K-DLM: A Domain-Adaptive Language Model Pre-Training Framework with Knowledge Graph

Jiaxin Zou[1], Zuotong Xie[1], Junhua Chen[2], Jiawei Hou[2], Qiang Yan[2], and Hai-Tao Zheng[1,3(✉)]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
{zoujx20,xiezt20}@mails.tsinghua.edu.cn, zheng.haitao@sz.tsinghua.edu.cn
[2] Department of Search and Application, Weixin Group, Tencent, China
{jeshuachen,jiaweihou,rolanyan}@tencent.com
[3] Pengcheng Laboratory, Shenzhen 518055, China

Abstract. Despite the excellent performance of pre-trained language models, such as BERT, on various natural language processing tasks, they struggle with tasks that require domain-specific knowledge. Integrating information from knowledge graphs through pre-training tasks is a common approach. However, existing models tend to focus on entity information at the word level and fail to capture the rich information in knowledge graphs. To address this issue, we propose a domain-adaptive language model pre-training framework with a knowledge graph (K-DLM). K-DLM can learn both word and lexical-semantic level entity information and relationships from the knowledge graph. It predicts entity categories and sememes for masked phrases, replaces entities in sentences according to the knowledge graph, and learns relationship information via contrastive learning. The evaluation on open-domain and domain-specific tasks demonstrates that K-DLM outperforms previous models, particularly in domain-specific contexts. Our findings highlight K-DLM as an excellent pre-training framework for knowledge-driven problems that leverage domain knowledge graphs.

Keywords: Pre-trained language model · Knowledge graph · Vertical domain search

1 Introduction

Pre-trained language models (PTMs) such as BERT [7], XLNet [26], and RoBERTa [15] have achieved promising results on various natural language processing tasks [17,22,28]. However, the domain-specific knowledge required for certain tasks is not sufficiently learned through pre-training on open-domain corpora. Incorporating external knowledge, such as knowledge graphs (KGs), can

J. Zou and Z. Xie—These authors contributed equally to this work.

L. Iliadis et al. (Eds.): ICANN 2023, LNCS 14257, pp. 447–459, 2023.
https://doi.org/10.1007/978-3-031-44216-2_37

enhance PTMs' performance on downstream tasks. Researchers have primarily focused on two approaches for integrating KGs into PTMs: embedding-based and task-based.

The embedding-based approach, such as ERNIE-Tsinghua [29], KEPLER [23] and KELM [1], employs entity embeddings or natural language descriptions of KGs for pre-training, while the task-based approach, such as LIBERT [10] and SentiLR [9], incorporates pre-training tasks to acquire factual knowledge. However, current task-based models only consider entity information at the word level, disregarding lexical-semantic level and relationship information. This limitation hampers their ability to capture comprehensive knowledge within KGs.

We introduce K-DLM, a domain-adaptive language model pre-training framework with a KG that combines embedding-based and task-based approaches. K-DLM utilizes the masked language model (MLM) from BERT, employing entity and phrase level masking to pre-train on Chinese corpora [19,20]. By integrating common sense knowledge base with the domain KG, K-DLM enhances both universal and specific knowledge. It employs soft-labeling to predict entity categories and sememes of phrases, while learning relationship information through supervised contrastive learning. Additionally, we propose a novel entity replacement strategy to create positive and negative samples for relationship learning. Our experiments demonstrate the superior performance of K-DLM over previous models, particularly on domain-specific tasks, making it an effective pre-training framework for knowledge-driven problems involving domain KGs.

In this paper, K-DLM's performance was evaluated on six tasks across 17 Chinese datasets in open and specific domains. Results show that K-DLM performs well on open-domain tasks, especially those involving sememes. Additionally, category information of entities is crucial for NER tasks. K-DLM also performs better on domain-specific tasks due to its ability to utilize relationship knowledge. Overall, the main contributions of K-DLM are as follows:

- We propose a domain-adaptive language model pre-training framework with a KG (K-DLM).
- K-DLM can fully capture word and lexical-semantic level entity information as well as relationship information in the KG.
- By incorporating external knowledge, K-DLM significantly outperforms previous models not only on all domain-specific tasks but also on most open-domain NLP tasks.

2 Related Works

2.1 Embedding-Based Approaches

KG embedding, as represented by TransE [2], models relationships by operating on low-dimensional embeddings of entities in KGs. ERNIE-Tsinghua [29] introduces the KG into pre-trained language models by combining the language and knowledge embeddings obtained by TransE. However, this approach presents a

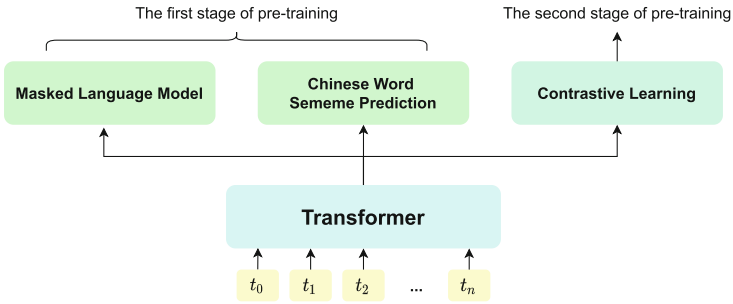

Fig. 1. Overview of the proposed Domain-Adaptive Language Model Pre-Training Framework with KG (K-DLM)

Heterogeneous Embedding Space (HES) problem where the language and knowledge embeddings are not obtained simultaneously. To address this issue more effectively, K-BERT [13] and CoLAKE [18] explicitly include knowledge triples in the training corpus to pre-train the language model and learn the knowledge representation concurrently. KEPLER [23] learns entity representation directly from entity description text and combines it with relationship embedding obtained by TransE. Furthermore, KELM [1] converts knowledge triples into fluent and natural sentences and adds them to the corpus of the pre-training model. By transforming heterogeneous KGs into text, the vector-space of knowledge representation becomes more consistent with that of language representation.

2.2 Task-Based Approaches

Since the release of BERT [7], various pre-training tasks have been proposed for different purposes, including learning external knowledge. ERNIE-Baidu [19] improved BERT's masking strategy to incorporate entity information from KGs. To overcome the limitations of predicting only single words when masked, ERNIE-Baidu masks all tokens that compose a complete phrase or entity simultaneously. SentiLR [9] extends MLM to Label-Aware MLM by adding emotional polarity to each word, while SenseBERT [11] predicts masked words and their super senses in WordNet [16] simultaneously to integrate semantic KGs into pre-trained language models. WKLM [24] replaces entities in the sentence with the same type of entities in Wikipedia and trains the model to recognize these replacements.

3 Method

In this section, we introduce K-DLM, a framework consisting of three steps: pre-processing and two-stage pre-training. The overall architecture of K-DLM is illustrated in Fig. 1.

3.1 Knowledge Graph Fusion

Before two-stage pre-training, we merge the domain-specific KG with HowNet, guided by two fundamental principles:

- We classify entities into their corresponding sememes in HowNet when the categories of entities in the KG align with those in HowNet.
- We retain the categories in the domain-specific KG when the categories of entities in the KG do not align with those in HowNet.

We refer to the categories of entities and sememes of phrases in the fused KG collectively as "sememes".

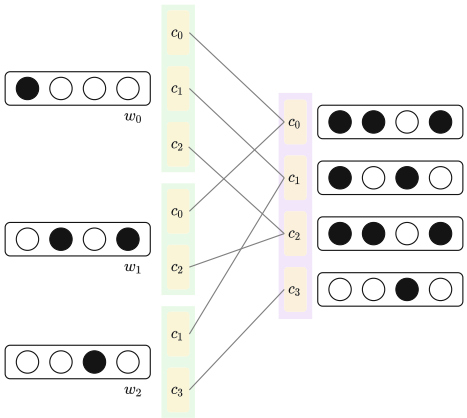

Fig. 2. Construction of mapping from characters to sememes. The set of possible sememes of each character is the union of possible sememes of all words composed of it

3.2 Masked Language Model and Chinese Word Sememe Prediction

During the first pre-training stage, K-DLM undergoes two tasks: masked language modeling and Chinese word sememe prediction. For masked language modeling, entity and phrase level masking strategies are employed. In the Chinese word sememe prediction task, K-DLM is trained to predict sememes in the fused KG. To accommodate KG integration, we modify the embedding layer and pre-training objective while utilizing the Transformer Encoder [21].

Embedding Layer. The embedding layer combines multiple embeddings to generate the input representation. We modify the input embedding E_{word} by summing four embeddings:

$$E_{word} = E_{tok} + E_{sem} + E_{seg} + E_{pos} \tag{1}$$

where E_{sem} and E_{pos} follow the original BERT. For E_{tok}, we utilize entity and phrase level masking strategies, masking entire words instead of individual Chinese characters. To capture the linguistic characteristics of Modern Chinese, we introduce a new split-and-merge mapping strategy. Let $X = (x_0, x_1, \cdots, x_n)$ denote the vocabulary index of a sentence $(c_0, c_1, \cdots, c_n)$, where n is the sentence length and $x_i \in \mathbb{R}^{D_W}$. E_{sem} is computed using a two-layer mapping:

$$E_{sem} = SMX \tag{2}$$

where $M \in \mathbb{R}^{D_S \times D_W}$ is a static mapping from characters to the union of corresponding sememes constructed with the fused KG, and $S \in \mathbb{R}^{d \times D_S}$ is a learnable mapping from sememes to the internal Transformer dimension d, where D_S is the size of the sememe vocabulary. Figure 2 illustrates an example of the construction process for M.

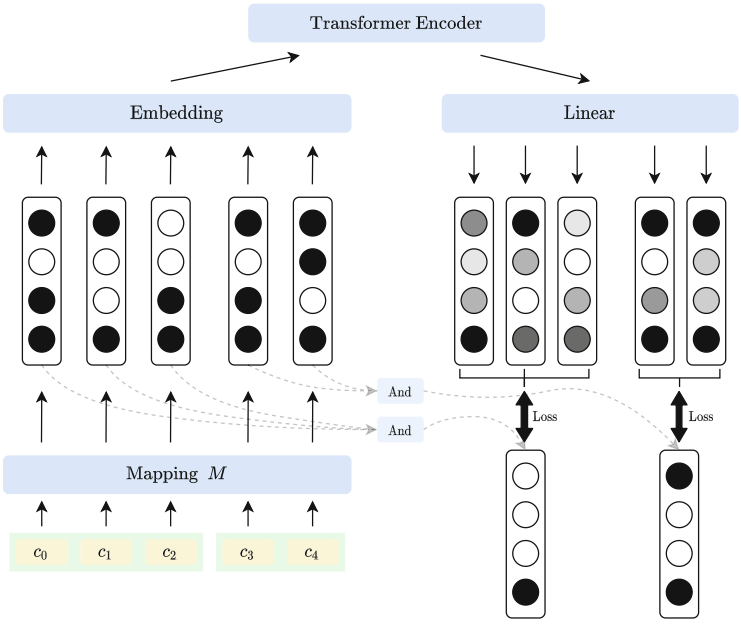

Fig. 3. Chinese word sememe prediction

Pre-Training Objective. We enhanced the original Masked Language Model (MLM) used in BERT by introducing entity and phrase level masking. Specifically, we mask all Chinese characters that belong to a complete word and require the model to recover the entire word during MLM pre-training. Besides, in our paper, we focus on predicting the allowed sememes of Chinese words, not char-

acters. Thus, we propose $\mathcal{L}_{CSP}$ for Chinese word sememe prediction:

$$\begin{aligned}\mathcal{L}_{CSP} = &- \log \sum_{s \in \bigcap\limits_{c \in w} PS(c)} p(s|context) \\ &- \sum_{s \in \bigcap\limits_{c \in w} PS(c)} \frac{1}{|\bigcap\limits_{c \in w} PS(c)|} \log p(s|context)\end{aligned} \tag{3}$$

where c is the character-level masked token and w is the whole Chinese word which c belongs to. The second penalty term enables the model to predict the given token c as all sememes in $PS(c)$ with possibility tending to equality, which enhanced the generalization ability. The output embedding of tokens belonging to the same word can be averaged according to the word segmentation boundary and used to predict all possible sememes of the whole Chinese word. We show an illustration of Chinese word sememe prediction in Fig. 3. Taking MLM task and CSP task together as objective of the first stage pre-training, we introduced the entities in the KG into the K-DLM:

$$\mathcal{L}_{stage1} = \mathcal{L}_{MLM} + \mathcal{L}_{CSP} \tag{4}$$

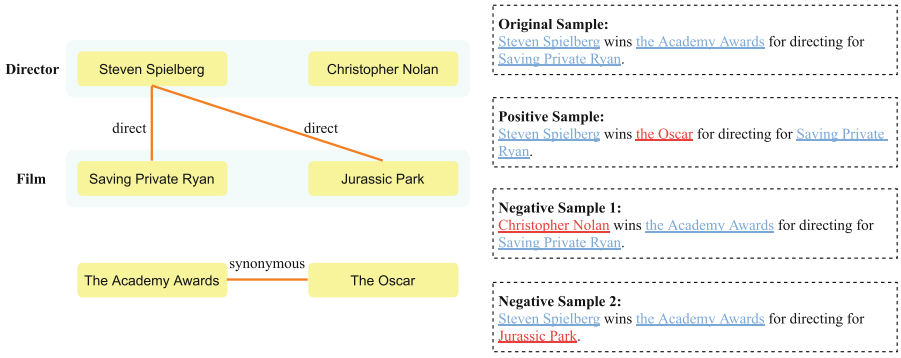

Fig. 4. The construction of positive and negative samples. Positive sample is obtained by synonymous replacement. Negative sample 1 illustrates entity replacement based on unique relationships, while negative sample 2 demonstrates replacement using non-unique relationships

3.3 Contrastive Learning

To incorporate entity relationships into K-DLM, we conduct the second stage of pre-training due to the inherent complexity of learning explicit relationships through MLM tasks. Following the contrastive framework of SimCLR [3], we introduce a novel replacement strategy for relationship types in the knowledge graph, generating positive and negative examples. Our second-stage pre-training objective employs a cross-entropy loss with in-batch negatives [4].

Replacement Strategy. In the fused KG, relationships encompass both general semantic relationships (e.g., synonyms, hypernyms, and hyponyms) and domain-specific relationships between entities. To connect corpus entities to the KG, we employ an off-the-shelf entity linking tool[1], discarding sentences without entities. For an entity e, we define positive ($P(e)$) and negative ($N(e)$) candidate sets. We propose three replacement strategies for relationships:

- **Semantic Relationship Replacement**: Synonymous replacements are considered positive samples, while hypernymous and hyponymous replacements are treated as semantic changes and used to construct negative samples. Specifically, synonyms of e are added to $P(e)$, and hypernyms and hyponyms of e are added to $N(e)$.
- **Unique Relationship Replacement**: This strategy is applied when a sentence contains multiple entities. We design a replacement strategy for each entity based on relationship uniqueness. For entities e_1 and e_2 in a sentence, assuming e_1 remains constant and e_2 is replaced, the relationship between e_1 and e_2 is denoted as r, with type t. If r is the unique relationship of type t for e_1, we select an entity of the same type as e_2 from the KG and add it to $N(e_2)$. To increase task difficulty, we calculate the edit distance between the original entity and each entity of the same type, randomly selecting from the 10 entities with the smallest edit distance as replacements.
- **Non-unique Relationship Replacement**: This strategy is employed when e_1 has relationships with other entities in the KG, excluding r, of type t. Now, r is considered a non-unique relationship for e_1. We randomly select an entity from all entities with the same relationship to e_1 and add it to $N(e_2)$.

Figure 4 illustrates the creation of positive and negative samples. The "direct" relationship between directors and films is used for comprehensibility purposes, but is not included in ServiceKG.

4 Experiments

In this section, we present the details of training setup and conduct experiments on 17 Chinese datasets, among which 13 are open-domain, and 4 are specific-domain, to answer the following research questions:

- **RQ1:** What is the role of sememes and types of phrases and entities in open-domain classification tasks?
- **RQ2:** How does our proposed method perform compared with others of introducing KG into pre-trained language model?
- **RQ3:** Could our proposed method benefit from domain KG in domain-specific tasks?

[1] Developed by WeChat search algorithm team.

4.1 Experiment Setup

Pre-training Corpora. To evaluate our proposed method and compare it with previous works [6,13], we pre-train our model using five Chinese corpora. These corpora include **WikiZh**, which is utilized to train **bert-base-chinese** in [7], **WebtextZh**, which is used to train **K-BERT** in [13], **Sogou baike**, **Baike QA**, and **NewsZh**. The total size of these corpora is similar to the pre-training corpus of **Chinese RoBERTa-wwm-ext** in [6], which is not publicly available. We construct our pre-training corpus, named **"ext"**, utilizing the above five corpora about encyclopedia, QA, and news to ensure impartial comparison with the pre-training corpus of **Chinese RoBERTa-wwm-ext**.

Knowledge Graph. We utilize Hownet[2] [8] as our source of common sense knowledge. Unlike K-BERT [13], we only integrate some concepts and sememes from Hownet into the domain KG, named ServiceKG. ServiceKG is constructed from billions of search logs and contains around 60k nodes and 200k relations of ten types of nodes and five types of relations. Due to copyright reasons, we cannot publish the complete ServiceKG. The sememes and relations from Hownet are used in all our experiments, including the results marked as **ServiceKG** in Table 3. Limited by model scale and computing power, we selected the top 97 sememes out of 2196, which cover over 70% of the sememe labels in Hownet.

Baselines. We compare our proposed K-DLM with five baselines in this paper:

- **Google BERT**, the official BERT (Chinese) pretrained on **Wikizh** [7].
- **Chinese RoBERTa-wwm-ext**, the RoBERTa-like BERT pretrained on **ext** corpus [6].
- **K-BERT**, the K-BERT pretrained on **WikiZh** and **WebtextZh**, utilizing HowNet as a KG [13].
- **RoBERTa-wwm**, our implementation of RoBERTa-wwm pretrained on **Wikizh** and **WebtextZh**.
- **RoBERTa-wwm-ext**, our implementation of RoBERTa-wwm-ext pretrained on our **ext** corpus.

Evaluation Benchmarks. To evaluate the performance of our proposed K-DLM, we conducted experiments on 17 datasets belonging to six natural language understanding tasks, as follows:

- **Natural Language Inference:** CMNLI [25], XNLI [5].
- **Winograd Schema Challenge:** CLUEWSC2020 [25].
- **Semantic Similarity:** AFQMC [25], LCQMC [14], CSL [25].
- **Sentiment Analysis:** Book-Review [13], Chnsenticorp [13], Shopping [13], Weibo [13].

[2] https://openhownet.thunlp.org.

- **Named Entity Recognition:** MSRA-NER [12], Finance-NER [13], Medicine-NER [13].
- **Text Multi-Class Classification:** TNEWS [25], IFLYTEK [25], Affair-CLS, Service-CLS.

Out of the 17 datasets mentioned above, Finance-NER, Medicine-NER, Affair-CLS, and Service-CLS are domain-specific, while the rest are open-domain. Affair-CLS and Service-CLS are five-class datasets from two vertical search scenarios within WeChat, a platform with billions of daily active users. Affair-CLS is related to government affairs search, while Service-CLS is related to service search. In our experiments, we perform coarse-grained binary classification by grouping labels 0 to 2 as weakly relevant and labels 3 and 4 as strongly relevant.

Training Details. To clearly demonstrate the effect of introducing the KG, we used the pre-training configuration of RoBERTa-wwm [6], which employs the WordPiece encoding scheme. Our hyperparameters aligned with those of Google BERT, where model size matched the $BERT_{base}$ configuration, which is L=12, H=768, A=12, with a total of 102M parameters. Inputs were constructed as **DOC-SENTENCES** as in RoBERTa-wwm. We utilized the LAMB optimizer [27] with a batch size scaled from 512 to 32K on 128 T V100 GPUs with 32GB VRAM. Our model underwent pre-training for 15625 steps with an initial learning rate of 5e-3 and implemented a warm-up strategy for the first 20% of steps, followed by a linear decay of the learning rate.

4.2 Overall Performance

In this section, we compared our proposed K-DLM with several baseline pre-trained language models above respectively.

Table 1. Results of various models on classification tasks in CLUE benchmark (Acc.%)

Models	AFQMC	TNEWS	IFLYTEK	CMNLI	CLUEWSC2020	CSL
Google BERT	74.16	56.09	60.37	79.47	59.60	79.63
Chinese RoBERTa-wwm-ext	74.30	57.51	**60.8**	80.70	67.20	80.67
RoBERTa-wwm-ext	**74.47**	57.26	60.37	80.31	80.92	81.07
K-DLM-ext	74.40	**57.76**	59.37	**81.21**	**85.20**	**81.30**

Table 2. Results of various models on seven open-domain tasks (F1% for MSRA-NER, Acc.% for others)

Models	Book-Review		Chnsenticorp		Shopping		Weibo		XNLI		LCQMC		MSRA-NER	
	Dev	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*
Google BERT	88.3	87.5	93.3	94.3	96.7	96.3	98.2	98.3	76.0	75.4	88.4	86.2	94.5	93.6
K-BERT(HowNet)	88.5	87.4	**95.4**	95.6	96.9	96.9	**98.3**	**98.4**	77.2	77.0	89.2	87.1	96.3	95.6
RoBERTa-wwm	89.2	88.1	95.1	94.9	96.9	97.0	97.7	98.0	77.6	77.7	88.7	87.0	96.3	95.3
K-DLM(HowNet)	**89.3**	**88.3**	95.3	**95.6**	**97.1**	**97.1**	97.9	97.7	**78.6**	**78.1**	**89.9**	**88.6**	**96.7**	**96.8**

Classification Tasks in CLUE Benchmark (RQ1). We compare our K-DLM with Chinese RoBERTa-wwm-ext [6] to highlight the advantages of incorporating sememe and category knowledge. Table 1 presents the results on the development set of six sentence classification tasks in the CLUE benchmark [25]. The performance of Google BERT and Chinese RoBERTa-wwm-ext is obtained from the public leaderboard of the CLUE benchmark, while we use the hyperparameters provided by Chinese RoBERTa-wwm-ext without further tuning. Our observations can be categorized as follows:

- Sememes help correctly classify sentences with categories based on commonsense phrases or entities, showing a positive impact on tasks such as TNEWS, CMNLI, CLUEWSC2020, and CSL.
- In contrast, when domain-specific entities consisting of common characters impact the category (AFQMC and IFLYTEK), sememes lead to incorrect category predictions, indicating that adding these entities into the KG can solve the issue.

Other Open-Domain Tasks (RQ2). To compare our K-DLM with K-BERT, which is equipped with HowNet, we pre-trained both **RoBERTa-wwm** and our **K-DLM** on the same corpus as K-BERT (HowNet) as stated in [13]. We then evaluated the models on seven open-domain tasks, with each dataset divided into *train*, *dev*, and *test* subsets. We fine-tuned the models on the *train* subset, selected the best model based on the *dev* subset, and evaluated its performance on the *test* subset. The experimental results are presented in Table 2.

- K-DLM did not show a significant performance improvement for sentiment analysis tasks (i.e., Book-Review, Chnsenticorp, and Shopping) because sentiment mainly relies on emotion words and negations rather than knowledge. Moreover, for colloquial-style sentences from social media (i.e., Weibo), inaccurate sememe predictions impaired the model's ability to judge emotions.
- K-DLM outperforms K-BERT on common knowledge-dependent tasks (XNLI, LCQMC, and MSRA-NER) by addressing a problem encountered by K-BERT. Fine-tuning K-BERT requires word segmentation and NER, which introduce errors and restrict knowledge utilization. In contrast, our character-level sememe incorporation in K-DLM benefits downstream tasks without entity linking and improves decision-making for entities outside the KG.

Table 3. Results of various models on specific-domain tasks (%)

Models	Finance-NER			Medicine-NER			Affair-CLS			Service-CLS		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
Google BERT	84.8	87.4	86.1	91.9	93.1	92.5	97.8	91.0	94.3	**77.1**	84.9	80.8
K-BERT(HowNet)	86.3	**88.5**	87.3	93.5	93.8	93.7	-	-	-	-	-	-
K-BERT(ServiceKG)	-	-	-	-	-	-	97.8	90.9	94.2	76.8	85.2	80.8
RoBERTa-wwm	86.5	87.9	87.2	93.7	94.5	94.1	97.9	91.3	94.5	76.5	85.9	80.9
K-DLM	**86.9**	88.2	**87.5**	**94.0**	**95.0**	**94.5**	**97.9**	**92.2**	**95.0**	75.5	**88.7**	**81.5**

Specific-Domain Tasks (RQ3). We conduct experiments on four specific-domain tasks to assess whether domain KG benefits K-DLM. Following the experiment setup in [13] for the Finance-NER and Medicine-NER tasks, we evaluate models equipped with HowNet on these tasks. For our self-developed **ServiceKG**, we fine-tune the **RoBERTa-wwm** as described in Sect. 4.2 using the method proposed by K-BERT with this KG, and obtain **K-BERT (ServiceKG)** for comparison with our K-DLM. The results are summarized in Table 3.

- In domain-specific NER tasks, HowNet's financial and medical knowledge aids entity identification in sentences. Our K-DLM outperforms K-BERT (HowNet) in terms of precision and F1 score, indicating its successful classification of entities into correct categories with the assistance of sememes.
- For our query intention classification tasks (Affair-CLS and Service-CLS), we focus more on the relationship between entities (i.e., services offered to something) than on the types of entities in the query. Therefore, the phrase and entity mask has no significant effect. Because queries are relatively short, adding entity relation entity triples into a query using K-BERT can cause semantic drift that cannot be ignored. In contrast, our **K-DLM** introduces relations by replacement without changing the sentence length, resulting in improved performance in short text classification.

5 Conclusion

In summary, our proposed K-DLM framework utilizes a Chinese soft-label scheme, split-and-merge mapping strategy, and replacement-based relation injection strategy for short text processing. This approach enhances the utilization of sememes and category information, leading to improved model performance in vertical domain query understanding while avoiding semantic drift. Experimental results demonstrate that using sememes of Chinese words enhances the performance of open-domain classification tasks relying on common knowledge. We conducted comparisons with alternative KG introduction methods to assess efficacy and applicability. Additionally, our approach enables adaptability to tasks across diverse domains by leveraging domain-specific KGs during pre-training.

Acknowledgments. This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, Beijing Academy of Artificial Intelligence (BAAI), the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), the Major Key Project of PCL for Experiments and Applications (PCL2021A06), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008).

References

1. Agarwal, O., Ge, H., Shakeri, S., Al-Rfou, R.: Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In: NAACL-HLT, pp. 3554–3565. Association for Computational Linguistics (2021)
2. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS, pp. 2787–2795 (2013)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020)
4. Chen, T., Sun, Y., Shi, Y., Hong, L.: On sampling strategies for neural network-based collaborative filtering. In: KDD, pp. 767–776. ACM (2017)
5. Conneau, A., et al.: XNLI: evaluating cross-lingual sentence representations. In: EMNLP, pp. 2475–2485. Association for Computational Linguistics (2018)
6. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for Chinese BERT. IEEE ACM Trans. Audio Speech Lang. Process. **29**, 3504–3514 (2021)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1), pp. 4171–4186. Association for Computational Linguistics (2019)
8. Dong, Z., Dong, Q.: HowNet - a hybrid language and knowledge resource. In: International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings 2003, pp. 820–824 (2003)
9. Ke, P., Ji, H., Liu, S., Zhu, X., Huang, M.: SentiLARE: sentiment-aware language representation learning with linguistic knowledge. In: EMNLP (1), pp. 6975–6988. Association for Computational Linguistics (2020)
10. Lauscher, A., Vulic, I., Ponti, E.M., Korhonen, A., Glavas, G.: Specializing unsupervised pretraining models for word-level semantic similarity. In: COLING, pp. 1371–1383. International Committee on Computational Linguistics (2020)
11. Levine, Y., et al.: SenseBERT: driving some sense into BERT. In: ACL, pp. 4656–4667. Association for Computational Linguistics (2020)
12. Levow, G.: The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In: SIGHAN@COLING/ACL, pp. 108–117. Association for Computational Linguistics (2006)
13. Liu, W., et al.: K-BERT: enabling language representation with knowledge graph. In: AAAI, pp. 2901–2908. AAAI Press (2020)
14. Liu, X., et al.: LCQMC: a large-scale Chinese question matching corpus. In: COLING, pp. 1952–1962. Association for Computational Linguistics (2018)
15. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019)
16. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
17. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100, 000+ questions for machine comprehension of text. In: EMNLP, pp. 2383–2392. The Association for Computational Linguistics (2016)
18. Sun, T., et al.: CoLAKE: contextualized language and knowledge embedding. In: COLING, pp. 3660–3670. International Committee on Computational Linguistics (2020)
19. Sun, Y., et al.: ERNIE: enhanced representation through knowledge integration. CoRR abs/1904.09223 (2019)

20. Sun, Y., et al.: ERNIE 2.0: a continual pre-training framework for language understanding. In: AAAI, pp. 8968–8975. AAAI Press (2020)
21. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
22. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: EMNLP, pp. 353–355. Association for Computational Linguistics (2018)
23. Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., Tang, J.: KEPLER: a unified model for knowledge embedding and pre-trained language representation. Trans. Assoc. Comput. Linguistics **9**, 176–194 (2021)
24. Xiong, W., Du, J., Wang, W.Y., Stoyanov, V.: Pretrained encyclopedia: weakly supervised knowledge-pretrained language model. In: ICLR. OpenReview.net (2020)
25. Xu, L., et al.: CLUE: a Chinese language understanding evaluation benchmark. In: COLING, pp. 4762–4772. International Committee on Computational Linguistics (2020)
26. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: NeurIPS, pp. 5754–5764 (2019)
27. You, Y., et al.: Large batch optimization for deep learning: training BERT in 76 minutes. In: ICLR. OpenReview.net (2020)
28. Zellers, R., Bisk, Y., Schwartz, R., Choi, Y.: SWAG: a large-scale adversarial dataset for grounded commonsense inference. In: EMNLP, pp. 93–104. Association for Computational Linguistics (2018)
29. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. In: ACL (1), pp. 1441–1451. Association for Computational Linguistics (2019)