

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

Xiaozhi Wang¹, Tianyu Gao¹, Zhaocheng Zhu^{2,3}, Zhiyuan Liu¹, Juanzi Li¹, Jian Tang^{2,4,5}

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China

²Mila - Québec AI Institute, Montréal, Canada

³Université de Montréal, Montréal, Canada

⁴HEC, Montréal, Canada

⁵CIFAR AI Research Chair

Abstract

Pre-trained language representation models (PLMs) learn effective language representations from large-scale unlabeled corpora. Knowledge embedding (KE) algorithms encode the entities and relations in knowledge graphs into informative embeddings to do knowledge graph completion and provide external knowledge for various NLP applications. In this paper, we propose a unified model for **Knowledge Embedding and Pre-trained Language Representation (KEPLER)**, which not only better integrates factual knowledge into PLMs but also effectively learns knowledge graph embeddings. Our KEPLER utilizes a PLM to encode textual descriptions of entities as their entity embeddings, and then jointly learn the knowledge embeddings and language representations. Experimental results on various NLP tasks such as the relation extraction and the entity typing show that our KEPLER can achieve comparable results to the state-of-the-art knowledge-enhanced PLMs without any additional inference overhead. Furthermore, we construct Wikidata5m, a new large-scale knowledge graph dataset with aligned text descriptions, to evaluate KE embedding methods in both the traditional transductive setting and the challenging inductive setting, which needs the models to predict entity embeddings for unseen entities. Experiments demonstrate our KEPLER can achieve good results in both settings.

1 Introduction

Pre-trained language representation models (PLMs) such as ELMo (Peters et al., 2018a), BERT (Devlin et al., 2019a) and XLNet (Yang et al., 2019) learn effective language representations from large-scale nonstructural and unlabelled

corpora and achieve great performance on various NLP tasks. However, they are typically lack of factual world knowledge (Petroni et al., 2019; Logan et al., 2019).

Recent works (Zhang et al., 2019; Peters et al., 2017; Liu et al., 2019a) utilize entity embeddings of large-scale knowledge bases to provide external knowledge for PLMs and improve their performance on various NLP tasks. However, they have some issues: (1) They use fixed entity embeddings learned by a separate knowledge embedding (KE) algorithm, which cannot be easily aligned with the language representations because they are essentially in two different vector spaces. (2) They require an entity linker to link the words in context to corresponding entities so that they can benefit from the entity embeddings, which makes them suffer from the error propagation problem. (3) Their sophisticated mechanisms to retrieve and use entity embeddings lead to additional inference overhead compared with vanilla PLMs.

Actually, knowledge embedding methods have a strong connection with NLP models. There are not only many works integrating knowledge embeddings into NLP models to improve the performance of NLP applications such as machine translation (Zareemoodi et al., 2018), reading comprehension (Mihaylov and Frank, 2018; Zhong et al., 2019) and dialogue system (Madotto et al., 2018), but also some early works use text as additional information (Xie et al., 2016; An et al., 2018) or jointly train the knowledge and text embedding in the same space (Wang et al., 2014; Toutanova et al., 2015; Han et al., 2016; Cao et al., 2017, 2018).

In this paper, we propose to learn knowledge embedding and language representation with a unified model and encode them into the same semantic space, which can not only better integrate knowledge into PLMs but also help to learn more informative knowledge embeddings with the effec-

Work in progress.

tive language representations. We propose **KEPLER**, which is short for “a unified model for **K**nowledge **E**mbedding and **P**re-trained **L**anguage **R**epresentation”. We collect informative textual descriptions for entities in the knowledge graph and utilizes a typical PLM to encode the descriptions as text embeddings, then we treat the description embeddings as entity embeddings and optimize a KE objective function on top of them. The key idea is to encode structural knowledge in the textual representation of entities using a PLM, which can generalize to unobserved entities in the knowledge graph.

Our KEPLER enjoys the following advantages: (1) We integrate world knowledge into PLMs with the supervision of the KE objective, which is more flexible for the PLMs, and encode the entity and text into the same space, which avoids the gap between the language representations and fixed entity embeddings. (2) We do not need an entity linker or additional mechanisms to retrieve corresponding entity embeddings, which avoids the error propagation problem and extra overhead. During inference, our KEPLER is exactly the same as standard PLMs, which can be adopted in a wide range of NLP applications. (3) Different from conventional KE methods, our KEPLER encodes textual entity descriptions as entity embeddings, which enables our model to infer knowledge embedding in the inductive setting (get entity embeddings for the unseen entities). This is especially useful for deployment, where the model may deal with unseen entities.

The existing KE datasets are relatively small-scale, which is not sufficient to pre-train a large model, and typically lack of description data and a data split for the inductive setting. Therefore, we construct Wikidata5m, a new large-scale knowledge graph dataset with aligned text description for each entity. Wikidata5m is a subset of Wikidata (Vrandečić and Krötzsch, 2014), a free knowledge base with about sixty million entities. To ensure each entity is informative and the knowledge base is as clean as possible, we only select the entities with corresponding Wikipedia pages. Our Wikidata5m contains five million entities and twenty million triplets. We also benchmark several classical KE methods on Wikidata5m to facilitate future research. To our knowledge, this is the first million-scale general knowledge graph dataset.

To summarize, our contribution is three-fold: (1) We propose to encode entities and texts into

the same space and jointly train the KE and language modeling objectives, and then get a better knowledge-enhanced PLM which avoids error propagation and additional overhead. Experimental results on various NLP tasks demonstrate the effectiveness of our KEPLER. (2) We encode textual descriptions as entity embeddings, which improves KE with textual information and enables inductive KE. (3) We introduce a new large-scale knowledge graph dataset Wikidata5m, which may promote the research on large-scale knowledge graph, inductive knowledge embedding and interactions between knowledge graph and NLP.

2 Related Work

Pre-trained Language Model There has been a long history of pre-training in NLP. Early works focus on distributed word representation (Collobert and Weston, 2008; Mikolov et al., 2013; Pennington et al., 2014), many of which are still often adopted in current models as word embeddings for their ability to capture syntactic and semantic information from large-scale corpora. Peters et al. (2018b) push this trend a step forward by using a bidirectional LSTM to capture contextualized word embeddings (ELMo) for richer semantic meanings under different circumstances.

Apart from those methods using pre-trained word embeddings as input features, there is another trend exploring pre-trained encoders. Dai and Le (2015) first propose to train an auto-encoder on unlabeled data, and then fine-tune it on downstream tasks. Howard and Ruder (2018) propose a universal language model (ULMFiT) based on AWD-LSTM (Merity et al., 2018). With the powerful Transformer (Vaswani et al., 2017) as its encoder, Radford et al. (2018) demonstrate a pre-trained generative model (GPT) and its effects, while Devlin et al. (2019b) release a pre-trained deep Bidirectional Encoder Representation from Transformers (BERT), achieving state-of-the-arts on dozens of benchmarks.

After Devlin et al. (2019b), similar pre-trained encoders spring up recently. Yang et al. (2019) propose a permutation language model (XLNet) based on TransformerXL (Dai et al., 2019). Later, Liu et al. (2019c) show that more data and more sophisticated parameter tuning would benefit pre-trained encoders a lot and release a new state-of-the-art model (Roberta). Other works explore how to add more tasks (Liu et al., 2019b) and more

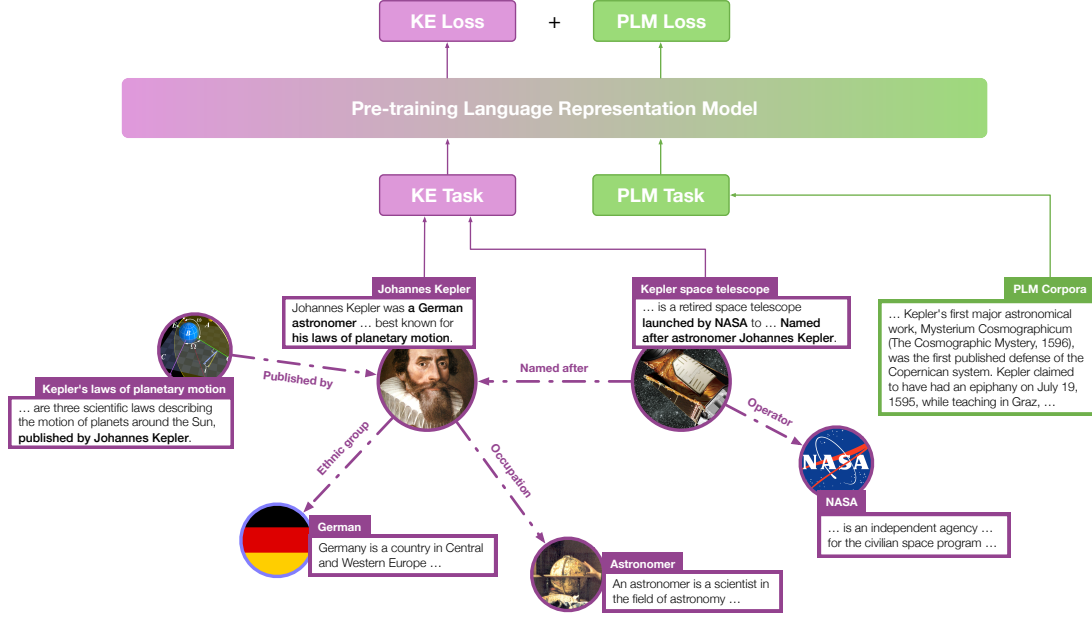


Figure 1: A demonstration for KEPLER structure. By jointly training with knowledge embedding (KE) and pre-training language representation model (PLM) objectives, our framework can implicitly incorporate knowledge into the language representation model.

parameters (Lan et al., 2019; Raffel et al., 2019) to pre-training models.

Recently some works attempt to incorporate knowledge information in pre-training. Zhang et al. (2019) introduce pre-processed knowledge embeddings into the Transformer architecture of BERT (Devlin et al., 2019b). With similar ideas, Peters et al. (2019) incorporate an integrated entity linker in their models. Besides, Logan et al. (2019); Hayashi et al. (2019) utilize relations between entities inside one sentence to help train better generation models. Despite the promising results those methods bring with knowledge-enhanced techniques, they either use fixed external knowledge information or have complex structures or pipelines to handle entities within sentences.

Knowledge Graph Embeddings In recent years knowledge embeddings have been extensively studied through predicting missing links in graphs. Conventional models define score functions for relation triples (h, r, t) and predict head or tail entities with scores of candidate entities. For example, TransE (Bordes et al., 2013) treats tail entities as translations of heads, while DistMult (Yang et al., 2015) use matrix multiplications as score functions and ComplEx (Trouillon et al., 2016) adopt complex operations based on it. RotatE (Sun et al., 2019a) combines the advantages of both of them.

Among these works, Xie et al. (2016) propose

to utilize entity descriptions as an external information source and introduce an entity description encoder to enhance the TransE score function. Though similar to our method, Xie et al. (2016) aim at utilizing entity descriptions to help knowledge representation learning, while we take entity descriptions as a tool to incorporate external knowledge in our model.

3 KEPLER Model

In this section, we introduce the structure of our KEPLER model, and how we combine two training goals of masked language modeling and knowledge representation learning.

3.1 Training Objectives

To incorporate world knowledge into our pre-trained language representation models (PLMs), we design a multi-task loss as shown in Figure 1 and Equation 1,

$$\mathcal{L} = \mathcal{L}_{KE} + \mathcal{L}_{LM}, \quad (1)$$

where \mathcal{L}_{KE} represents knowledge embedding loss and \mathcal{L}_{LM} represents language model loss. Since our PLMs are involved in both tasks, jointly optimizing the two objectives could implicitly integrate knowledge from external graphs with text encoders, while keeping the strong abilities of PLMs for syntactic and semantic understanding.

More specifically, we adopt a general \mathcal{L}_{KE} format using negative sampling,

$$L = -\log d_r(\mathbf{h}, \mathbf{t}) + \sum_{i=1}^n \frac{1}{n} \log d_r(\mathbf{h}'_i, \mathbf{t}'_i), \quad (2)$$

where (h, r, t) is the correct triple from knowledge graphs and (h'_i, r, t'_i) are negative sampling triples. d_r is the score function, for which we have many choices. Different from conventional knowledge embedding methods, for entity embeddings \mathbf{h} and \mathbf{t} , instead of looking up in embedding tables, we use PLMs as our text encoders to extract entity representations from their descriptions.

For \mathcal{L}_{LM} , many alternatives for pre-trained language representation can be used, e.g., masked language model (Devlin et al., 2019b). Note that those two tasks only share the text encoder and for each mini-batch, text sampled for \mathcal{L}_{KE} and \mathcal{L}_{LM} is not (necessarily) the same.

3.2 Model Details

Though we have many alternatives of model structures and training objectives to choose under KEPLER framework, here for better clarity, we introduce a specific one that we use in experiments.

Model Structure We use the transformer architecture (Vaswani et al., 2017) as in (Devlin et al., 2019b; Liu et al., 2019c), which we will not address in details. To be more specific, we use RoBERTa_{BASE} codes and checkpoints¹ in all our experiments since it is one of the state-of-the-art pre-trained models with acceptable computing requirements. Besides the training data and hyperparameters, one of the major differences between RoBERTa and BERT is that RoBERTa uses Byte-Pair Encoding (BPE) (Sennrich et al., 2016) to better tokenize rare words.

Given a sequence of tokens x_1, x_2, \dots, x_N , the input format is $[\text{CLS}], x_1, x_2, \dots, x_N, [\text{EOS}]$, where $[\text{CLS}]$ and $[\text{EOS}]$ are two special tokens. Model output at $[\text{CLS}]$ is often used as the sentence representation.

PLM Objective Inspired by BERT (Devlin et al., 2019b), MLM randomly selects 15% of input tokens, among which 80% are masked with the special mark $[\text{MASK}]$, 10% are replaced by another

random token, and the rest remain unchanged. Under MLM, models try to predict the correct tokens and a cross-entropy loss is calculated over the selected positions.

We adopt the pre-trained checkpoint of RoBERTa_{BASE} for the initialization of our model. However, we still keep MLM as one of our objectives to avoid catastrophic forgetting (McCloskey and Cohen, 1989) while training towards the KRL loss. Note that experiments show that only further pre-training from RoBERTa_{BASE} checkpoint does not bring promotion, suggesting that the combination of the two tasks contributes most to the performance.

KE Objective We use the loss formula from (Sun et al., 2019b) as our KE objective, which takes negative sampling (Mikolov et al., 2013) for efficient optimization:

$$L = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^n \frac{1}{n} \log \sigma(d_r(\mathbf{h}'_i, \mathbf{t}'_i) - \gamma), \quad (3)$$

where (h, r, t) is the correct triple, (h'_i, r, t'_i) are negative sampling triples, γ is the margin, σ is the sigmoid function, and d_r is the score function, for which we choose to follow TransE (Bordes et al., 2013) for its simplicity and efficiency,

$$d_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_p \quad (4)$$

where we take the norm p as 1. Due to the limit of computing resources, we take the negative sampling size n as 1. The negative sampling policy is to fix the head entity and randomly sample a tail entity, and vice versa.

Different from conventional KE methods, we do not have an entity embedding lookup table. Instead, we use our KEPLER model to encode the corresponding entity descriptions and take the $[\text{CLS}]$ outputs as the entity embeddings.

3.3 Downstream Tasks

Like all BERT-like models, we fine-tune KEPLER on downstream tasks and use $[\text{CLS}]$ output for sentence-level prediction and the outputs of all tokens for sequence labelling tasks (Devlin et al., 2019b). For supervised relation extraction and few-shot relation extraction, we follow the approaches from (Baldini Soares et al., 2019) and (Gao et al., 2019) respectively.

¹<https://github.com/pytorch/fairseq>

Dataset	#entity	#relation	#training	#validation	#test
FB15K	14,951	1,345	483,142	50,000	59,071
WN18	40,943	18	141,442	5,000	5,000
FB15K-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134
Wikidata5m	4,594,485	822	20,542,906	40,641	41,028

Table 1: Statistics of Wikidata5m compared with existing widely-used benchmarks.

4 Wikidata5m

We construct a new large-scale knowledge graph dataset with aligned text descriptions. Our dataset is built by integrating Wikidata (Vrandečić and Krötzsch, 2014), a large-scale open knowledge base, with Wikipedia. Each entity in the knowledge graph is aligned with its text description in Wikipedia pages. In the following sections, we will first introduce the data collection steps, and then give the benchmarks of popular KE methods on this dataset.

4.1 Data Collection

We pull the latest dump of Wikidata² and Wikipedia³ from their websites respectively. We remove pages whose first paragraphs contain fewer than 5 words. For each entity, we align it to a Wikipedia page with the MediaWiki *wbgetentities* action API. The first section of Wikipedia pages is extracted as the description for entities. Entities that have no corresponding Wikipedia pages are discarded.

To construct the knowledge graph, we retrieve all the statements in entity pages, and map the entities and relations in statements to their canonical IDs in Wikidata. A statement is considered to be a valid triplet if both of its entities can be aligned with Wikipedia pages and its relation has a non-empty page in Wikidata. The final knowledge graph dataset contains 4,813,455 entities, 822 relations and 21,344,269 triplets, where each entity has a text description. Statistics of our Wikidata5m dataset and four widely-used datasets are showed in Table 1. Top-5 entity categories are listed in Table 3. We can see that our Wikidata5m is much larger than existing knowledge graph datasets, covering all sorts of domains.

²<https://www.wikidata.org>

³<https://en.wikipedia.org>

Subset	#entity	#relation	#triplet
Training	4,579,609	822	20,496,514
Validation	7,374	199	6,699
Test	7,475	201	6,894

Table 2: Statistics of Wikidata5m inductive setting.

Entity Type	Occurrence	Percentage
Human	1,517,591	31.5%
Taxon	363,882	7.56%
Wikimedia list	118,823	2.47%
Film	114,266	2.37%
Human Settlement	110,939	2.30%
Total	2,225,501	46.2%

Table 3: Top-5 entity categories in Wikidata5m.

4.2 Data Split

The data split statistics for the conventional transductive setting are also shown in Table 1.

In this work, we also evaluate models on the challenging inductive setting, which requires the models to produce entity embeddings for entities which are not seen at the training time and also do link predictions for the unseen entities. So we provide a data split for the inductive setting evaluation. The statistics for the inductive setting data split are shown in Table 2. In the inductive setting, the entities and triplets in training, validation and test sets are mutually disjoint, while in the transductive setting, only the triplet sets are mutually disjoint.

4.3 Benchmarks

To assess the challenges of Wikidata5m, we benchmark several popular knowledge graph embedding models on the dataset. Since the conventional knowledge graph embedding models are inherently transductive, we split the triplets of knowledge graph into train, valid and test sets. Each model is trained on the training set and evaluated on the link

Method	MR	MRR	HITS@1	HITS@3	HITS@10
TransE (Bordes et al., 2013)	109370	0.253	0.170	0.311	0.392
DistMult (Yang et al., 2015)	211030	0.253	0.208	0.278	0.334
ComplEx (Trouillon et al., 2016)	244540	0.281	0.228	0.310	0.373
Simple (Kazemi and Poole, 2018)	115263	0.296	0.252	0.317	0.377
RotatE (Sun et al., 2019b)	89459	0.290	0.234	0.322	0.390

Table 4: Performance of different knowledge graph embedding models on Wikidata5m.

prediction task.

We conduct 5 knowledge graph embedding models, including TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), Simple (Kazemi and Poole, 2018) and RotatE (Sun et al., 2019b). Because their original implementations do not scale to Wikidata5m, we benchmark these methods using the multi-GPU implementation in GraphVite (Zhu et al., 2019). The performance of link prediction is evaluated in the filtered setting, where test triplets are ranked against all candidate triplets that are not observed in the knowledge graph. We report the standard metrics of Mean Rank (MR), Mean Reciprocal Rank (MRR) and Hits at N (HITS@N).

Table 4 shows the benchmarks of popular methods on Wikidata5m.

5 Experiments

In this section, we introduce the experiment settings and experimental results of KEPLER on various NLP and KE tasks.

5.1 Pre-training settings

In experiments, we choose RoBERTa (Liu et al., 2019c) as our base model and implement our methods in the fairseq framework (Ott et al., 2019) for pre-training. Due to the computing resource limit, we choose RoBERTa_{BASE} architecture and use the released roberta.base⁴ parameters to initialize our model.

In our pre-training procedure, we only use the English Wikipedia corpus to save time and also for a fair comparison with previous knowledge-enhanced PLMs (Zhang et al., 2019; Peters et al., 2019).

⁴<https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.md>

Model	P	R	F-1
BERT _{BASE} *	67.23	64.81	66.0
BERT _{LARGE} +	-	-	70.1
ERNIE _{BASE} *	69.97	66.08	67.97
MTB _{LARGE} +	-	-	71.50
RoBERTa _{BASE}	70.07	70.63	70.35
KnowBERT _{BASE} #	71.60	71.40	71.50
KEPLER _{BASE}	70.43	73.02	71.70

Table 5: Results on the relation classification dataset TACRED (%). Results with *, + and # are from Zhang et al. (2019), Baldini Soares et al. (2019) and Peters et al. (2019) respectively. BASE and LARGE identify whether the model uses a base or large version BERT-like architecture.

5.2 NLP Tasks

In this section, we introduce how our KEPLER can be used as a knowledge-enhanced PLM on various NLP tasks and its performance compared with state-of-the-art models.

Relation Classification

Relation classification is an important NLP task that requires models to classify relation types between two given entities from text. We evaluate our model and baselines on two commonly-used datasets: TACRED (Zhang et al., 2017) and FewRel (Han et al., 2018). TACRED covers 42 relation types and contains 106,264 sentences. FewRel is a few-shot relation classification dataset, which has 100 relations and 700 instances for each relation.

Here we follow the relation extraction fine-tuning procedure from Zhang et al. (2019), where four special tokens are added before and after entity mentions in the sentence to highlight where the entities are. Then we take the [CLS] output as the sentence representation for classification.

Table 5 shows results of various models on TACRED, from which we can see that our model

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Proto (BERT _{BASE})	80.68	89.60	71.48	82.89
Proto (RoBERTa _{BASE})	88.60	93.81	80.83	92.28
Proto (KEPLER)	89.77	95.86	82.93	92.70
PAIR (BERT _{BASE})	88.97	94.3	81.43	87.93
PAIR (RoBERTa _{BASE})	88.74	93.90	83.08	89.16
PAIR (KEPLER)	90.87	95.40	84.74	90.31
MTB (BERT _{LARGE})	93.86	97.06	89.20	94.27

Table 6: Accuracies (%) on FewRel dataset. “Proto” indicates Prototypical Networks (Snell et al., 2017) used in Han et al. (2018). “PAIR” is proposed in Gao et al. (2019) and “MTB” is from Baldini Soares et al. (2019).

Model	P	R	F-1
UFET *	68.8	53.3	60.1
BERT _{BASE}	76.4	71.0	73.6
RoBERTa _{BASE}	76.6	73.7	75.1
ERNIE ⁺	78.4	72.9	75.6
KnowBERT [#]	78.6	73.7	76.1
KEPLER	77.2	74.2	75.7

Table 7: Entity typing results on OpenEntity (%). Models with * + # come from Choi et al. (2018); Zhang et al. (2019); Peters et al. (2019) respectively. Our KEPLER performs better than RoBERTa_{BASE}, and achieves comparable results with other state-of-the-art models.

achieves state-of-the-art on this benchmark. Note that some baselines use the LARGE version of pre-trained language models while we still take the BASE architecture. We have gained a large promotion over our base model (RoBERTa_{BASE}) while staying a little bit advanced over other competitive methods (even if they use a LARGE architecture).

Our model has also shown strength on FewRel dataset. We use Prototypical Networks (Snell et al., 2017) and PAIR (Gao et al., 2019) as the base frameworks and try out different kinds of pre-trained models as encoders. As shown in Table 6, for both frameworks, our models have superior performance over others. We have also compared with current state-of-the-art MTP (Baldini Soares et al., 2019), which outperforms us a little. But note that MTP uses a large version of BERT while we use the base version, and also it carries out a new pre-training task specifically targeting relation extraction, while ours is a general way to combine knowledge and natural language which would benefit all knowledge-related tasks.

Method	MR	MRR	HITS@1	HITS@3	HITS@10
KEPLER	30.8387	0.217	0.0	0.360	0.692

Table 8: Performance of KEPLER on inductive setting in Wikidata5m.

Entity Typing

Entity typing requires models to classify given entity mentions into pre-defined entity types. For this task, we evaluate all the models on OpenEntity (Choi et al., 2018) following the setting from Zhang et al. (2019), which focuses on nine general entity types.

Evaluation results are demonstrated in Table 7. For now we have achieved better results than RoBERTa, and ERNIE and KnowBERT show slightly better results than ours. It is mainly due to that we use different ways of extracting entity representations. KnowBERT adds special tokens before and after the mention and uses the output of the token before the mention as the representation for typing, while ours, for now, directly uses [CLS]. We will try this better way of entity representation in the future.

5.3 Knowledge Embedding

In this section, we show how our KEPLER works as a KE model, and evaluate it on our Wikidata5m dataset in inductive setting.

We do not use the existing KE benchmarks because they are lack of high-quality text descriptions for their entities and they do not have a reasonable data split for the inductive setting.

Inductive Setting

We evaluate the generalization ability of our KEPLER by testing it on the inductive setting in Wikidata5m (as described in Section 4.2), which re-

quires it to produce effective entity embeddings for the unseen entities. The results are shown in Table 8.

6 Conclusion and Future Work

In this paper, we propose KEPLER, a unified model for knowledge embedding and pre-trained language representation. We jointly train the knowledge embedding and language representation objectives on top of the language representation model. Experimental results on extensive tasks demonstrate the effectiveness of our model.

In the future, we will: (1) Evaluate whether our model can recall factual knowledge with more tasks. (2) Try variations of existing models, such as highlighting entity mentions in descriptions or changing knowledge embedding form, to get better understanding of how KEPLER works and bring more promotion for downstream tasks.

References

- Bo An, Bo Chen, Xianpei Han, and Le Sun. 2018. [Accurate text-enhanced knowledge graph representation learning](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 745–755, New Orleans, Louisiana. Association for Computational Linguistics.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. [Matching the blanks: Distributional similarity for relation learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. [Translating embeddings for modeling multi-relational data](#). In *Proceedings of NIPS*, pages 2787–2795.
- Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Chengjiang Li, Xu Chen, and Tiansi Dong. 2018. [Joint representation learning of cross-lingual words and entities via attentive distant supervision](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 227–237, Brussels, Belgium. Association for Computational Linguistics.
- Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. [Bridge text and knowledge by learning multi-prototype entity mention embedding](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1623–1633, Vancouver, Canada. Association for Computational Linguistics.
- Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. [Ultra-fine entity typing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.
- Ronan Collobert and Jason Weston. 2008. [A unified architecture for natural language processing: Deep neural networks with multitask learning](#). In *Proceedings of ICML*, pages 160–167.
- Andrew M Dai and Quoc V Le. 2015. [Semi-supervised sequence learning](#). In *Proceedings of NIPS*, pages 3079–3087.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. [Transformer-XL: Attentive language models beyond a fixed-length context](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of NAACL-HLT*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. [FewRel 2.0: Towards more challenging few-shot relation classification](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6251–6256, Hong Kong, China. Association for Computational Linguistics.
- Xu Han, Zhiyuan Liu, and Maosong Sun. 2016. [Joint representation learning of text and knowledge for knowledge graph completion](#). *arXiv preprint arXiv:1611.04125*.
- Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. [FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

- Hiroaki Hayashi, Zecong Hu, Chenyan Xiong, and Graham Neubig. 2019. [Latent relation language models](#). *arXiv preprint arXiv:1908.07690*.
- Jeremy Howard and Sebastian Ruder. 2018. [Universal language model fine-tuning for text classification](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Seyed Mehran Kazemi and David Poole. 2018. [Simple embedding for link prediction in knowledge graphs](#). In *Advances in Neural Information Processing Systems*, pages 4284–4295.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. [Albert: A lite bert for self-supervised learning of language representations](#). *arXiv preprint arXiv:1909.11942*.
- Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. [K-bert: Enabling language representation with knowledge graph](#). *arXiv preprint arXiv:1909.07606*.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. [Multi-task deep neural networks for natural language understanding](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. [Barack’s wife hillary: Using knowledge graphs for fact-aware language modeling](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.
- Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. [Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.
- Michael McCloskey and Neal J Cohen. 1989. [Catastrophic interference in connectionist networks: The sequential learning problem](#). In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. [Regularizing and optimizing lstm language models](#). In *Proceedings of ICLR*.
- Todor Mihaylov and Anette Frank. 2018. [Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Proceedings of NIPS*, pages 3111–3119.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [Glove: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. [Semi-supervised sequence tagging with bidirectional language models](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. [Deep contextualized word representations](#). In *Proceedings of NAACL-HLT*, pages 2227–2237.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. [Knowledge enhanced contextual word representations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#) *arXiv preprint arXiv:1909.01066*.

- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#). In *Proceedings of Technical report, OpenAI*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *arXiv preprint arXiv:1910.10683*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Jake Snell, Kevin Swersky, and Richard Zemel. 2017. [Prototypical networks for few-shot learning](#). In *Advances in Neural Information Processing Systems*, pages 4077–4087.
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019a. [Rotate: Knowledge graph embedding by relational rotation in complex space](#).
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019b. [Rotate: Knowledge graph embedding by relational rotation in complex space](#). *arXiv preprint arXiv:1902.10197*.
- Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoi-fung Poon, Pallavi Choudhury, and Michael Gamon. 2015. [Representing text for joint embedding of text and knowledge bases](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal. Association for Computational Linguistics.
- 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of NIPS*, pages 5998–6008.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: a free collaborative knowledge base](#).
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. [Knowledge graph and text jointly embedding](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1591–1601, Doha, Qatar. Association for Computational Linguistics.
- Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. [Representation learning of knowledge graphs with entity descriptions](#). In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI’16*, pages 2659–2665. AAAI Press.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. [Embedding entities and relations for learning and inference in knowledge bases](#).
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). *arXiv preprint arXiv:1906.08237*.
- Poorya Zareemoodi, Wray Buntine, and Gholamreza Haffari. 2018. [Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Melbourne, Australia. Association for Computational Linguistics.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. [ERNIE: Enhanced language representation with informative entities](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.
- Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. [Improving question answering by commonsense-based pre-training](#). In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 16–28. Springer.
- Zhaocheng Zhu, Shizhen Xu, Jian Tang, and Meng Qu. 2019. [Graphvite: A high-performance cpu-gpu hybrid system for node embedding](#). In *The World Wide Web Conference*, pages 2494–2504. ACM.