT _{RoBERTa}	78.7	72.7	75.6
KEPLER _{Wiki}	77.8	74.6	76.2
CoLAKE	77.0	75.7	76.4
DKPLM	79.2	75.9	77.5

Table 2: The performance of models on Open Entity (%).

Model	Precision	Recall	F1
CNN	70.30	54.20	61.20
PA-LSTM (Zhang et al. 2017)	65.70	64.50	65.10
C-GCN (Zhang and Qi 2018)	69.90	63.30	66.40
BERT	67.23	64.81	66.00
RoBERTa	70.80	69.60	70.20
ERNIE _{BERT}	70.01	66.14	68.09
KnowBERT	71.62	71.49	71.53
DKPLM	72.61	73.53	73.07

Table 3: The performance of models on TACRED (%).

PLMs, KEPLMs and other task-specific models. From the results, we can observe: the KEPLMs outperform task-specific models and the plain PLMs. In addition, our DKPLM model with injected long-tail entity knowledge achieves a large performance gain compared to baselines (+2.2% Precision, +0.2% Recall and +1.1% F1).

Relation Extraction: The Relation Extraction task (RE) aims to determine the fine-grained semantic relation between the two entities in a given sentence. We use a benchmark RE dataset TACRED (Zhang et al. 2017) to evaluate our model’s performance. The relation types in TACRED is 42 and we adopt the micro averaged metrics and macro averaged metrics for evaluation. As shown in Table 3, the performance of knowledge-injected models are much higher, and our model achieves new state-of-the-art performance (+1.46% F1), which implies injecting long-tail entities’ knowledge triple into PLMs for RE is also very effective.

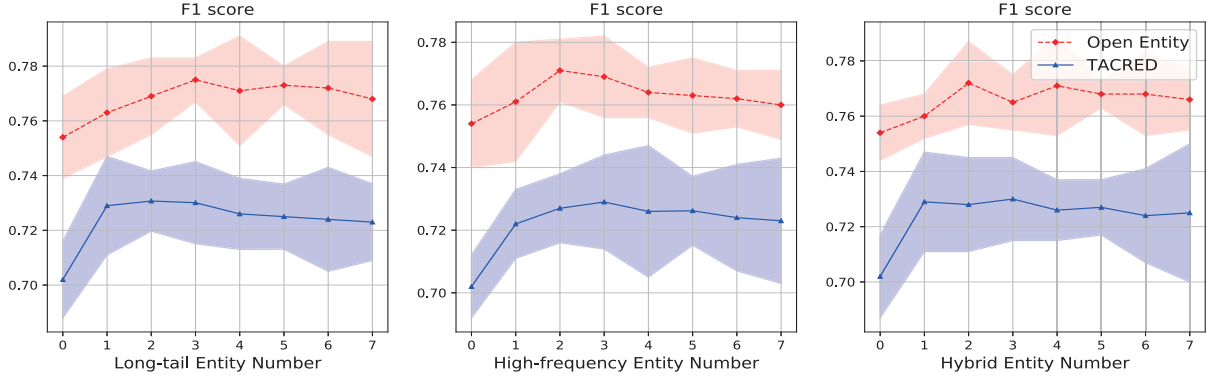


Figure 4: The influence of different injected numbers of long-tail entities and high-frequency entities.

Model	Pre-training	Fine-tuning	Inference
RoBETa_base	9.60	7.09	1.55
BERT_base	8.46	6.76	0.97
ERNIE-THU	14.71	8.19	1.95
KEPLER	18.12	7.53	1.86
CoLAKE	12.46	8.02	1.91
DKPLM	10.02	7.16	1.61

Table 4: The running time (s) in the three stages of various PLMs and KEPLMs over 1000 random samples.

Analysis of Running Time

In this section, we compare DKPLM with other models on pre-training, fine-tuning and inference time. Specifically, we choose 1000 samples randomly from the pre-training corpus and two knowledge-aware tasks. The fine-tuning and inference time is the average time of the two datasets, respectively.

As shown in Table 4, we have the following observations. (1) The running time of the three stages of plain PLMs are consistently shorter than existing KEPLMs due to the smaller size of model parameters. Specifically, existing KEPLMs contain the knowledge encoder module to project the knowledge embedding space to the contextual semantic space. (2) The running time of our model (especially during model fine-tuning and inference) is very similar to that of plain PLMs. The reason for the slightly longer time is that DKPLM adds a few projection parameters to align the knowledge triple representations. This experiment shows that DKPLM is useful for online applications due to its fast inference speed.

Influence of Long-tail and High-frequency Entities

We evaluate DKPLM using different injected entity numbers, and consider three types including long-tail entities only, high-frequency entities only and a mixture of these entities. We choose TACRED and Open Entity datasets and report the F1 metrics over testing sets to verify the effectiveness of knowledge injection. As shown in Figure 4, we can observe that: (1) Injecting knowledge triples into long-tail entities is better than high-frequency entities. (2) The state-of-the-art performance can be obtained by injecting knowledge to

Model	TACRED	Open Entity
DKPLM	77.5%	73.07%
- Long-tail Entity Detection	77.3%	72.89%
- Pseudo Token Embedding	76.7%	72.35%
- Knowledge Decoding	77.1%	72.54%

Table 5: Ablation study on two tasks (testing sets).

a fewer entities rather than all the entities. (3) Our results are consistent with Zhang et al. (2021a) in that injecting too much knowledge may hurt the performance.

Ablation Study

We report DKPLM’s performance in two knowledge-aware testing sets to perform the ablation study on the F1 metric. As shown in Table 5, we can conclude that (1) our proposed three mechanisms are effective in contributing to the complete DKPLM model. (2) The model’s performance declines significantly when removing the “Pseudo Token Embedding” mechanism. Here, the external knowledge of detected long-tail entities is not injected into the model. DKPLM degenerates to relying entirely on entity-level information to decode knowledge triples, leading to model confusion due to the sparsity of the knowledge of long-tail entities.

Conclusion and Future Work

In this paper, we propose a novel KEPLMs to decouple knowledge injection and fine-tuning for knowledge-enhanced language understanding named DKPLM. In DKPLM, we design three entity-related mechanisms to inject the knowledge information into the PLMs with minimum extra parameters for the real-world scenarios, namely knowledge-aware long-tail entity detection, pseudo token embedding injection and relational knowledge decoding. The experiments show that our model achieves the state-of-the-art performance over zero-shot knowledge probing tasks and knowledge-aware downstream tasks. Future work includes (1) selecting more effective knowledge triples from large-scale KGs to inject external knowledge into the PLMs, and (2) utilizing noised knowledge triples to further enhance the language understanding abilities of PLMs.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. This work is supported by the Alibaba Group through Alibaba Research Intern Program.

References

- Ba, L. J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *CoRR*, abs/1607.06450.
- Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*, 2787–2795.
- Broscheit, S. 2019. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. In Bansal, M.; and Villavicencio, A., eds., *CoNLL*, 677–685. Association for Computational Linguistics.
- Cao, B.; Lin, H.; Han, X.; Sun, L.; Yan, L.; Liao, M.; Xue, T.; and Xu, J. 2021. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. In *ACL*, 1860–1874.
- Chang, T. A.; Xu, Y.; Xu, W.; and Tu, Z. 2021. Convolutions and Self-Attention: Re-interpreting Relative Positions in Pre-trained Language Models. In *ACL*, 4322–4333.
- Choi, E.; Levy, O.; Choi, Y.; and Zettlemoyer, L. 2018. Ultra-Fine Entity Typing. In *ACL*, 87–96.
- Colon-Hernandez, P.; Havasi, C.; Alonso, J. B.; Huggins, M.; and Breazeal, C. 2021. Combining pre-trained language models and structured knowledge. *CoRR*, abs/2101.12294.
- Cui, B.; Li, Y.; Chen, M.; and Zhang, Z. 2019. Fine-tune BERT with Sparse Self-Attention Mechanism. In *EMNLP*, 3539–3544.
- Cui, L.; Cheng, S.; Wu, Y.; and Zhang, Y. 2021. On Commonsense Cues in BERT for Solving Commonsense Tasks. In *ACL*, 683–693.
- Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.
- Ferragina, P.; and Scaiella, U. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*, 1625–1628.
- He, Y.; Zhu, Z.; Zhang, Y.; Chen, Q.; and Caverlee, J. 2020. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *EMNLP*, 4604–4614.
- Jean, S.; Cho, K.; Memisevic, R.; and Bengio, Y. 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. In *ACL*, 1–10.
- Jiang, H.; Zhang, D.; Cao, T.; Yin, B.; and Zhao, T. 2021. Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data. In *ACL*, 1775–1789.
- Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguistics*, 8: 64–77.
- Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A Structured Self-Attentive Sentence Embedding. In *ICLR*.
- 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, X.; He, P.; Chen, W.; and Gao, J. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. In *ACL*, 4487–4496.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.
- Malik, V.; Sanjay, R.; Nigam, S. K.; Ghosh, K.; Guha, S. K.; Bhattacharya, A.; and Modi, A. 2021. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *ACL*, 4046–4062.
- Peters, M. E.; Neumann, M.; IV, R. L. L.; Schwartz, R.; Joshi, V.; Singh, S.; and Smith, N. A. 2019. Knowledge Enhanced Contextual Word Representations. In *EMNLP*, 43–54.
- Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P. S. H.; Bakhtin, A.; Wu, Y.; and Miller, A. H. 2019. Language Models as Knowledge Bases? In *EMNLP*, 2463–2473.
- Pörner, N.; Waltinger, U.; and Schütze, H. 2019. BERT is Not a Knowledge Base (Yet): Factual Knowledge vs. Name-Based Reasoning in Unsupervised QA. *CoRR*, abs/1911.03681.
- Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.
- Shen, Y.; Ma, X.; Tan, Z.; Zhang, S.; Wang, W.; and Lu, W. 2021. Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition. In *ACL*, 2782–2794.
- Su, Y.; Han, X.; Zhang, Z.; Li, P.; Liu, Z.; Lin, Y.; Zhou, J.; and Sun, M. 2020. Contextual knowledge selection and embedding towards enhanced pre-trained language models. *arXiv e-prints*, arXiv–2009.
- Sun, T.; Shao, Y.; Qiu, X.; Guo, Q.; Hu, Y.; Huang, X.; and Zhang, Z. 2020. CoLAKE: Contextualized Language and Knowledge Embedding. In *COLING*, 3660–3670.
- Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *CoRR*, abs/1904.09223.
- Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019a. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR*.
- Wang, C.; Liu, X.; and Song, D. 2020. Language Models are Open Knowledge Graphs. *CoRR*, abs/2010.11667.

- Wang, R.; Tang, D.; Duan, N.; Wei, Z.; Huang, X.; Ji, J.; Cao, G.; Jiang, D.; and Zhou, M. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *ACL*, 1405–1418.
- Wang, W.; Bi, B.; Yan, M.; Wu, C.; Xia, J.; Bao, Z.; Peng, L.; and Si, L. 2020. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. In *ICLR*.
- Wang, X.; Gao, T.; Zhu, Z.; Liu, Z.; Li, J.; and Tang, J. 2019b. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *CoRR*, abs/1911.06136.
- Xu, Z.; Guo, D.; Tang, D.; Su, Q.; Shou, L.; Gong, M.; Zhong, W.; Quan, X.; Jiang, D.; and Duan, N. 2021. Syntax-Enhanced Pre-trained Model. In *ACL*, 5412–5422.
- Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NIPS*, 5754–5764.
- Zhang, N.; Deng, S.; Cheng, X.; Chen, X.; Zhang, Y.; Zhang, W.; Chen, H.; and Center, H. I. 2021a. Drop Redundant, Shrink Irrelevant: Selective Knowledge Injection for Language Pretraining. In *IJCAI*.
- Zhang, N.; Deng, S.; Li, J.; Chen, X.; Zhang, W.; and Chen, H. 2020. Summarizing Chinese Medical Answer with Graph Convolution Networks and Question-focused Dual Attention. In *EMNLP*, 15–24.
- Zhang, T.; Cai, Z.; Wang, C.; Qiu, M.; Yang, B.; and He, X. 2021b. SMedBERT: A Knowledge-Enhanced Pre-trained Language Model with Structured Semantics for Medical Text Mining. In *ACL*, 5882–5893. Association for Computational Linguistics.
- Zhang, Y.; and Qi, P. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *EMNLP*, 2205–2215.
- Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *EMNLP*, 35–45.
- 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.