KELM: Knowledge Enhanced Pre-Trained Language Representations with Message Passing on Hierarchical Relational Graphs

Yinquan Lu, Haonan Lu, Guirong Fu, Qun Liu

1 Huawei Technologies Co., Ltd.
2 OPPO Guangdong Mobile Telecommunications Co., Ltd.
3 ByteDance
4 Huawei Noah’s Ark Lab

luyinquan1@huawei.com, luhaonan@oppo.com, fuguirong@bytedance.com, qun.liu@huawei.com

Abstract

Incorporating factual knowledge into pre-trained language models (PLM) such as BERT is an emerging trend in recent NLP studies. However, most of the existing methods combine the external knowledge integration module with a modified pre-training loss and re-implement the pre-training process on the large-scale corpus. Re-pretraining these models is usually resource-consuming, and difficult to adapt to another domain with a different knowledge graph (KG). Besides, those works either cannot embed knowledge context dynamically according to textual context or struggle with the knowledge ambiguity issue. In this paper, we propose a novel knowledge-aware language model framework based on fine-tuning process, which equips PLM with a unified knowledge-enhanced text graph that contains both text and multi-relational sub-graphs extracted from KG. We design a hierarchical relational-graph-based message passing mechanism, which can allow the representations of injected KG and text to mutually update each other and can dynamically select ambiguous mentioned entities that share the same text. Our empirical results show that our model can efficiently incorporate world knowledge from KGs into existing language models such as BERT, and achieve significant improvement on the machine reading comprehension (MRC) tasks compared with other knowledge-enhanced models.

Introduction

Pre-trained language models benefit from the large-scale corpus and can learn complex linguistic representation (Devlin et al. 2019; Liu et al. 2019b; Yang et al. 2020). Although they have achieved promising results in many NLP tasks, they neglect to incorporate structured knowledge for language understanding. Limited by implicit knowledge representation, existing PLMs are still difficult to learn world knowledge efficiently (Poerner, Waltinger, and Schütze 2019; Yu et al. 2020). For example, hundreds of related training samples in the corpus are required to understand the fact “ban means an official prohibition or edict against something” for PLMs.

By contrast, knowledge graphs (KGs) explicitly organize the above fact as a triplet “(ban, hypernyms, prohibition)”. Although domain knowledge can be represented more efficiently in KG form, entities with different meanings share the same text may happen in a KG (knowledge ambiguity issue). For example, one can also find “(ban, hypernyms, moldovan monetary unit)” in WordNet (Miller 1995). Recently, many efforts have been made on leveraging heterogeneous factual knowledge in KGs to enhance PLM representations. These models generally adopt two methods: (1) Injecting pre-trained entity embeddings into PLM explicitly, such as ERNIE (Zhang et al. 2019), which injects entity embeddings pre-trained on a knowledge graph by using TransE (Bordes et al. 2013). (2) Implicitly learning factual knowledge by adding extra pre-training tasks such as entity-level mask, entity-based replacement prediction, etc. (Wang et al. 2020c; Sun et al. 2020). Some studies use both of the above two methods such as CokeBERT (Su et al. 2020).

However, as summarized in Table 1 most of the existing knowledge-enhanced PLMs need to re-pretrain the models based on an additional large-scale corpus, they mainly encounter two problems below: (1) Incorporating external knowledge during pretraining is usually resource-consuming and difficult to adapt to other domains with different KGs. By checking the third column of Table 1 one can see that most of the pretrain-based models use Wiki-related KG as their injected knowledge source. These models also use English Wikipedia as pre-training corpus. They either use an additional entity linking tool (e.g. TAGME (Ferragina and Scaiella 2010)) to align the entity mention in the text to a single mentioned entity in a Wiki-related KG uniquely or directly treat hyperlinks in Wikipedia as entity annotations. These models depend heavily on the one-to-one mapping relationship between Wikipedia corpus and Wiki-related KG, thus they never consider handling knowledge ambiguity issue. (2) These models with explicit knowledge injection usually use algorithms like BILINEAR (Yang et al. 2015) to obtain pre-trained KG embeddings, which contain information about graph structure. Unfortunately, their knowledge context is...
few works [Qiu et al. 2019; Yang et al. 2019] concentrate on injecting external knowledge based on fine-tuning PLM on downstream tasks, which is much easier to change the injected KGs and adapt to relevant domain tasks. They either cannot consider multi-hop relational information, or struggle with knowledge ambiguity issue. How to fuse heterogeneous information dynamically based on the fine-tuning process on the downstream tasks and use the information of injected KGs more efficiently remains a challenge.

To overcome the challenges mentioned above, we propose a novel framework named KELM, which injects world knowledge from KGs during the fine-tuning phase by building a Unified Knowledge-enhanced Text Graph (UKET) that contains both injected sub-graphs from external knowledge and text sentence. The method extends the input sentence by extracting sub-graphs centered on every mentioned entity from KGs. In this way, we can get a Unified Knowledge-enhanced Text Graph as shown in Fig. 1 which is made of three kinds of graphs: (1) The injected knowledge graphs, referred to as the “KG only” part; (2) The graph about entity mentions in the text and mentioned entities in KGs, referred to as the “entity link to token” part. Entity mentions in the text are linked with mentioned entities in KGs by string matching, so one entity mention may trigger several mentioned entities that share the same text in the injected KGs (e.g. “Ford” in Fig. 1). (3) The “text only” part, where the input text sequence is treated as a fully-connected word graph just like classical Transformer architecture [Vaswani et al. 2017].

Based on this unified graph, we design a novel Hierarchical relational-graph-based Message Passing (HMP) mechanism to fuse heterogeneous information on the output layer of PLM. The implementation of HMP is via a Hierarchical Knowledge Enhancement Module as depicted in Fig. 2 which also consists of three parts, and each part is designed for solving the different problems above: (1) For reserving the structure information and dynamically embedding injected knowledge, we utilize a relational GNN (e.g. rGCN [Schlichtkrull et al. 2017]) to aggregate and update representations of extracted sub-graphs for each injected KG (corresponding to the “KG only” part of UKET). All men-

Table 1: A brief summary and comparison of recent knowledge-enhanced PLMs. The full names of some abbreviations are as follows. MLM: masked language model, NSP: next sentence prediction, Ent: entity, Rel: relation, CLS: classification, Sent: sentence, ATT: attention. Comments/descriptions of features are written in parentheses. Desired properties are written in bold.
tioned entities and their K-hop neighbors in sub-graphs are initialized by pre-trained vectors obtained from the classical knowledge graph embedding (KGE) method (we adopt BILINEAR here). In this way, knowledge context can be dynamically embedded, the structural information about the graph is also kept; (2) For handling knowledge ambiguity issue and selecting relevant mentioned entities according to the input context, we leverage a specially designed attention mechanism to weight these ambiguous mentioned entities by using the textual representations of words/tokens to query the representations of their related mentioned entities in KGs (corresponding to the “entity link to token” graph of UKET). The attention score can help to select knowledge according to the input sentence dynamically. By concatenating the outputs of this step with the original outputs of PLM, we can get a knowledge-enriched representation for each token; (3) For further interactions between knowledge-enriched tokens, we employ a self-attention mechanism that operates on the fully-connected word graph (corresponding to the “text only” graph of UKET) to allow the knowledge-enriched representation of each token to further interact with others.

We conduct experiments on the MRC task, which requires a system to comprehend a given text and answer questions about it. In this paper, to prove the generalization ability of our method, we evaluate KELM on both the extractive-style MRC task (answers can be found in a span of the given text) and the multiple-response-items-style MRC task (each question is associated with several choices for answer-options, the number of correct answer-options is not pre-specified). MRC is a challenging task and represents a value path towards natural language understanding (NLU). With the rapid increment of knowledge, NLU becomes more difficult since the system needs to absorb new knowledge continuously. Pre-training models on large-scale corpus is inefficient. Therefore, fine-tuning the knowledge-enhanced PLM on the downstream tasks directly is crucial in the application.

Related Work

Knowledge Graph Embedding

We denote a directed knowledge graph as $G(\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ and $\mathcal{R}$ are sets of entities and relations, respectively. We also define $\mathcal{F}$ as a set of facts, a fact stored in a KG can be expressed as a triplet $(h, r, t) \in \mathcal{F}$, which indicates a relation $r$ pointing from the head entity $h$ to tail entity $t$, where $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$. KGE aims to extract topological information in KG and to learn a set of low-dimensional representations of entities and relations by knowledge graph completion task (Yang et al. 2015; Lu and Hu 2020).

Multi-relational Graph Neural Network

Real-world KGs usually include several relations. However, traditional GNN models such as GCN (Kipf and Welling 2017), and GAT (Velčišković et al. 2018) can only be used in the graph with one type of relation. Schlichtkrull et al. (2017) Haonan et al. (2019) generalizes traditional GNN models by performing relation-specific aggregation, making it possible to encode relational graphs. The use of multi-relational GNN makes it possible to encode injected knowledge embeddings dynamically in SKG and CokeBERT.

2Our anonymous code will be available at https://github.com/nlp-anonymous-happy/anonymous-KG-guided-NLP.
Joint Language and Knowledge Models

Since BERT was published in 2018, many efforts have been made for further optimization, basically focusing on the design of the pre-training process and the variation of the encoder. For studies of knowledge-enhanced PLMs, they also fall into the above two categories or combine both of them sometimes. Despite their success in leveraging external factual knowledge, the gains are limited by computing resources, knowledge ambiguity issue, and the expressivity of their methods for the fusion of heterogeneous information, as summarized in Table 1 and the introduction part.

Recent studies notice that the architecture of Transformer treats input sequences as fully-connected word graphs, thus some of them try to integrate injected KGs and textual context into a unified data structure. Here we argue that UKET in our KELM is different from the WK graph proposed in CoLAKE/K-BERT. These two studies heuristically convert textual context and entity-related sub-graph into input sequences, both entities and relations are treated as input words of the PLM, then they leverage a Transformer with a masked attention mechanism to encode those sequences from the embedding layer and pre-train the model based on the large-scale corpus. Unfortunately, it is not trivial for them to convert the second or higher order neighbors related to textual context (Su et al. 2020), the structural information about the graph is lost. UKET differs from the WK graph of CoLAKE/K-BERT in that, instead of converting mentioned entities, relations, and text into a sequence of words and feeding them together into the input layer of PLM (they unify text and KG into a sequence), UKET unifies text and KG into a graph. Besides, by using our UKET framework, the knowledge fusion process of KELM is based on the representation of the last hidden layer of PLM, making it possible to directly fine-tune the PLM on the downstream tasks without re-pretraining the model. SKG also utilizes relational GNN to fuse information of KGs and text representation encoded by PLM. However, SKG only uses GNN to dynamically encode the injected KGs, which corresponds to part one of Fig. 1 Outputs of SKG are made by directly concatenating outputs of graph encoder with the outputs of PLM. It cannot select ambiguous knowledge and forbids the interactions between knowledge-enriched tokens corresponding to part two and part three of Fig. 1 respectively. KT-NET uses a specially designed attention mechanism to select relevant knowledge from KGs. For example, it treats all synsets of entity mentions within the WN18 as candidate KB concepts. This limits the ability of KT-NET to select the most relevant mentioned entities. Moreover, the representations of injected knowledge are static in KT-NET, they cannot dynamically change according to textual context, the information about the original graph structure in KG is also lost.

Methodology

The architecture of KELM is shown in Fig. 2. It consists of three main modules: (1) PLM Encoding Module; (2) Hierarchical Knowledge Enhancement Module; (3) Output Module.

PLM Encoding Module

This module utilizes PLM (e.g., BERT) to encode text to get textual representations for passages and questions. An input example of the MRC task includes a paragraph and a question with a candidate answer, represented as a single sequence of tokens of the length $n$: $T=\{[CLS], Q, (A), [SEP], P, [SEP]\}={t_i}_{i=1}^{n}$, where $Q$, $A$ and $P$ represent all tokens for question, candidate answer and paragraph, respectively $[SEP]$ and $[CLS]$ are special tokens in BERT and defined as a sentence separator and a classification token, respectively. $i$-th token in the sequence is represented by $\vec{t_i} \in \mathbb{R}^{d_t}$, where $d_t$ is the last hidden layer size of used PLM.

Hierarchical Knowledge Enhancement Module

This module is the implementation of our proposed HMP framework, the knowledge fusion process of KELM is different from the WK graph proposed in CoLAKE/K-BERT. These two studies heuristically convert textual context and entity-related sub-graph into input sequences, both entities and relations are treated as input words of the PLM, then they leverage a Transformer with a masked attention mechanism to encode those sequences from the embedding layer and pre-train the model based on the large-scale corpus. Unfortunately, it is not trivial for them to convert the second or higher order neighbors related to textual context (Su et al. 2020), the structural information about the graph is lost. UKET differs from the WK graph of CoLAKE/K-BERT in that, instead of converting mentioned entities, relations, and text into a sequence of words and feeding them together into the input layer of PLM (they unify text and KG into a sequence), UKET unifies text and KG into a graph. Besides, by using our UKET framework, the knowledge fusion process of KELM is based on the representation of the last hidden layer of PLM, making it possible to directly fine-tune the PLM on the downstream tasks without re-pretraining the model. SKG also utilizes relational GNN to fuse information of KGs and text representation encoded by PLM. However, SKG only uses GNN to dynamically encode the injected KGs, which corresponds to part one of Fig. 1 Outputs of SKG are made by directly concatenating outputs of graph encoder with the outputs of PLM. It cannot select ambiguous knowledge and forbids the interactions between knowledge-enriched tokens corresponding to part two and part three of Fig. 1 respectively. KT-NET uses a specially designed attention mechanism to select relevant knowledge from KGs. For example, it treats all synsets of entity mentions within the WN18 as candidate KB concepts. This limits the ability of KT-NET to select the most relevant mentioned entities. Moreover, the representations of injected knowledge are static in KT-NET, they cannot dynamically change according to textual context, the information about the original graph structure in KG is also lost.

---

1 A subset of WordNet.
2 Refer the example given in the case study of KT-NET, the most relevant concept for the word “ban” is “forbidding_NN_1” (with the probability of 86.1%), but not “ban_NN_4”.
3 Depending on the type of MRC task (extractive-style v.s. multiple-response-items-style), candidate answer A is not required in the sequence of tokens for the extractive-style MRC task.
we define the full hierarchical graph consisting of all three parts \( \{G^q_s\}_{q=1}^{|Q|}, \{G^m_q\}_{q=1}^{|Q|} \), and \( G_t \), as Unified Knowledge-enhanced Text Graph (UKET).

**Dynamically Embedding Knowledge Context** We use pre-trained vectors obtained from the KGE method to initialize representations of entities in \( G^q_s(e, R^q_s) \). Considering the structural information of injected knowledge graph forgotten during training, we utilize \( |Q| \) independent GNN encoders (i.e., \( q_1, \ldots, q_\|Q\| \) in Fig. 2) which is the case of injecting two independent KGs in our experiment setting to dynamically update entity embeddings of \( |Q| \) injected KGs. We use rGCN to model the multi-relational nature of the knowledge graph. To update \( i \)-th node of \( q \)-th KG in \( l \)-th rGCN layer:

\[
\hat{s}^{q,(l+1)}_i = \sigma \left( \sum_{r \in \mathcal{N}_r^i} \sum_{j \in \mathcal{N}_j^r} \frac{1}{|\mathcal{N}_i^j|} W_r^{q,(l)} \hat{s}^{q,(l)}_j \right)
\]

Where \( \mathcal{N}_r^i \) is a set of neighbors of \( i \)-th node under relation \( r \in \mathcal{T}^q \), \( W_r^{q,(l)} \) is trainable weight matrix at \( l \)-th layer and \( \hat{s}^{q,(l+1)}_i \) is the hidden state of \( i \)-th node at \((l+1)\)-th layer. After \( L \) updates, \(|Q|\) sets of embeddings are obtained. The output of the \( q \)-th KG can be represented as \( S^q \in \mathbb{R}^{|E^q| \times d_q} \), where \( |E^q| \) and \( d_q \) are the numbers of nodes of extracted subgraph and the dimension of pre-trained KGE, respectively.

**Dynamically Selecting Semantics-Related Mentioned Entities** To handle the knowledge ambiguity issue, we introduce an attention layer to weight these ambiguous mentioned entities by using the textual representations of tokens (outputs of Section 3.1) to query their semantics-related mentioned entities representations in KGs. Here, we follow the attention mechanism of GAT to update each entity mention embedding in \( G^{q,\ell}_m \):

\[
\tilde{x}^{q}_j = \sigma \left( \sum_{j \in \mathcal{N}_j^q} \alpha^{q,\ell}_{ij} W_q \tilde{s}^{q}_j \right)
\]

Where \( \tilde{x}^{q}_j \) is the output embeddings from the \( q \)-th rGCN in the previous step. \( \tilde{x}^{q}_j \) is the hidden state of \( i \)-th entity mention \( x^{q}_j \) in \( \mathcal{X}^q \), and \( \mathcal{N}_j^q \) is a set of neighbors of \( x^{q}_j \) in \( G^q_m \). \( W_q \in \mathbb{R}^{d_{a^q} \times d_q} \) is a trainable weight matrix, we set \( d_{a^q} = d_{a^m} = d_q \) (thus \( \tilde{x}^{q}_j \in \mathbb{R}^{d_q} \)), \( \sigma \) is a nonlinear activation function. \( \alpha^{q,\ell}_{ij} \) is the attention score that weights ambiguous mentioned entities in the \( q \)-th KG:

\[
\alpha^{q,\ell}_{ij} = \frac{\exp(\text{LeakyReLU}(\tilde{s}^{q}_i W_q \tilde{s}^{q}_j)))}{\sum_{k \in \mathcal{N}_k^q} \exp(\text{LeakyReLU}(\tilde{s}^{q}_i W_q \tilde{s}^{q}_k)))}
\]

The representation \( \tilde{h}^{q}_t \) with a dimension of \( d_q \) is projected to the dimension of \( d_{exp} \), before using it to query the related mentioned entity embeddings of \( S^q \): \( \tilde{h}^{q}_t = W^{q}_{\text{proj}} \tilde{h}^{q}_t \), where \( W^{q}_{\text{proj}} \in \mathbb{R}^{d_{exp} \times d_q} \). \( \sigma^q \in \mathbb{R}^{2d_q} \) is a trainable weight vector.

\[ T \] is the transposition operation and \( || \) is the concatenation operation.

Finally, we concatenate outputs of \( |Q| \) KGs with textual context representation to get final knowledge-enriched representation:

\[
\tilde{h}^{q}_t = [\tilde{h}^{q}_{t,1}^{q}, \ldots, \tilde{h}^{q}_{t,|Q|}] \in \mathbb{R}^{d_{exp} + d_1 + \cdots + d_{|Q|}}
\]

If token \( t_i \) can’t match any entity in \( q \)-th KG (say \( t_i \notin \mathcal{X}^q \)), we fill \( \hat{x}^{q}_i \) in Eq[3] with zeros. Note that mentioned entities in KGs are not always useful, to prevent noise, we follow [Yang and Mitchell 2017]’s work and add an extra sentinel node linked to each entity mention in \( G^q_t \). The sentinel node is initialized by zeros and not trainable, which is the same as the case of no retrieved entities in the KG. In this way, according to the textual context, KELM can dynamically select mentioned entities and avoid introducing knowledge noise.

**Interaction Between Knowledge-enriched Token Embeddings** To allow knowledge-enriched tokens’ representations to propagate to each other in the text, we use a fully-connected word graph \( G_t \), with knowledge-enriched representations from outputs of the previous step, and employ the self-attention mechanism similar to KT-NET to update token’s embedding. The final representation for \( i \)-th token in the text is \( \tilde{h}^{q}_t \in \mathbb{R}^{d_{exp} + d_1 + \cdots + d_{|Q|}} \).

**Output Module**

**Extractive-style MRC task** A simple linear transformation layer and softmax operation are used to predict start and end positions of answers. For \( i \)-th token, the probabilities to be the start and end position of answer span are:

\[
p^s_i = \frac{\exp(w^T_i \tilde{h}^{q}_t)}{\sum_{j=1}^{|Q|} \exp(w^T_j \tilde{h}^{q}_t)} \quad p^e_i = \frac{\exp(w^T_i \tilde{h}^{q}_t)}{\sum_{j=1}^{|Q|} \exp(w^T_j \tilde{h}^{q}_t)},
\]

where \( w_s, w_e \in \mathbb{R}^{d_{exp} + d_1 + \cdots + d_{|Q|}} \) are trainable vectors and \( n \) is the number of tokens. The training loss is calculated by the log-likelihood of the true start and end positions:

\[
L = -\frac{1}{N} \sum_{i=1}^{N} (\log p^s_{y^s_i} + \log p^e_{y^e_i}),
\]

where \( N \) is the total number of examples in the dataset, \( y^s_i \) and \( y^e_i \) are the true start and end positions of \( i \)-th query’s answer, respectively. During inference, we pick the span \((a, b)\) with maximum \( p^s_a p^e_b \) where \( a \leq b \) as predicted answer.

**Multiple-response-items-style MRC task** Since answers to a given question are independent of each other, to predict the correct probability of each answer, a fully connected layer followed by a sigmoid function is applied on the final representation of [CLS] token in BERT.

**Experiments**

**Datasets**

In this paper, we empirically evaluate KELM on both two types of MRC benchmarks in SuperGLUE [Wang et al. 2020a]: ReCoRD [Zhang et al. 2018] (extractive-style) and MultiRC [Khashabi et al. 2018] (multiple-response-items-style). Detailed descriptions of the two datasets can be found in Appendix A. On both datasets, the test set is not public, one has to submit the predicted results to the organization to get the final test score. Since frequent submissions to probe the unseen test set are not encouraged, we only submit our best model once for each of the datasets, thus the statistics of the results (e.g., mean, variance, etc.) are not applicable. We use Exact Match (EM) and (macro-averaged) F1 as the evaluation metrics.
External Knowledge We adopt knowledge sources the same as used in KT-NET: WordNet and NELL (Carlson et al. 2010). Representations of injected knowledge are initialized by resources provided by (Yang and Mitchell 2017), where the embeddings of WordNet were pre-trained on a subset consisting of 151,442 triplets with 40,943 synsets and 18 relations, and the embeddings of NELL were pre-trained on a subset with 180,107 entities and 258 concepts. The size of these embeddings is 100. We retrieve related knowledge from the two KGs in a given sentence: (1) For WordNet, we look up mentioned entities in the WordNet by string matching operation, and link all tokens in the same word to the retrieved mentioned entities (tokens are tokenized by Tokenizer of BERT). Then, we extract all 1-hop neighbors for each mentioned entity and construct sub-graphs. In this paper, our experiment results are based on the 1-hop case. However, our framework can be generalized to multi-hop easily, and we leave this for future work. (2) For NELL, we link entity mentions to the whole KG, and return associated concepts.

Experimental Setups

Baselines and Comparison Setting Because we use BERTlarge as the base model in our method, we use it as our primary baseline for all tasks. For fair comparison, we mainly compare our results with two fine-tune-based knowledge-enhanced models: KT-NET and SKG, which also evaluate their results on ReCoRD with BERTlarge as the encoder part. As mentioned in the original paper of KT-NET, KT-NET mainly focuses on the extractive-style MRC task. We also evaluate KT-NET on the multiple-response-items-style MRC task and compare the results with KELM. We evaluate our approach in three different KB settings: KELMWordNet, KELMNELL, and KELMBoth, to inject KG from WordNet, NELL, and both of the two, respectively (The same as KT-NET). Implementation details of our model are presented in Appendix B.

| Model                | Dev   | Test  |
|----------------------|-------|-------|
| BERTlarge            | 70.2  | 71.3  |
| SKG+BERTlarge        | 70.9  | 72.2  |
| KT-NETWordNet        | 70.6  | -     |
| KT-NETNELL           | 70.5  | -     |
| KT-NETBoth           | 71.6  | 74.0  |
| KELMWordNet          | 75.4  | 75.9  |
| KELMNELL             | 74.8  | 75.5  |
| KELMBoth             | 75.1  | 76.2  |

Table 2: Performance comparison on ReCoRD. The effectiveness of injecting external knowledge are shown.

Results

The results for the extractive-style MRC task and multiple-response-items-style MRC task are given in Table 2 and Table 3, respectively. The scores of other models are taken directly from the leaderboard of SuperGLUE and literature (Qiu et al. 2019; Yang et al. 2019). In this paper, our implementation is based on a single model, and hence comparing with ensembles based models is not considered. Best results are labeled in bold and the second best are underlined.

Results on the dev set of ReCoRD show that: (1) KELM outperforms BERTlarge, irrespective of which external KG is used. Our best KELM offers a 5.2/3.7 improvement in EM/F1 over BERTlarge. (2) KELM outperforms previous SOTA knowledge-enhanced PLM (KT-NET) by +3.8 EM/+2.3 F1. In addition, KELM outperforms KT-NET significantly in all three KB settings. On the dev set of MultiRC, the best KELM offers a 3.6 improvement in EM over KT-NET. Although the performance on F1 drop a little compared with KT-NET, we still get a gain of +2.9 (EM+F1) over the former SOTA model.

Results on the test set further demonstrate the effectiveness of KELM and its superiority over the previous works. On ReCoRD, it significantly outperforms the former SOTA knowledge-enhanced PLM (finetuning based model) by +3.2 EM/+1.9 F1. And on MultiRC, KELM offers a 3.1/0.8 improvement in EM/F1 over BERTlarge, and achieves a gain of +1.5 (EM+F1) over KT-NET.

Case Study

This section uses an example in ReCoRD to show how KELM avoids knowledge ambiguity issue and selects the most relevant mentioned entities adaptively w.r.t the textual context. Recall that given a token $t_i$, the importance of a mentioned entity $m^q_j$ in $q$-th KG is scored by the attention weight $a^q_{ij}$ in Eq. 2. To illustrate how KELM can select the most relevant mentioned entities, we analyze the example that was also used in the case study part of KT-NET.

The question of this example is “Sudan remains a XXX-designated state sponsor of terror and is one of six countries subject to the Trump administration’s ban”, where the “XXX” is the answer that needs to be predicted. The case study in KT-NET shows the top 3 most relevant concepts from WordNet for the word “ban” are “forbidding.n.01”, “proscription.n.01”, and “ban.v.02”, with the weights of 3.1/0.8. However, our framework can be generalized to multi-hop easily, and we leave this for future work.
Table 4: Case study. Comparisons between the golden label with the most relevant mentioned entity in WordNet. The importance of selected mentioned entities is provided in the parenthesis.

| Word in text (prototype) | The most relevant mentioned entity in WordNet (predicted) | Golden mentioned entity |
|-------------------------|----------------------------------------------------------|-------------------------|
| ford                    | ford.n.05 (0.56)                                         | ford.n.05               |
| pardon                  | pardon.v.02 (0.86)                                        | pardon.v.02             |
| nixon                   | nixon.n.01 (0.74)                                         | nixon.n.01              |
| lead                    | lead.v.03 (0.73)                                          | lead.v.03               |
| outrage                 | outrage.n.02 (0.62)                                       | outrage.n.02            |

Table 5 of Appendix.

Further Analysis and Discussion

KELM incorporates knowledge in KGs into the representations in the last hidden layer of PLM (Refer to Methodology Section). It is essentially a model-agnostic, KG-agnostic, and task-agnostic framework for enhancing language model representations with factual knowledge from KGs. It can be used to enhance any PLM, with any injected KGs, on any downstream task. Besides the two Q&A-related MRC tasks we mentioned in the main text, we also evaluate KELM on COPA and SQuAD 1.1 based on BERTlarge, results are presented in Appendix D and Appendix F, respectively. To demonstrate KELM is a model-agnostic framework, we also implement KELM based on RoBERTa large and evaluate it on ReCoRD. The experiment is presented in Appendix E. Improvements achieved by KELM over all vanilla base PLM models indicate the effectiveness of injecting external knowledge.

However, the improvements of KELM over RoBERTa on ReCoRD and BERT on SQuAD 1.1 are marginal compared with the ones on ReCoRD/MultiRC/COPA (BERTlarge based). The reason behind this is that pretrained model on in-domain unlabeled data could boost performance on downstream tasks. Passages in ReCoRD are collected from articles in CNN/Daily Mail, while BERT is pre-trained on BookCorpus and English Wikipedia. RoBERTa not only uses the corpus that used in BERT (16 GB), but also an additional corpus collected from the CommonCrawl News dataset (76 GB). ReCoRD is in-domain for RoBERTa but is out-of-domain for BERT. Similarly, SQuAD 1.1 is created from Wikipedia, it is an in-domain downstream task for both BERT and RoBERTa. This partially explains why RoBERTa achieves a much larger improvement over BERT on the result of ReCoRD (71.3 → 88.4 in EM on test set) than the one on SQuAD 1.1 (84.1 → 88.9). A similar analysis can be also found in T5 (Raffel et al. 2020). From our empirical results, we can summarize that general KG (e.g. WordNet) can not help too much for the PLMs pretrained on in-domain data. But it can still improve the performance of the model when the downstream tasks are out-of-domain. Further detailed analysis can be found in our appendix.

Finding a popular NLP task/dataset that is not related to the training corpus of modern PLMs is difficult. Pre-training on large-scale corpus is always good if we have unlimited computational resources and plenty of in-domain corpus. It has been evident that the simple finetuning of PLM is not sufficient for domain-specific applications. KELM can provide people another choice when they do not have such a large-scale in-domain corpus and want to incorporate incremental domain-related structural knowledge into the domain-specific applications.

Conclusion

In this paper, we have proposed KELM for MRC, which enhances PLM representations with structured knowledge from KGs based on the fine-tuning process. Via a unified knowledge-enhanced text graph, KELM can embed the injected knowledge dynamically, and select relevant mentioned entities in the input KGs. In the empirical analysis, KELM shows the effectiveness of fusing external knowledge into representations of PLM and demonstrates the ability to avoid knowledge ambiguity issue. Injecting emerging factual knowledge into PLM during finetuning without re-pretraining the whole model is quite important in the application of PLMs and is still barely investigated. Improvements achieved by KELM over vanilla baselines indicate a

*A commonsense causal reasoning task in SuperGLUE

*An extractive-style MRC task.
potential direction for future research.

Appendix A

Dataset Details

ReCoRD (an acronym for the Reading Comprehension with Commonsense Reasoning Dataset) is a large-scale dataset for extractive-style MRC requiring commonsense reasoning. There are 100,730, 10,000, and 10,000 examples in the training, development (dev), and test set, respectively. An example of the ReCoRD consists of three parts: passage, question, and answer. The passage is formed by the first few paragraphs of an article from CNN or Daily Mail, with named entities recognized and marked. The question is a sentence from the rest of the article, with a missing entity specified as the golden answer. The model needs to find the golden answer among the entities marked in the passage. Questions that can be easily answered by pattern matching are filtered out. By the design of the process of data collection, one can see that to answer the questions, external background knowledge and ability of reasoning are required.

MultiRC (Multi-Sentence Reading Comprehension) is a multiple-response-items-style MRC dataset of short paragraphs and multi-sentence questions that can be answered from the content of the paragraph. Each example of MultiRC includes a question that associates with several choices for answer-options, and the number of correct answer-options is not pre-specified. The correct answer is not required to be a span in the text. The dataset consists of 10K questions (6k multiple-sentence questions), about 60% of this data make training/dev data. Paragraphs in the dataset have diverse provenance by being extracted from 7 different domains such as news, fiction, historical text etc., and hence are expected to be more complicated in their contents as compared to single-domain datasets.

Appendix B

Implementation Details

Our implementation is based on HuggingFace (Wolf et al., 2020) and DGL (Wang et al., 2020b). For all three settings of KELM, parameters of the encoding layer of BERT$_{large}$ are initialized with pre-trained model released by Google. Other trainable parameters in HMP are randomly initialized. The total number of trainable parameters of KELM is 340.4M (Roughly the same as BERT$_{large}$, which has 340M parameters). Since including all neighbors around mentioned entities in WordNet is not efficient, for simplicity, we use top 3 most common relations in WordNet in our experiment (i.e. hyponym, hypernym, derivationally related forms). For both datasets, we use a “two stage” fine-tune strategy to achieve our best performance, the FullTokenizer built in BERT is used to segment input words into wordpieces.

For ReCoRD, the maximum length of answer during inference is set to 30, and the maximum length of question is set to 64. Questions longer than that are truncated. The maximum length of input sequence $T$ is set to 384. Input sequences longer than that are segmented into chunks with a stride of 128. Fine-tuning our model on ReCoRD costs about 18 hours on 4 V100 GPU with a batch size of 48. We freeze parameters of BERT and use Adam optimizer with a learning rate of 1e-3 to train our knowledge module in the first stage. The maximum number of training epochs of the first stage is 10. The purpose of this is to provide a good weight initialization for our HMP. For the second stage, the pre-trained BERT parameters and our HMP part will be fine-tuned together. The max number of training epochs is chosen from $\{4, 6, 8\}$. The learning rate is set to be $2e-5$ with a warmup over the first 6% of max steps, and linear decay until up to max epochs. For both two stages, early stopping is applied according to the best EM+F1 score on the dev set every 500 steps.

For MultiRC, the maximum length of input sequence $T$ is set to 256. The summation of length of question ($Q$) and length of candidate answer ($A$) is not limited. Paragraph ($P$) is truncated to fit the maximum length of input sequence. Fine-tuning KELM on MultiRC needs about 12 hours on 4 V100 GPU with a batch size of 48. For the first stage fine-tuning, learning rate is $1e-4$ and the maximum number of training epochs is 10. For the second stage, the max number of training steps is chosen from $\{10000, 15000, 20000\}$. The learning rate is set to be $2e-5$ with a warmup over the first 10% of max steps.

Appendix C

Supplementation of the Case Study Section

We provide definitions of the top 3 most relevant mentioned entities in WordNet that correspond to the word examples mentioned in Case Study Section. Descriptions are obtained by using NLTK (Loper and Bird, 2002). By comprehending the motivating example in the case study section, we can see that KELM can correctly select the most relevant mentioned entities in the KG.

Appendix D

Experiment on Commonsense Causal Reasoning Task

To further explore the generalization ability of KELM, we also evaluate our method on COPA (Roemmele, Bejan, and Gordon, 2011) (Choice of Plausible Alternatives), which is also a benchmark dataset in SuperGLUE and can be used for evaluating progress in open-domain commonsense causal reasoning. COPA consists of 1000 questions, split equally into development and test sets of 500 questions each. Each question is composed of a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. Similar to the previous two MRC tasks, the development set is publicly available, but the test set is hidden. One has to submit the predicted results for the test set to SuperGLUE to retrieve the final test score. Since the implementation of KELM is based on BERT$_{large}$, we use it as our baseline for the comparison. The result of BERT$_{large}$ is directly taken from the leaderboard of SuperGLUE. Table 6 shows the experiment results. The injected KG is WordNet here, and we use accuracy as the evaluation metric.
Table 5: Definitions of mentioned entities in WordNet corresponding to the word examples in the case study. The importance of mentioned entities is provided in the parenthesis. “sentinel” is meaningless, which is used to avoid knowledge noise.

| Word in text (prototype) | Mentioned entity in WordNet | Definition |
|-------------------------|-----------------------------|------------|
| ban                     | ban.n.04 (0.72)             | an official prohibition or edict against something |
|                         | ban.v.02 (0.21)             | prohibit especially by legal means or social pressure |
|                         | ban.v.03 (0.06)             | forbid the public distribution of a movie or a newspaper |
| ford                    | ford.n.09 (0.56)            | 38th President of the United States; appointed vice president and succeeded Nixon when Nixon resigned (1913-) |
|                         | ford.n.07 (0.24)            | a shallow area in a stream that can be forded |
|                         | ford.v.01 (0.08)            | cross a river where it’s shallow |
| pardon                  | pardon.v.02 (0.36)          | a warrant granting release from punishment for an offense |
|                         | pardon.n.02 (0.04)          | grant a pardon to |
| nixon                   | nixon.n.01 (0.74)           | vice president under Eisenhower and 37th President of the United States; resigned after the Watergate scandal in 1974 (1913-1994) |
|                         | nixon.sent (0.26)           | - |
| lead                    | lead.v.03 (0.73)            | tend to or result in |
|                         | lead.n.03 (0.12)            | evidence pointing to a possible solution |
|                         | lead.v.04 (0.05)            | travel in front of; go in advance of others |
| outrage                 | outrage.n.02 (0.62)         | a wantonly cruel act |
|                         | outrage.sent (0.38)         | - |

Table 6: Performance comparison on COPA. The effectiveness of injecting external knowledge (WordNet) are shown.

The huge improvement over the baseline in this task demonstrates that knowledge in WordNet is indeed helpful for BERT to improve the generalization ability to the out-of-domain downstream task.

Appendix E

KELM: A framework of finetune-based model-agnostic knowledge-enhanced PLM

We implement KELM based on the RoBERTa_{large}, which has a similar number of trainable parameters as BERT_{large} but uses nearly 10 times of training corpus than BERT_{large}. Since the performances of RoBERTa on the leaderboard of SuperGLUE are based on ensembling, we also fine-tune RoBERTa_{large} on ReCoRD to produce the results of a single model. Comparisons of the results can be found in Table 7, where you can also see an improvement there. However, that improvement is not as significant as we observed in BERT_{large}. Reasons are two-fold: (1) Passages in ReCoRD are collected from articles in CNN/Daily Mail, while BERT is pre-trained on BookCorpus and English Wikipedia. RoBERTa not only uses the corpus that used in BERT (16 GB), but also an additional corpus collected from the CommonCrawl News dataset (76 GB). ReCoRD dataset is in-domain for RoBERTa but is out-of-domain for BERT. It seems that the improvements of KELM with injecting general KGs (e.g. WordNet) on the in-domain downstream tasks are not as large as the out-of-domain downstream tasks. A similar phenomenon can be also observed in the experiment of SQuAD 1.1 (Refer to Appendix E).

(2) The same external knowledge (WordNet, NELL) can not help RoBERTa_{large} too much, since RoBERTa is pre-trained on a much larger corpus than BERT. knowledge in WordNet/NELL has been learned in RoBERTa.

We also list the results of LUKE (Yamada et al. 2020) in Table 7. LUKE is a pretrain-based knowledge enhanced PLM and uses Wiki-related golden entities (one-to-one mapping) as the injected knowledge source (about 500k entities\textsuperscript{11}). It has more 128 M parameters than the total number of parameters of the vanilla RoBERTa. As we summarized in Table 1 in the main text, the pre-training task is also different compared with RoBERTa. Although LUKE gets better results compared with vanilla RoBERTa and KELM, it needs 16 NVIDIA Tesla V100 GPUs and the training takes approximately 30 days. Relying on hyperlinks in Wikipedia as golden entity annotations, lacking the flexibility to adapt the external knowledge of other domains, and needing re-pretraining when incorporating knowledge, these limitations hinder the abilities of applications.

Appendix F

Experiment on SQuAD 1.1

SQuAD 1.1 (Rajpurkar et al. 2016) is a well known extractive-style MRC dataset that consists of questions created by crowdworkers for Wikipedia articles. It contains 100,000+ question-answer pairs on 536 articles. We implement KELM based on the BERT_{large}, and compare our results on the development set of SQuAD 1.1 with KT-NET (Best result of KT-NET is based on injecting WordNet only). Results are shown in Table 8.

Results on KELM show an improvement over vanilla BERT. Both BERT and RoBERTa use English Wikipedia as the corpus for pretraining. Since SQuAD is also created from Wikipedia, it is an in-domain downstream task for both BERT and RoBERTa (while ReCoRD dataset is in-domain\textsuperscript{13}). For KELM, we only use 40943 entities in WordNet and 258 concepts in NELL.

Table 7: Comparison of the effectiveness of injecting external knowledge between BERT and RoBERTa. [*] Results are from our implementation.

| Model               | Dev | Test |
|---------------------|-----|------|
| BERT_{large}        |     | 70.6 |
| KELM_{RoBERTa_{large}} | 76.1| 78.0 |

Table 8: Results on KELM, BERT, and RoBERTa on the development set of SQuAD 1.1.
for RoBERTa but is out-of-domain for BERT). This explains why RoBERTa achieves a much larger improvement over BERT on the result of ReCoRD (71.3 → 88.4 in EM on test set) than the one on SQuAD 1.1 (84.1 → 88.9). The rest of the improvement is because RoBERTa uses 10 times of training corpus than BERT and different pre-training strategies they used.

Interestingly, we find the performance of KELM on SQuAD 1.1 is sub-optimal compared with KT-NET. As we mentioned in the last paragraph of the Related Work Section, KT-NET treats all synsets of entity mentions within the WN18 as candidate KB concepts. Via a specially designed attention mechanism, KT-NET can directly use all 1-hop neighbors of the mentioned entities. Although this limits the ability of KT-NET to select the most relevant mentioned entities indirectly via the HMP mechanism, the knowledge-enriched representation in Eq.3 is obtained by simple concatenation of the embeddings from different KGs. Too much knowledge incorporation may divert the sentence from its correct meaning (Knowledge noise issue). We expect these two potential improvements to be a promising avenue for future research.

### Appendix H

#### Limitations and Further Improvements of KELM

Limitations for KELM are two-fold: (1) Meanings of mentioned entities in different KGs that share the same entity mentions in the text may conflict with each other. Although HMP can help to select the most relevant mentioned entities in a single KG, there’s no mechanism to guarantee the selections across different KGs; (2) Note the knowledge-enriched token in the third subgraph of UKET. While KT-NET only contains the third subgraph with a modified knowledge integration module.

### References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In NIPS.

Carlson, A.; Betteridge, J.; Kiesel, B.; Settles, B.; Jr., E. H.; and Mitchell, T. 2010. Toward an Architecture for Never-Ending Language Learning. In Proceedings of the Conference on Artificial Intelligence (AAAI), 1306–1313. AAAI Press.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ferragina, P.; and Scaiella, U. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). Proceedings of the 19th ACM international conference on Information and knowledge management.

Haonan, L.; Huang, S. H.; Ye, T.; and Xiuyan, G. 2019. Graph Star Net for Generalized Multi-Task Learning. arXiv:1906.12330.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL).

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2019a. K-BERT: Enabling Language Representation with Knowledge Graph. arXiv:1909.07606.

### Appendix G

#### Further Discussions About the Novelty

w.r.t SKG/KT-NET

UKET defined in KELM consists of three subgraphs in a hierarchical structure, each subgraph corresponds to one sub-process of our proposed HMP mechanism and solves one problem presented in the Hierarchical Knowledge Enhancement Module part of Methodology Section. SKG only uses GNN to dynamically encode the extracted KG which corresponds to the first part of UKET, it can not solve the knowledge ambiguity issue and forbids interactions among knowledge-enriched tokens. KT-NET defines a similar graph as the third part of UKET. However, the first and second subgraphs of UKET are absent. The second subgraph of UKET is independent of ideas of KT-NET and SKG, thus KELM is not a simple combination of these two methods. We are the first to unify text and KG into a graph and to propose this hierarchical message passing framework to incorporate two heterogeneous information. SKG/KT-NET can be interpreted as parts of the ablation study of components of KELM. The result of SKG is ablation with the component only related to the first subgraph of UKET. While KT-NET only contains the third subgraph with a modified knowledge integration module.
Zhang, S.; Liu, X.; Liu, J.; Gao, J.; Duh, K.; and Durme, B. V. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. arXiv:1810.12885.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced Language Representation with Informative Entities. arXiv:1905.07129.