KLMo: Knowledge Graph Enhanced Pretrained Language Model with Fine-Grained Relationships

Lei He, Suncong Zheng, Tao Yang, Feng Zhang
Tencent AI Platform Department, China
{bettyleihe,congzheng,rigorosyang,jayzhang}@tencent.com

Abstract

Interactions between entities in knowledge graph (KG) provide rich knowledge for language representation learning. However, existing knowledge-enhanced pretrained language models (PLMs) only focus on entity information and ignore the fine-grained relationships between entities. In this work, we propose to incorporate KG (including both entities and relations) into the language learning process to obtain KG-enhanced pretrained Language Model, namely KLMo. Specifically, a novel knowledge aggregator is designed to explicitly model the interaction between entity spans in text and all entities and relations in a contextual KG. An relation prediction objective is utilized to incorporate relation information by distant supervision. An entity linking objective is further utilized to link entity spans in text to entities in KG. In this way, the structured knowledge can be effectively integrated into language representations. Experimental results demonstrate that KLMo achieves great improvements on several knowledge-driven tasks, such as entity typing and relation classification, comparing with the state-of-the-art knowledge-enhanced PLMs.

1 Introduction

Knowledge Graph (KG) with entities and relations provides rich knowledge for language learning (Wang et al., 2017, 2014). Recently, researchers have explored to incorporate KG information into PLMs (Devlin et al., 2018; Radford et al.) to enhance language representations, such as ERNIE-THU (Zhang et al., 2019), WKLM (Xiong et al., 2019), KEPLER (Wang et al., 2019), KnowBERT (Peters et al., 2019), BERT-MK (He et al., 2019) and KALM (Rosset et al., 2020), . However, they only utilize entity information and ignore the fine-grained relationships between entities. The fine-grained semantic information of relations between entities is also critical to language representation learning. Taking Figure 1 as example, for entity typing, without explicitly knowing the fine-grained relation *Guest* between *Lang Lang* and *Trio of Happiness*, which is different from the relation *Performer* between *Lang Lang* and *BBC Proms*, it's impossible to correctly predict the type of *Trio of Happiness* as *TV Show*, since the input sentence literally implies that *Trio of Happiness* belongs to the same type as *BBC Proms*. The fine-grained relations between entities in KG provide specific constraint on entities, thus can play an important role in language learning for knowledge-driven tasks.

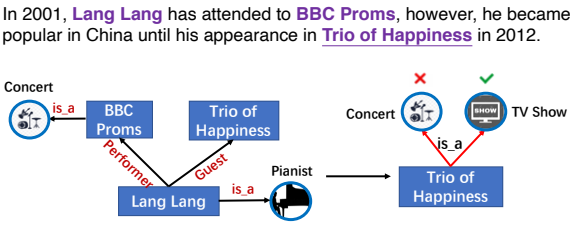

Figure 1: An illustrative example of incorporating knowledge into PLMs. The relations in KG is critical to correctly predict the type of *Trio of Happiness*.

To explicitly incorporate entities and fine-grained relations in KG into PLMs, one main challenge we are faced with is the Text-Knowledge Alignment (TKA) problem: it's difficult to make token-relation and token-entity alignments for the fusion of text and knowledge. To handle this problem, the KG-enhanced pretrained language model (KLMo) is proposed to integrate KG (i.e. both entities and fine-grained relations) into the language representation learning. The main component of KLMo is a knowledge aggregator, which is responsible for text and knowledge information fusion from two individual embedding spaces, i.e. token embedding space and KG embedding space. The knowledge aggregator models the interaction between entity spans in text and all entities and relations in a contextual KG via an entity span-level cross-KG attention to make tokens attend to highly

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4536–4542
November 7–11, 2021.

related entities and relations in KG. Based on the KG-enhanced token representations, a relation prediction objective is utilized to predict the relation of each pair of entities in text based on the distant supervision of KG. Furthermore, an entity linking objective is utilized to predict entities in KG based on the corresponding entity spans in text. The relation prediction and entity linking objectives are the key to the integration of KG information into text representations.

We conduct experiments on two Chinese knowledge-driven NLP tasks, i.e. entity typing and relation classification. The experimental results demonstrate that KLMo obtains large improvements over BERT and existing knowledge-enhanced PLMs, by taking full advantage of a structured KG including both entities and fine-grained relations. We also will publish a Chinese entity typing dataset for the evaluation of Chinese PLMs.

2 Model Description

As shown in Figure 2, KLMo is designed as a multi-layer Transformer-based (Vaswani et al., 2017) model, which accepts a token sequence and the entities and relations in its contextual KG as input. The token sequence is firstly encoded by a multi-layer Transformer-based text encoder. The output of the text encoder is further used as input for the knowledge aggregator that fuses the knowledge embeddings of entities and relations into the token sequence to obtain KG-enhanced token representations. Based on the KG-enhanced representations, novel relation prediction and entity linking objectives are jointly optimized as the pre-training objectives, which help incorporate high-related entity and relation information in the KG into the text representations.

2.1 Knowledge Aggregator

As shown in Figure 2, the knowledge aggregator is designed as an M-layer knowledge encoder to integrate knowledge in KG into language representation learning. It accepts the hidden embeddings of the token sequence and the knowledge embeddings of the entities and relations in KG as input, and fuses text and KG information from two individual embedding spaces. The knowledge aggregator contains two separate multi-head attentions: token-level self-attention and knowledge graph attention (Veličković et al., 2017), which encodes the input text and the KG independently. The entity

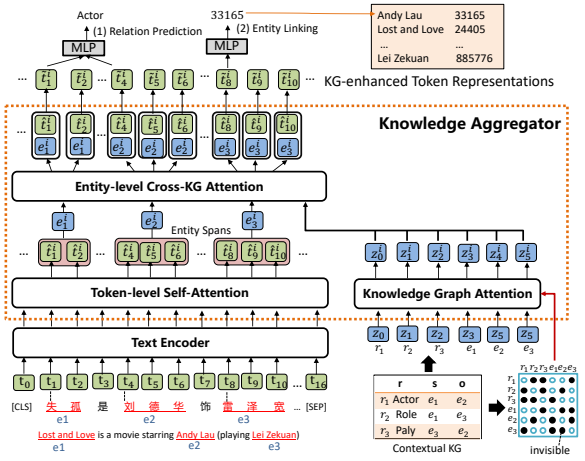

Figure 2: Overview of the model architecture.

representation is computed by pooling over all tokens in an entity-span. Then the aggregator models the interaction between entity spans in text and all entities and relations in a contextual KG through an entity-level cross-KG attention to incorporate knowledge into the text representations.

Knowledge Graph Attention As the entities and relations in a KG composes a graph, it's critical to considering the graph structure during knowledge representation learning. We first represent entities and relations in the contextual KG by TransE (Bordes et al., 2013) and then translate them into an entity and relation embedding sequence $\{\boldsymbol{z}_0, \boldsymbol{z}_1, ..., \boldsymbol{z}_q\}$, served as the input for the knowledge aggregator. Then the knowledge aggregator encodes the entity and relation sequence by a knowledge graph attention which considers its graph structure by importing a visible matrix M into the traditional self-attention mechanism (Liu et al., 2020). The visible matrix M only allows adjacent entities and relations in the KG to be visible to each other during representation learning, as shown in the right bottom of Figure 2.

Entity-level Cross-KG Attention To compute the KG-enhanced entity representations, given an entity mention list $C_e = \{(e_0, start_0, end_0), ..., (e_m, start_m, end_m)\}$, the knowledge aggregator first computes the entity span representations $\{\hat{\boldsymbol{e}}^i_0, ..., \hat{\boldsymbol{e}}^i_m\}$ by pooling over all tokens in an entity-span with self-attentive span pooling method from (Lee et al., 2017). The entity span embeddings $\{\hat{\boldsymbol{e}}^i_0, ..., \hat{\boldsymbol{e}}^i_m\}$ can be expanded to all tokens $\{\hat{\boldsymbol{e}}^i_0, ..., \hat{\boldsymbol{e}}^i_n\}$ by making $\hat{\boldsymbol{e}}^i_j = \hat{\boldsymbol{t}}^i_j$ for tokens not in any entity spans, where $\hat{\boldsymbol{t}}^i_j$ denotes the representation of the j-th token from the

token-level self-attention.

In order to model the interaction between entity spans in text and all entities and relations in a contextual KG, the aggregator performs an entity-level cross-KG attention to allow tokens attend to highly related entities and relations in KG, thus computes the KG-enhanced entity representations. Specifically, the entity-level cross-KG attention in the i-th aggregator is performed by contextual multi-head attention between the entity span embeddings $\{\hat{\boldsymbol{e}}_0^i, ..., \hat{\boldsymbol{e}}_n^i\}$ as the query and the entity and relation embeddings $\{\boldsymbol{z}_0^i, ..., \boldsymbol{z}_q^i\}$ as the key and value.

KG-enhanced Token Representations To inject the KG-enhanced entity information into the token representations, the i-th layer of the knowledge aggregator computes the KG-enhanced token representations $\{\tilde{\boldsymbol{t}}_0^i, ..., \tilde{\boldsymbol{t}}_n^i\}$ by adopting an information fusion operation between $\{\hat{\boldsymbol{t}}_0^i, ..., \hat{\boldsymbol{t}}_n^i\}$ and $\{\boldsymbol{e}_0^i, ..., \boldsymbol{e}_n^i\}$. For the j-th token, the fusion operation is defined as follows:

$$\begin{aligned} \boldsymbol{u}_j^i &= \sigma(\boldsymbol{W}_u^i \hat{\boldsymbol{t}}_j^i + \boldsymbol{W}_e^i \boldsymbol{e}_j^i + \boldsymbol{b}_u^i) \\ \tilde{\boldsymbol{t}}_j^i &= \sigma(\boldsymbol{W}_t^i \boldsymbol{u}_j^i + \boldsymbol{b}_t^i) \end{aligned} \quad (1)$$

where $\boldsymbol{u}_j^i$ represents the hidden state integrating the information from both token and entity. σ is a non-linear activation function. $\boldsymbol{W}_*^i$ and $\boldsymbol{b}_*^i$ are learnable weights and biases respectively. The KG-enhanced token representation $\{\tilde{\boldsymbol{t}}_0^i, ..., \tilde{\boldsymbol{t}}_n^i\}$ is fed into the next layer of knowledge aggregator as input.

2.2 Pre-training Objectives

To incorporate KG knowledge into the language representation learning, KLMo adopts a multi-task loss function as the training objective:

$$\mathcal{L} = \mathcal{L}_{RP} + \mathcal{L}_{EL} + \mathcal{L}_{MLM} \quad (2)$$

In addition to the loss of masked language model $\mathcal{L}_{MLM}$ (Devlin et al., 2018; Li et al., 2020), an relation prediction loss $\mathcal{L}_{RP}$ and an entity linking loss $\mathcal{L}_{EL}$ are integrated to predict the entities in KG based on the corresponding KG-enhanced tokens representations $\{\tilde{\boldsymbol{t}}_0^M, ..., \tilde{\boldsymbol{t}}_n^M\}$.

For each pair of entity spans, we utilize the relation between their corresponding entities in the KG as the distant supervision for relation prediction. The relation prediction and entity linking objectives are the key to the integration of relations and entities in KG into the text. Since the number of entities in KG is quite large for the $Softmax$ operation in entity-linking objective, we handle this problem by only predicting entities in the same batch instead of all entities in KG. To prevent KLMo from completely remembering entity mentions while predicting rather than relying on textual contexts, we randomly mask 10% of entities with a special [MASK] token in the input text.

Modal	Precision	Recall	F1	Acc
BERT	81.76	80.11	80.92	80.06
WKLM	82.71	80.28	81.47	80.17
ERNIE	82.66	81.39	82.02	81.18
KLMo	82.68	**84.33**	**83.50**	**81.75**

Table 1: Results on Entity Typing.

3 Experiments

This section presents the details of KLMo pre-training and its finetuning on two specific knowledge-driven NLP tasks: entity typing and relation classification. We pretrain KLMo by a Chinese corpus of Baidu Baike's webpages and the Baike Knowledge Graph. Details of the pretraining corpus and experimental settings are described in Appendix A. [1]

3.1 Baselines

We compare KLMo with the state-of-the-art PLMs pretrained on the same Baidu Baike corpus: (1) BERT-Base Chinese (Devlin et al., 2018), which is further pretrained on the Baidu Baike corpus for one epoch.(2) ERNIE-THU (Zhang et al., 2019), a pioneering and typical work in this field, which incorporates entity knowledge into the PLM. (3) WKLM (Xiong et al., 2019), a weakly supervised Knowledge-enhanced PLM using entity replacement predictions to incorporate the entity knowledge, which provides the state-of-the-art results on several knowledge-driven tasks.

3.2 Entity Typing

Dataset In this work, we create a Chinese entity typing dataset, which is a completely manually-annotated dataset containing 23,100 sentences and 28,093 annotated entities distributed in 15 fine-grained categories of media works, such as Movie, Show and TV Play. We split the dataset into a training set with 15,000 sentences and a test set with 8,100 sentences. The detail statistics of the dataset

[1]Our code and datasets are available at: https://github.com/lei-nlp/KLMo

Modal	Precision	Recall	F1
CNN	-	-	20.56
BERT	15.94	35.12	21.93
WKLM	16.32	**36.96**	22.64
ERNIE	18.18	34.29	23.76
KLMo	**20.90**	31.24	**25.05**

Table 2: Results on Relation Classification.

Modal	Precision	Recall	F1
BERT	81.76	80.11	80.92
KLMo	82.68	84.33	83.50
w/o KG	82.30	83.02	82.66

Table 3: Ablation study on Entity Typing.

and the finetuning settings are shown in Appendix B.1.

Results We evaluate various pretrained models for entity typing under precision, recall, micro-F1 and accuracy metrics. The results are shown in Table 1. The following observations can be found: (1) All knowledge-enhanced PLMs generally perform much better than the BERT baseline on all measures, which shows that entity knowledge is beneficial to entity type predication with limited annotated resources. (2) Compared with the existing knowledge-enhanced PLMs, KLMo largely improves the recall score over WKLM and ERNIE, leading to an improvement of 1.58 and 0.57 on micro-F1 respectively. This indicates that fine-grained relationships between entities help KLMo to predict appropriate categories for more entities.

3.3 Relation Classification

Dataset The CCKS 2019 Task 3 Inter-Personal Relational Extraction (IPRE) dataset (Han et al., 2020) is used for the evaluation on relation classification. The training set is automatically labeled by distant supervision, and the test set is manually annotated. There are 35 relations (including a null-relation class "*NA*"), where "*NA*" accounts for nearly 86% in the training set and 97% in the test set. The detail statistics of the dataset and finetuning settings are shown in Appendix B.2.

Results We adopt precision, recall and micro-F1 as the evaluation measures. The results are shown in Table 2. In addition to BERT baseline, we also compare KLMo with an official CNN baseline, which gets CNN output as the sentence embedding and feed it into a relation classifier. From Table 2, we can see that both CNN and BERT baseline models do not perform well, which indicates the high difficulty of the dataset. This ascribes to the large number of noisy labels in the training set automatically generated by distant supervision.

Although the dataset are very difficult, we can still observe that: (1) All knowledge-enhanced PLMs largely improve the precision and micro-F1 scores over BERT baseline, which shows that both entity information and KG information can enhance language representations and accordingly prompt the performance of relation classification. (2) KLMo largely improves the precision score over WKLM and ERNIE, leading to an improvement of 2.41 and 1.29 on micro-F1 respectively, which demonstrates that fine-grained relations in KG help KLMo avoid fitting on noisy labels and predict relations correctly.

3.4 Effects of KG Information

Most NLP tasks only provide text inputs and the entity linking itself is a hard task. Thus, we investigate the effects of KG entities and relations for KLMo on entity typing. w/o KG refers to finetuning KLMo without the input of KG entities and relations. Table 3 shows the results of the ablation study. Without KG input for funetuning, KLMo still largely outperforms BERT on both precision and recall scores, leading to an improvement by 1.74 on micro-F1. Compared with KLMo finetuning with KG, KLMo without KG witnesses a small decrease of 0.84 on micro-F1 measure. This demonstrates that KG information has been integrated into KLMo during pre-training. For most specific NLP tasks, KLMo can be finetuned in a similar way as BERT.

4 Conclusion

In this paper, we propose a novel KG-enhanced pretrained language model KLMo to explicitly integrate KG entities and fine-grained relations into the language representation learning. Accordingly, the novel knowledge aggregator is designed to handle the heterogeneous information fusion and text-knowledge alignment problems. Further, the relation prediction and entity linking objectives are jointly optimized to encourage the knowledge information integration. The experiment results show that KLMo outperforms the other state-of-the-art knowledge-enhanced PLMs, which validates the intuition that fine-grained relationships in KG can

enhance the language representation learning and benefit some knowledge-driven NLP tasks.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*, pages 1–9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xianpei Han, Zhichun Wang, Jiangtao Zhang, Qinghua Wen, Wenqi Li, Buzhou Tang, Qi Wang, Zhifan Feng, Yang Zhang, Yajuan Lu, et al. 2020. Overview of the ccks 2019 knowledge graph evaluation track: Entity, relation, event and qa. *arXiv preprint arXiv:2003.03875*.

Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 139–144.

Bin He, Di Zhou, Jinghui Xiao, Qun Liu, Nicholas Jing Yuan, Tong Xu, et al. 2019. Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Knowledge-aware language model pretraining. *arXiv preprint arXiv:2007.00655*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. Kepler: A unified model for knowledge embedding and pre-trained language representation. *arXiv preprint arXiv:1911.06136*.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2019. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *arXiv preprint arXiv:1912.09637*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

A Pre-training Settings

A.1 Pretraining Corpus

Baike Knowledge Graph Baike Knowledge Base is a generic-domain Chinese knowledge base, which contains 226 concept types, more than 100 million entities, and 2.2 billion triples. Each entity in Baike Knowledge Base is aligned to a webpage from a variety of sources, such as Baidu Baike, Sogou Baike and douban. To pretrain a KG-enhanced Chinese language model, we extract a subset of this knowledge base to build a Baike KG using the following rules: 1) removing entities not from Baidu Baike articles; 2) removing low-popular entities (smaller than 200); 3) only keeping fact triples whose both entities are Baidu Baike entities. The final Baike KG contains 2,466,069 entities, 390 relations and 9,859,314 triples.

Chinese Pretraining Corpus KLMo mainly adopts Baidu Baike's webpages for pre-training, which contains encyclopedia articles written in formal Chinese language. Entities in articles can be

Dataset	Entity Typing Dataset			Linked Entity Typing Dataset		
	#Sentences	#Entities	#Types	#Linked_Sentences	#Linked_Entities	Sent_Ratio
Training Set	15,000	18,180	15	7606	11,313	50.7%
Testing Set	8100	9913	15	4165	6268	51.4%

Table 4: Statistics of the Chinese Entity Typing Dataset.

	#Sentences	#Relations	#Linked_Sentences	#Linked_Entities	Sent_Ratio
Training Set	287,351	35	132,739	232,882	46.2%
Testing Set	38,417	35	15,906	27,233	41.4%

Table 5: Statistics of Chinese Relation Classification Dataset.

extracted by anchor links and aligned to Baike KG entities. After preprocessing the corpus, a large formatted dataset containing 7.8B tokens, 174M sentences, 21M entities and 1.2M relations is generated for the pre-training of KLMo. Sentences having less than 5 words or 2 entities are discarded.

A.2 Implementation Details

In the experiment, we first obtain the knowledge representations trained on Baike KG triples by TransE (Bordes et al., 2013) algorithm using the OpenKE toolkit (Han et al., 2018). The representations are used to initialize the entity and relation embeddings in KLMo. The embedding dimension is set to 100 and the epoch number is set to 5000.

As for the pre-training of KLMo, due to the expensive cost of pre-training from scratch, we inherit the parameters of BERT-Base Chinese to initialize the Transformer blocks for token encoding, while the parameters for entity and relation encoding modules are all randomly initialized. The number of text encoder layers L and knowledge aggregator layers M are both 6. The hidden size of token embeddings d_t, knowledge embeddings d_z and entity span embeddings d_e are set to 768, 100 and 100. The number of token-oriented attention heads A_t, KG-oriented attention heads A_z and entity span-level attention heads A_e are set to 12, 4 and 12 respectively. The pre-training of KLMo runs 3 epochs on 4 NVIDIA Tesla V100 (32GB) GPUs with the batch size of 128, the max sequence length of 512 and the learning rate of 5e-5.

B Finetuning Settings

B.1 Entity Typing

To evaluate the performance of KLMo, two knowledge-driven tasks, i.e. entity typing and relation classification, are performed in this work. Given a sentence with an entity mention, the entity typing task is to label the mention with its fine-grained semantic type.

Dataset Entity typing is not a new task. However, to our best knowledge, there is no public benchmark dataset available on Chinese fine-grained entity typing. Therefore, In this work, we create a Chinese entity typing dataset, which is a completely manually-annotated dataset containing 23,100 sentences and 28,093 annotated entities distributed in 15 fine-grained categories of media works, such as Movie, Show and TV Play. We split the dataset into a training set with 15,000 sentences and a test set with 8,100 sentences. The detail statistics of the dataset are shown in Table 4.

Finetuning The Chinese entity typing dataset lacks of KG entity annotations, thus we first use an entity linker tool accompanied with Baike Knowledge Base to recognize entity mentions in sentences and link them to their corresponding Baike KG entities. The statistics of the linked entity typing dataset are shown in Table 4. Over 50% of sentences contain at least one linked KG entity mention in both training set and test set. To finetune KLMo for entity typing, we use the representation of the first token of each entity span to predict its entity type. The model is finetuned for 10 epochs on the training set with the batch size of 128, the max sequence length of 256 and the learning rate of 2e-5.

B.2 Relation Classification

We also compare the results of various pretrained models on the task of relation classification. Given a pair of entities in a sentence, the relation classification task is to determine the relation type between the pair of entities.

Dataset The CCKS 2019 Task 3 Inter-Personal Relational Extraction (IPRE) dataset (Han et al., 2020) is used for the evaluation on relation classification. The training set is automatically labeled by distant supervision, and the test set is manually annotated. There are 35 relations (including a null-relation class "*NA*"), where "*NA*" accounts for nearly 86% in the training set and 97% in the test set. The detail statistics of the dataset are shown in Table 5.

Finetuning The IPRE dataset also lacks of KG entity annotations, and we recognize and link entity mentions to their corresponding Baike KG entities in the same way as we do for the entity typing dataset. The statistics of the linked dataset are shown in Table 5. Over 40% of sentences contain at least one linked KG entity mention. To finetune KLMo for relation classification, we concatenate the representations of the first token of the two candidate entity spans. The model is finetuned for 10 epochs with the batch size of 128, the max sequence length of 256 and the learning rate of 2e-5.