Improving Chinese knowledge enhancement models with unified respresentation space

Fei Li[a,b], Zhengyi Chen[b], Yanyan Wang[*b], and Yin Xu[b], Ticheng Duan[b]
[a]School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230000, China; [b]Innovation + Research Institute, GuoChuang Cloud Technology Ltd., Hefei, 230000, China

ABSTRACT

Knowledge Enhanced Pre-trained Language Models have recently improved context-aware representations through external knowledge and linguistic knowledge from grammatical or syntactic analysis. The mismatch between text and knowledge graph embeddings in the feature space cannot be resolved in the fine-tuning phase and input module knowledge augmentation. In this paper, we revisit and advance the development of natural language understanding in Chinese and propose an improving Chinese knowledge Enhancement Models with Unified Respresentation Space. Specifically, knowledge embedded in the knowledge graph triples is effectively injected into it based on a novel pre-training task and knowledge-aware masking strategy. We conducted extensive experiments in seven Chinese nature language process tasks to evaluate the proposed model. The experimental results show that Our Model understands the external knowledge more deeply. We also demonstrate the effectiveness of the proposed method by ablation experiments.

Keywords: Knowledge enhancement, knowledge-aware masking strategy, pretrained language model, representation space

1. INTRODUCTION

Pre-trained Language Models (PLMs) significantly improves various downstream nature language process tasks by performing self-supervised pre-training on a large-scale text corpus to capture semantic knowledge [1]. Compared with PLMs, KGs contain a large number of structured facts, which can be effectively embedded into continuous entity and relationship representations using knowledge embedding (KE) methods [2]. Most knowledge injection methods can be divided into three categories: (1) injecting knowledge into the input sequence[3][4][5]; (2) using two encoders to represent knowledge and source text separately, and then fusing the two representation spaces using a knowledge fusion module[6]; (3) modifying the MLM task or adding a knowledge representation learning pre-training task. The first method is straightforward, but breaks the context integrity of the input sequence. Obviously, the MLM task is not suitable for connecting to the input sequence's trigrams. A fatal flaw of the second method is that directly merging the two representation spaces trained by two different encoders is cumbersome and unreasonable. For the third method, the ELECTRA model did not address the problem of external knowledge injection. Therefore, although modifying pre-training tasks requires training large-scale knowledge-enhanced pre-trained language models (KEPLMs) from scratch, which requires high-performance computing devices and takes a long time. Modifying pre-training tasks is the only method that can jointly train external and internal knowledge.

To address the above issues, we propose a model. First, we design a pre-training task to unify two different types of input into a unified representation space. It learns to identify important words in the input sequence based on clues, adds knowledge about words to the end, masks important words in the input sequence, and forces the model to understand the content from external knowledge. To achieve this pre-training task, we specifically design a knowledge-aware masking strategy for it. Second, we pay special attention to Chinese, as it is one of the most widely used languages. Experimental results show that our proposed model improves PLM's semantic understanding by modifying the mask strategy. The contributions of this paper are as follows.

[*] Corresponding author: wang.yanyan@ustcinfo.com. Tel: +86-13731991874.
fli312@mail.ustc.edu.cn(F. Li);chen.zhengyi@kdgcsoft.com(Z.Y.Chen);wang.yanyan@ustcinfo.com(Y.Y.Wang); xu.yin@ustcinfo.com (Y. Xu); duan.ticheng@ustcinfo.com(T.C.D).

International Conference on Automation and Intelligent Technology (ICAIT 2024), edited by R. Usubamatov, S. Feng, X. Mei, Proc. of SPIE Vol. 13401, 134010A · © The Authors. Published under a Creative Commons Attribution CC-BY 3.0 License
doi: 10.1117/12.3047851

- We present a novel KEPLM to inject the knowledge into PLMs, which train representation by jointly knowledge prediction and internal knowledge understanding tasks.
- We propose a method to filter knowledge-related keywords. This method selects the most critical word for the model through two perspectives: semantic substitutability and word frequency.
- Our experiments on several public Chinese benchmark datasets for text classification and question answering show that our model performs better than existing top methods.

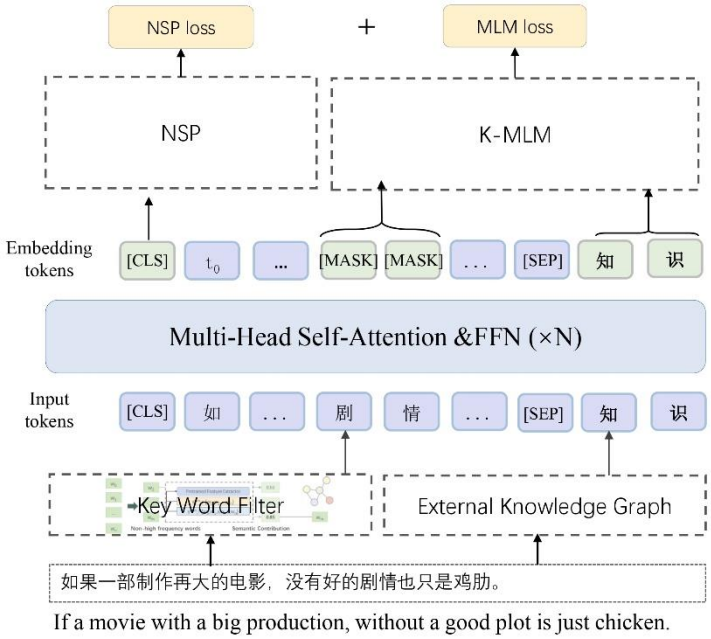

Figure 1. The framework of our model. (1) Key word filter: Select keywords from the input sequence and find relevant knowledge points of the word. (2) Mask strategy: Mask the important word, and add the related knowledge at the end of the input sequence.

2. RELATED WORK

In this section, we will provide a comprehensive overview of the previous studies conducted on two crucial aspects: Pre-trained Language Models (PLMs) and Knowledge-Enhanced Pre-trained Language Models (KEPLMs). Pretrained language models based purely on data are often unstable and challenging to interpret. To solve these problems, knowledge-enhanced pre-trained language models have been proposed in recent years. By effectively integrating knowledge, these models can improve stability and interpretability [7]. We have summarized the recent KE-PLMs into the following three types:

Inject Knowledge in Input Sequence: The current research focuses on three options: (1) Find the triples in the knowledge graph that correspond to the entity words in the input sequence and insert them into the input sequence. For example, Baidu's proposed ERNIE 3.0 [8] takes pre-trained corpus and knowledge graph as inputs before the Embedding layer. (2) Insert entity descriptions corresponding to entity words in the input sequence[9]. For example, [3] appends questions and candidate answers at the end of the input sequence. (3) Organize the text into a graph structure and incorporate subgraphs from a knowledge graph. Then, flatten the structure back into a text sequence for use in the PLM. For example, CoLAKE [10] splices and flattens the knowledge subgraphs and word graphs as the input sequence. K-BERT [4] splices and flattens the knowledge subgraphs and input sentences into a sentence tree as input.

Special Knowledge Fusion Module: This approach for gaining knowledge involves using a pre-trained encoder to learn the context, and a knowledge encoder to learn about external knowledge. These two representations are then combined using a knowledge fusion module. There are three options for adding the knowledge fusion module: (1) On top of the entire PLM. The KEPLER[9] model has two parts: one part learns how words are used in sentences, and another part learns

how different ideas are related to each other. However, KEPLER uses a different Encoder Layers Network to learn the representations of both knowledge sources [11]. (2) Between the Transformer layers of PLM. For example, the JAKET [12] model enhances knowledge by injecting the results of the knowledge encoder into each Transformer Layer. (3) Inside the Transformer layers of PLM. e.g., ERNIE-THU[11] added a knowledge fusion module after multiple encoders as a whole.

Modify Pre-training Task: The current research focuses on two directions: (1) modifying MLM tasks, (2) adding knowledge-related pre-training tasks. For example, MacBERT[13] uses MLM as a correction strategy. i.e., replacing [MASK] with words similar to the target word as mask characters, reducing the gap between pre-training-and fine-tuning. PERT [14] is a method that replaces MLM with a task that predicts the order of words in a sentence. MLM and its variants dominate the field, but researching other pre-training tasks is meaningful. In this paper, we argue that adjusting the masking strategy and adding a pre-training task to achieve the goal of fusing external and internal contextual knowledge into the same representation space.

3. OUR METHOD

An overview of Our Model is depicted in Fig. 1. As we can see that the proposed model shares identical neural architecture and pretraining tasks with BERT. However, the proposed model uses a new pre-training task: Knowledge-based Masked LM (K-MLM) task, thus there is a slight difference in the input, mask strategy and training objective. We mask low-frequency and low-semantic-substitutable keywords based on Zipf's law and the principle that word importance is inversely proportional to substitutability. Inspired by the input knowledge injection, we add the external knowledge at the end of the input sequence to urge the model to understand the masked keyword meaning of the sentence. Thus the training objective of Our Model is to jointly learn the losses of the K-MLM and NSP task. In the following sections, we will elaborate on the design of each auxiliary module.

3.1 Pre-training KEBERT

3.1.1 Task #1: Knowledge-based Masked LM (K-MLM)

The MLM task and the NSP task help the model learn the internal knowledge of the input sequence. Specifically, MLM randomly selects 15% of the tokens, and each of the selected tokens has an 80% probability of being replaced with [MASK], a 10% probability of being replaced with one other token, and a 10% probability of not being replaced [15]. However, neither of the two tasks can directly and rationally utilize the external knowledge. Therefore, how to guide the model to understand the external knowledge is an important issue [16]. To deal with above problems, we propose a knowledge-based masked task as a pre-training task, which is shown to be much more effective in injecting the external knowledge than MLM and NSP.

3.1.2 Task #2: Next Sentence Prediction (NSP)

The sentence prediction (NSP) objective is inherited from BERT. In the pre-training phase, the NSP task trains the model to understand the relationship between two sentences. Specifically, for each sentence pair, in addition to retaining the correct original contextual sentence, there is a 50% probability that another sentence will be randomly selected as the following sentence to form the negative sample [15]. Regarding network structure, the MLM task will classify each sentence vector. Finally, a Negative Log-Likelihood Loss L_NSP is calculated over these sentence pairs.

3.2 Training objectives

In order to silky smoothly integrate the representation space of factual knowledge and the representation space of internal knowledge in PLMs, specific definitions are shown in Equation 1.

$$L = L_{NSP} + L_{K-MLM} \tag{1}$$

where $L_{K-MLM}$and L_{NSP} are the losses for K-MLM and NSP correspondingly. Jointly optimizing the two objectives can smoothly integrate knowledge from external KGs into the contextual model

3.3 Knowledge-aware masking strategy

To adapt the K-MLM task, we design a knowledge-aware masking strategy. Specifically, we first filtering essential words in a sentence based on their frequency and semantic contribution. These words are then masked out with token [MASK].

In the following section, we will introduce the knowledge-aware important word entity detection module and the corresponding mask strategy.

The masking strategy of BERTbase consists of adding unique tokens and masking language models (MLM). It is assumed that knowledge is appended to the input sequence. In this case, external knowledge and input tokens are equally masked by [MASK], the model cannot distinguish between external knowledge and input tokens. Therefore, we propose a unified space to learn content and external knowledge simultaneously.

To address these issues, we propose modifying the existing masking strategy by replacing knowledge-related words with "[MASK]" and keeping other words masked with their original probabilities. This ensures that the model cannot see important words. To compensate for the information loss, we provide the related knowledge in the triad related to this critical word in the input sequence, as shown in Fig. 1. By introducing this masking strategy, we force the model to use the knowledge information and increase the difficulty of understanding the original input sequence.

4. EXPERIMENTS

4.1 Pre-training setup

We largely follow the training recipe of the common Chinese-KEPLMs, where we illustrate as follows. All models are trained from scratch.

- Pre-training Data: To ensure the success of Our Model, we constructed a diverse and high-quality Chinese corpus in 11 different categories, totaling 733MB. As far as we know, the Chinese corpus of ERNIE 3.0 is 4TB, and the pre-training Chinese corpus of MacBERT and PERT is 20GB.
- Knowledge Data: Furthermore, we have incorporated the OpenKG open-source abstract knowledge graph as an external knowledge base and have selected three tuples related to the corpus for introduction into the pre-training process. In total, the knowledge graph that we adapted comprises 52,179 non-repetitive tuples.
- Vocabulary: We created a corresponding vocabulary based on the Chinese corpus we used, which has a size of 13,142. This is smaller than the dictionary size of other PLMs, which is usually 21,128.
- Hyper-parameters: During the pre-training period, we set the maximum input sequence length to 512.
- Training Device: The training was done on a single GPU A100 (40G Memory)

We train Our Model with two sizes: Our Model-tiny (4-layer, 6-heads, 512-dim) and Our Model-base (6-layer, 8-heads, 768-dim), which are smaller than BERT settings.

4.2 Fine-tune setup

We choose the seven popular Chinese NLU datasets[17].

1. Text Classification (TC): AFQMC, TNEWS, IFLYTEK, OCNLI, WSC

2. Question Answering (QA): CMRC 2018, C3.

Specifically, the text categorization task involves selecting one or more appropriate categories to label text from a given passage. AFQMC [18]is a Chinese sentence pair semantic similarity judgment dataset containing sentence pairs from different domains. TNEWS[19] is a Chinese news headline classification dataset containing 15 news categories. IFLYTEK [20] is a Chinese long-text categorization dataset containing 119 application domains. OCNLI [21] is a Chinese natural language reasoning dataset containing three logical relations: entailment, contradiction, and neutrality. WSC [22] is a Chinese denotational disambiguation dataset containing several ambiguous pronouns and their antecedents. The question-answer task is extracting answers from given questions and text passages. CMRC2018 [23] is a Chinese machine reading comprehension dataset containing extracted questions and answers, and C3[5] is a cross-domain multiple-choice Chinese machine reading comprehension dataset that contains both dialog and article text types.

We selected the BERT-base-Chinese model as well as three other KEPLM models that are similar to our ideas, aligned the hyperparameters, and compared them on five text classification tasks and three QA tasks in CLUE. These models contain: BERTbase, MacBERT, PERT, ERNIE3. Our Modeltiny model only used 200,000 pre-training corpus, which is basically

on par with the other KEPLM models. In addition, to ensure experimental robustness, we repeat experiment five times with different random seeds and report the average scores. The fine-tuning script references the original BERT script.

Table 1. Comparitive experiments.

Models	TC					QA	
	AFQMC	TNEWS	IFLYTEK	OCNLI	WSC	CMRC	C3
MacBERT	68.93	11.31	23.18	29.68	36.58	67.58	73.64
PERT	31.00	10.41	21.68	30.48	63.42	53.19	26.37
ERNIE 3.0	68.87	16.29	18.27	30.67	63.19	19.04	73.71
$\mathrm{OurModel}_{tiny}$	69.10	16.81	14.97	34.68	63.49	60.11	63.81
$\mathrm{OurModel}_{\mathrm{base}}$	**69.18**	**19.89**	16.32	**36.05**	**64.47**	**77.53**	**73.74**

4.3 General experimental

The CLUE benchmark comprises machine reading comprehension, text classification and question-answering tasks. Table 1 shows the results for all the tasks. We can observe that even in the case of smaller corpus pre-training, the model obtains similar results as in the case of large corpus pre-training.

- We primarily tested these models on the AFQMC, TNEWS, IFLYTEK, OCNLI, and WSC datasets. The experimental results show that our model has a significant advantage on the TNEWS and OCNLI datasets, as these datasets require the model to understand world knowledge and contextual information. On the OCNLI and TNEWS datasets, our model brought about improvements by effectively injecting external knowledge. However, on the AFQMC and WSC datasets, our model did not bring significant improvements, as sentence similarity computation mainly relies on internal contextual knowledge, and coreference resolution requires the model to have a deeper understanding of sentence structure. On the IFLYTEK dataset, our model performed weaker than the other model, which we speculate is due to the relatively small size of the dataset and the large number of categories (119), suggesting that a larger-scale pre-trained corpus might be necessary.
- The results of the Question Answering task. We mainly tested these models on the CMRC2018 and C3 datasets. The results show that our model significantly improves performance in CMRC2018, but not in C3. This is mainly due to the fact that the CMRC2018 dataset contains richer semantics and stronger contextual relationships, while the semantic changes in the C3 dataset are not very obvious. This is mainly due to the fact that our pre-trained data does not cover tasks like C3.

5. CONCLUSION

In this paper, we revisit Chinese knowledge-enhanced pre-trained language models to test whether the technologies in these state-of-the-art models can be effectively applied to non-English languages as well. We propose a new model that incorporates a knowledge-based masked language model (K-MLM) into the basic pre-training task, effectively verifying that injecting external knowledge enhances the semantic understanding capabilities of language models. Experiments on multiple Chinese natural language processing datasets show that the proposed model achieves comparable results to existing models with much less training data. Further research indicates that by introducing knowledge graph triplets and adopting novel pre-training tasks along with knowledge-aware masking strategies, our model can more deeply understand external knowledge and perform at the level of the latest research findings, especially in scenarios where contextual knowledge is sparse. Regarding interpretability, since our model aims to integrate internal and external knowledge within a unified representation space, it provides a more intuitive way to understand how the model synthesizes information from different sources to make predictions. In terms of stability, through ablation studies, we confirm the importance of designing specific knowledge injection tasks during the pre-training phase, which helps ensure the consistency and robustness of the model. Therefore, our approach has certain advantages over other existing models in terms of interpretability and stability. Future work will explore using user feedback as additional supervisory signals to enhance the training process of KEPLMs.

REFERENCES

[1] Zhang, T., Dong, J., Wang, J., Wang, C., Wang, A., Liu, Y., Huang, J., Li, Y., He, X., 2022. Revisiting and advancing chinese natural language understanding with accelerated heterogeneous knowledge pre-training.

[2] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data, in: Neural Information Processing Systems.

[3] Bian, N., Han, X., Chen, B., Sun, L., 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation.

[4] Liu, W., Zhou, P., Zhao, Z., Wang, Z., Wang, P., 2019b. K-bert: Enabling language representation with knowledge graph.

[5] Sun, K., Yu, D., Yu, D., Cardie, C., 2020a. Investigating prior knowledge for challenging chinese machine reading comprehension. Transactions of the Association for Computational Linguistics 8, 141–155.

[6] Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J., 2019b. Rotate: Knowledge graph embedding by relational rotation in complex space.

[7] Stevens, S., Su, Y., 2020. An investigation of language model interpretability via sentence editing. arXiv e-prints.

[8] Sun, Y., Wang, S., Feng, S., Ding, S., Wang, H., 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation

[9] Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., Tang, J., 2019. Kepler: A unified model for knowledge embedding and pre-trained language representation.

[10] Sun, T., Shao, Y., Qiu, X., Guo, Q., Zhang, Z., 2020b. Colake: Contextualized language and knowledge embedding

[11] Sun, Y., Shuohuan, W., Yukun, L., Feng, S., Chen, X., Zhang, H., Tian, X., Danxiang, Z., Tian, H., Wu, H., 2019a. Ernie: Enhanced representation through knowledge integration. Cornell University - arXiv, Cornell University - arXiv.

[12] Yu, D., Zhu, C., Yang, Y., Zeng, M., 2022. Jaket: Joint pre-training of knowledge graph and language understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 11630–11638URL: http://dx.doi.org/10.1609/aaai.v36i10.21417, doi:10.1609/aaai.v36i10.21417.

[13] Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G., 2019a. Pre-training with whole word masking for chinese bert.

[14] Cui, Y., Yang, Z., Liu, T., 2022. Pert: Pre-training bert with permuted language model.

[15] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[16] Emelin, D., Bonadiman, D., Alqahtani, S., Zhang, Y., Mansour, S., 2022. Injecting domain knowledge in language models for task-oriented dialogue systems. arXiv preprint arXiv:2212.08120.

[17] Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., et al., 2020. Clue: A chinese language understanding evaluation benchmark. arXiv preprint arXiv:2004.05986.

[18] Ma, H., Guo, H., 2022. External knowledge and data augmentation enhanced model for chinese short text matching, in: International Conference on Neural Information Processing, Springer. pp. 76–87.

[19] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X., 2020. Pre-trained models for natural language processing: A survey. Science China Technological Sciences 63, 1872–1897.

[20] Wang, Y., Wang, Y., Hu, H., Zhou, S., Wang, Q., 2023. Knowledge-graph-and gcn-based domain chinese long text classification method. Applied Sciences 13, 7915.

[21] Hu, H., Richardson, K., Xu, L., Li, L., Kübler, S., Moss, L.S., 2020. Ocnli: Original chinese natural language inference. arXiv preprintarXiv:2010.05444.

[22] Zheng, H., Li, L., Dai, D., Chen, D., Liu, T., Sun, X., Liu, Y., 2021. Leveraging word-formation knowledge for chinese word sense disambiguation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 918–923.

[23] Cui, Y., Liu, T., Che, W., Xiao, L., Chen, Z., Ma, W., Wang, S., Hu, G., 2019b. A span-extraction dataset for Chinese machine reading comprehension, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp.5883–5889. URL: https://aclanthology.org/D19-1600, doi:10.18653/v1/D19-1600.