When Few-Shot Learning Meets Large-Scale Knowledge-Enhanced Pre-training: Alibaba at FewCLUE

Ziyun Xu[1,2], Chengyu Wang[1], Peng Li[1], Yang Li[1], Ming Wang[1,3], Boyu Hou[1,4], Minghui Qiu[1(✉)], Chengguang Tang[1], and Jun Huang[1]

[1] Alibaba Group, Hangzhou, Zhejiang 311121, China
{xuziyun.xzy,chengyu.wcy,jerry.lp,ly200170,jinpu.wm,houboyu.hby, minghui.qmh,chengguang.tcg,huangjun.hj}@alibaba-inc.com
[2] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
[4] College of Computer Science, Chongqing University, Chongqing 400044, China

Abstract. With the wide popularity of Pre-trained Language Models (PLMs), it has been a hot research topic to improve the performance of PLMs in the few-shot learning setting. FewCLUE is a new benchmark to evaluate the few-shot learning ability of PLMs over nine challenging Chinese language understanding tasks, which poses significant challenges to the learning process of PLMs with very little training data available. In this paper, we present our solution to FewCLUE tasks by means of large-scale knowledge-enhanced pre-training over massive texts and knowledge triples, together with a new few-shot learning algorithm for downstream tasks. Experimental results show that the generated models achieve the best performance in both limited and unlimited tracks of FewCLUE. Our solution is developed upon the PyTorch version of the EasyTransfer toolkit and will be released to public.

Keywords: Pre-trained Language Model · Knowledge-enhanced Pre-trained Language Model · Knowledge Graph · Few-shot learning

1 Introduction

Recent years has witnessed the successful application of large-scale Pre-trained Language Models (PLMs) for solving various Natural Language Processing (NLP) tasks, based on the "pre-training and fine-tuning" paradigm [7,12,17]. To ensure high model accuracy for downstream tasks, it is necessary to obtain a sufficient amount of training data, which is often the bottleneck in low-resource scenarios.

In the literature, two types of approaches have been proposed to address the above-mentioned problem. The first approach is to inject factual knowledge

L. Wang et al. (Eds.): NLPCC 2021, LNAI 13029, pp. 422–433, 2021.
https://doi.org/10.1007/978-3-030-88483-3_34

from Knowledge Graphs (KGs) into PLMs. PLMs pre-trained on large-scale unstructured corpora pay little attention to semantic information of important entity mentions and their relations expressed in texts. In contrast, Knowledge-Enhanced PLMs (KEPLMs) can significantly improve the plain PLMs in terms of the language understanding abilities, by fusing both unstructured knowledge from texts and structured knowledge from KGs [6,14,26,31,32]. The other type of approaches is to leverage the few-shot learning abilities of PLMs. As ultra-large PLMs such as GPT-3 [3] are proved to have the abilities to solve an NLP task with very few training samples, it has become a hot research topic to design prompts for fine-tuning BERT-style PLMs in the few-shot learning setting [8, 11,20].

Motivated by this research trend, FewCLUE[1] is established as a new benchmark of few-shot learning for Chinese Language Understanding Evaluation. It contains nine challenging few-shot NLP tasks, covering a wide range of topics, such as sentiment analysis, natural language inference, keyword recognition and coreference resolution. To solve the FewCLUE tasks, we first train a large-scale PLM that digests the world knowledge from Knowledge Graphs (KGs), with the resulting model named KEBERT. After continual pre-training, KEBERT is then adapted to specific downstream tasks based on the proposed Fuzzy-PET few-shot learning algorithm. Specifically, Fuzzy-PET employs the Fuzzy Verbalizer Mapping (FVM) mechanism that gives the underlying PLM more generalization power during few-shot learning. The results show that our approach effectively solves FewCLUE tasks, producing the highest score among all the teams in the competition.

In summary, we make the following contributions in this paper:

- We introduce a novel knowledge-enhanced pre-trained model named KEBERT, which digests both the unstructured knowledge from massive text corpora and the structured knowledge from KGs.
- We propose a few-shot learning algorithm named Fuzzy-PET to improve the generalization abilities of PLMs for few-shot learning.
- Our solution achieves the best performance in both limited and unlimited tracks of FewCLUE, which will be released to public.

The rest of this paper is as follows. Section 2 gives a brief overview on related work. Section 3 presents our solution to FewCLUE tasks based on KEPLMs and few-shot learning. Experimental results are reported in Sect. 4. Finally, we give the concluding remarks and discuss possible extensions in Sect. 5.

2 Related Work

In this section, we summarize the related work on three aspects: PLMs, KEPLMs and few-shot learning for PLMs.

[1] https://github.com/CLUEbenchmark/FewCLUE.

2.1 Pre-trained Language Models

Recently, PLMs have achieved significant improvements on various NLP tasks based on the "pre-training and fine-tuning" paradigm [17]. Among these PLMs, BERT [7] is probably most influential, which learns bidirectional contextual representations by transformer encoders. RoBERTa [12] improves the pre-training process of BERT by several optimization techniques such as dynamic masking, larger sequence length and byte-level byte-pair encoding (BPE). Other PLMs based on transformer encoder architectures include ALBERT [10], Transformer-XL [6], XLNet [29], StructBERT [25] and many others. Apart from transformer encoders, the encoder-decoder architectures of transformers have also been exploited in PLMs for modeling generative NLP tasks. Typical PLMs of this type include T5 [18], UniLM [1], etc.

As pre-training large-scale PLMs is computationally expensive, a lot of efforts have also devoted into efficient distributed pre-training. Mixed precision [13] uses half-precision or mixed-precision representations of floating points for model training. This technique can be further improved through Quantization Aware Training [9], where the weights are quantized during training and the gradients are approximated with the straight-through estimator. Additionally, gradient checkpointing [4] is also frequently applied to save memory by extra computation. 3D parallelism [19] combines model parallelism (tensor slicing) and pipeline parallelism with data parallelism in complex ways to efficiently scale models by fully leveraging computing resources in clusters.

2.2 Knowledge-Enhanced Pre-trained Language Models

As shown in various studies, PLMs pre-trained on large-scale unstructured corpora only can capture the basic lexical and syntactical knowledge of languages. However, the lack of semantic information of important entity mentions in texts may affect the performance of these PLMs significantly [26]. Recently, KEPLMs are proposed to utilize the structured knowledge from KGs to enhance the language understanding abilities of PLMs. Here, we summarize the recent KEPLMs into the following two types.

The first type is knowledge enhancement by entity embeddings. In the literature, ERNIE-THU [32] injects entity embeddings into the deep language token representations via a knowledge-encoder and text-encoder modules. Entity embeddings are obtained by the existing knowledge embedding algorithms, such as TransE [2]. KnowBERT [14] uses the knowledge attention and recontextualization technique (KAR) and entity-linking mechanisms to inject the knowledge embeddings to PLMs. The goal of entity linking here is to inject the knowledge into the PLMs with higher accuracy and less noise.

The other type of approaches can be categorized as knowledge enhancement by relation triple descriptions. These works encode knowledge description texts into PLMs, which refer to the texts converted from relation triples to replace the large-scale entity embeddings. For example, E-BERT [31] and KEPLER [26] encode entity description texts through the general text encoder such as the

transformer encoder [22]. This method learns context-aware token representations and knowledge representations jointly into a unified semantic space.

2.3 Few-Shot Learning for Pre-trained Language Models

The emergence of the ultra-large PLM GPT-3 [3] shows that it has the few-shot learning abilities with prompting texts provided. For BERT-style models, Pattern-Exploiting Training (PET) [20] converts a variety of few-shot NLP tasks into cloze questions, with manually defined patterns (also called prompts in following works) as additional inputs. The fine-tuned PLMs generate predicted masked language tokens that are further mapped into class labels by pre-defined mappings. To construct prompts automatically, Gao et al. [8] generates prompts for PLMs from the T5 model [18]. AutoPrompt [21] is an automated approach to generate prompts using token-based gradient searching. These approaches focus on discrete prompts in the form of natural languages only. P-tuning [11] learns continuous prompt embeddings, which can be optimized with fully differentiable parameters. Our work on few-shot learning is extended from PET [20] and also considers fuzzy verbalizers to make the underlying PLMs more generalized to unseen data instances during training, hence improving the testing performance.

3 The Proposed Approach

In this section, we begin with a brief summary on FewCLUE tasks. After that, detailed techniques of pre-training and few-shot learning are elaborated.

3.1 Task Description

FewCLUE is a new benchmark of few-shot learning for Chinese Language Understanding Evaluation. It styles after CLUE [27] and SuperGLUE [23] with nine challenging tasks covering a wide range of language understanding topics including text classification, language inference, idiom comprehension and co-reference resolution, etc. Each task consists of five independent labeled subsets, each of which has 16 examples per class for training and the same amount of data for validation, as well as additional unlabeled examples. Models are evaluated based on the averaged test performance trained over five subsets for each task. Table 1 presents the summary of FewCLUE tasks.

3.2 Knowledge-Enhanced Pre-training

As reported in [26], PLMs pre-trained on texts may only "understand" the literal meanings of input texts, lacking the deep understanding of the background knowledge of entities and relations. Without additional knowledge, it is challenging for the underlying PLM to solve few-shot learning tasks with high accuracy, which is the case for FewCLUE tasks.

Table 1. A brief summary of FewCLUE tasks.

Dataset	Train	Dev	Test public	Test private	Labels	Unlabeled	Task
Single sentence classification tasks							
eprstmt	32	32	610	753	2	19565	Sentiment analysis
csldcp	536	536	1784	2999	67	18111	Long text classification
tnews	240	240	2010	1500	15	20000	Short text classification
ifytek	928	690	1749	2279	119	7558	Long text classification
Sentence pair classification tasks							
ocnli	32	32	2520	3000	3	20000	NLI
bustm	32	32	1772	2000	2	4251	Semantic similarity
Reading comprehension tasks							
chid	42	42	2002	2000	7	7585	Multiple choice (idiom)
csl	32	32	2828	3000	2	19841	Keyword recognition
cluewsc	32	32	976	290	2	0	Coreference resolution

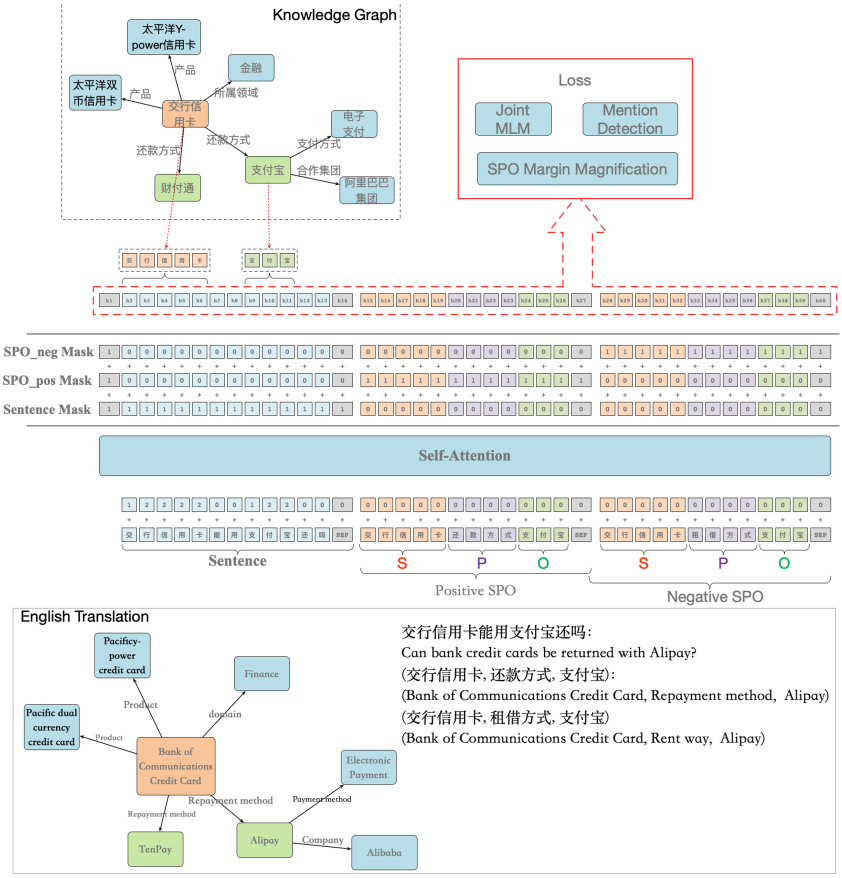

Fig. 1. The high-level architecture of KEBERT.

In this work, we introduce a new Knowledge-Enhanced BERT named KEBERT. The high-level architecture is presented in Fig. 1. To encode knowledge effectively, each pre-training instance is in the form of "Text-Positive SPO-Negative SPO" triples, where SPO refers to the Subjective-Property-Object relation triples stored in KGs. The knowledge of the positive SPO is semantically consistent with the input text, while the negative SPO is related to the input text but provides incorrect information. To facilitate knowledge understanding, three new pre-training tasks are proposed:

- Mention detection, which enhances the KEPLM's understanding abilities of entity mentions in the text;
- Joint MLM (Masked Language Modeling) of text and knowledge, which emphasizes the information sharing process between the structured knowledge and the unstructured texts, and improves the semantic understanding abilities of the model;
- SPO margin magnification, which is designed to widen the semantic gap of representations between the positive SPO and the negative SPO, and makes the underlying PLM be aware of the correctness of the knowledge.

In the future, we will release more technical details of KEBERT to public.

3.3 Continual Pre-training

We further design three variants of the MLM pre-training tasks to improve the robustness of the model over different few-shot tasks.

Whole-Word Masking. The whole-word masking strategy [5] is widely used for pre-training Chinese PLMs. Here, we use all unlabeled texts of each task as the pre-training corpus for whole-word masking. The purpose is to make the model more adaptive to the task-related data.

EFL-Based Pre-training. Previous work has shown the effectiveness of intermediate training on the GLUE benchmark [15], as well as few-shot matching tasks [30], which is a continual pre-training method that supplements PLMs with intermediate supervised tasks. Specifically, the Entailment as Few-Shot Learner (EFL) algorithm [24] is proposed to reformulate all language understanding tasks as entailment tasks. In our work, we follow the EFL baseline provided by the CLUE committee and use the CMNLI dataset [27] for continual pre-training. However, instead of directly using the "[CLS]" head to predict class labels, we create a prompt and a verbalizer (following PET [20]), to re-design the task as an MLM task, which is well aligned with other pre-training tasks.

MRC-Enhanced MLM. During the algorithm design process, we find that the PET algorithm [20] trained over vanilla KEPLMs may perform poorly over Machine Reading Comprehension (MRC) tasks. We argue that the problem is that PLMs are insensitive to the prompt formats of these tasks. Thus, we propose a variant of the MLM task to enhance the PLM's power for MRC tasks. For

each sentence, we extract the keywords based on TF-IDF scores, mask one of the keywords and randomly create several false options with the same length of the masked word. Next, we create a special prompt to force the model to choose the correct word to fill in the masked position. An example is shown in Table 2.

Table 2. An example of MRC-enhanced MLM. English translations of Chinese texts are also provided.

Input Text	卓越的生命之花需要用美育浇灌。 (The exceptional flower of life needs to be watered by aesthetic education.)
Masked Word	浇灌 (watered)
Other Options	篮球，苹果，跳跃 (basketball, apple, jump)
Output Text	卓越的生命之花需要用美育[MASK][MASK]。候选词：篮球，苹果，浇灌，跳跃 (The exceptional flower of life needs to be [MASK] by aesthetic education. Candidates: basketball, apple, watered, jump)

3.4 Fuzzy-PET Algorithm for Few-Shot Fine-Tuning

For the few-shot fine-tuning process for a specific task, we propose the Fuzzy-PET algorithm by extending PET [20], in order to improve the model's generalization abilities during inference time. Specifically, Fuzzy-PET employs the *Fuzzy Verbalizer Mapping* (FVM) mechanism that allows multiple masked language tokens to be mapped into the same class label for one pattern. Let V be the full vocabulary set, L be the class label set, and $\mathbf{x}$ be an arbitrary input text sequence. We further define a pattern function P as a mapping from $\mathbf{x}$ to $P(\mathbf{x})$ such that $P(\mathbf{x})$ is a text sequence consisting of $\mathbf{x}$, the prompting texts and a masked language token.[2] Hence, the textual input to our PLM w.r.t. the text sequence $\mathbf{x}$ is $P(\mathbf{x})$.

To facilitate the many-to-one mappings from multiple masked language tokens to the same class label, in contrast to [20], we define the *Reverse Fuzzy Verbalizer* (RFV) f as a function: $f : v \rightarrow l$ such that $v \in V$ is the predicted result of the masked language token and $l \in L$. The RFV f gives a one-to-one mapping from a masked token to a class label. By handcrafting a collection of RFVs $F = \{f_1, f_2, \cdots\}$, it is straightforward to establish the multiple fuzzy mappings from masked tokens to class labels. Formally, denote $M(v|P(\mathbf{x}))$ as the un-normalized score that the underlying PLM assigns to $v \in V$ at the masked position. F_l is a subset of the RFV collection F such that the class label of each

[2] For the Chinese language, we can use multiple masked tokens to generate model outputs in the form of multiple Chinese characters. For simplicity, in the algorithm description, we assume there is only one masked token.

RFV $f \in F$ is $l \in L$. Given the input $P(\mathbf{x})$ and the class label l, the sum of un-normalized scores $s(\mathbf{x}, l)$ for the class label l is as follows:

$$s(\mathbf{x}, l) = \sum_{f \in F_l} M(v|P(\mathbf{x})). \tag{1}$$

Hence, the predicted probability distribution $\Pr(l|\mathbf{x})$ over all the class labels is then defined by the softmax function:

$$\Pr(l|\mathbf{x}) = \frac{e^{s(\mathbf{x},l)}}{\sum_{l' \in L} e^{s(\mathbf{x},l')}}. \tag{2}$$

During the training process of Fuzzy-PET, we use the cross-entropy loss of the true and predicted probability distributions as the loss function to fine-tune the model.

4 Experiments

In this section, we present the experimental results on the FewCLUE tasks. Detailed analysis of our approach is also provided.

4.1 Experimental Details

We conduct our experiments using the PyTorch version of the EasyTransfer toolkit [16]. In the pre-training steps, we use CLUECorpus2020 [28], which contains 100 GB Chinese raw texts retrieved from Common Crawl.[3] We also employ 100 million high-quality Chinese knowledge SPOs from our in-house KG as the structured knowledge source. We pre-train two KEBERT models for limited and unlimited tracks of FewCLUE, namely KEBERT$_{large}$ and KEBERT$_{xlarge}$. The model configurations are shown in Table 3. All models are pre-trained with the described objectives. The training for all experiments are parallelized across 32 A100 GPUs, with the techniques of mixed precision, gradient checkpointing and 3D parallelism applied.

Table 3. Detailed model configurations.

Model	Layers	Attention head	Hidden size	Filter size	Total params
KEBERT$_{large}$	24	16	1024	4096	340M
KEBERT$_{xlarge}$	24	32	2048	8192	1.3B

During few-shot learning, we apply grid search to tune the best learning rate and the batch size for each individual task. The search space of the learning rate is from 1e−4 to 5e−5, and the search space of the batch size is from 4 to 32. The maximum learning epoch is 80 for all the tasks.

[3] https://commoncrawl.org/.

Table 4. The overall performance of KEBERT$_{xlarge}$ and KEBERT$_{large}$ over the private test set.

Task	eprstmt	csldcp	tnews	iflytek	ocnli	bustm	chid	csl	cluewsc	Score
Third place	**87.73**	60.26	**73.07**	45.25	66.18	70.40	57.30	56.94	60.76	63.11
Second place	85.50	59.17	72.84	44.04	67.77	**74.16**	58.05	60.75	**66.69**	64.53
KEBERT$_{large}$	85.92	57.60	72.69	44.78	70.72	61.45	**73.83**	**68.37**	59.93	65.33
KEBERT$_{xlarge}$	85.84	**60.78**	71.67	**46.07**	**71.86**	69.21	72.51	65.85	56.83	**66.13**

4.2 Experimental Results

In the competition, we win the first place in both limited and unlimited tracks of FewCLUE. Specifically, KEBERT$_{large}$ is used for the limited track submission, where the size of the PLM should be no larger than the RoBERTa-large model [12]. KEBERT$_{xlarge}$ is for the unlimited track submission. Table 5 lists the patterns that we use for each task. Please note that we apply variants of the PET algorithm on the *chid* and *cluewsc* tasks for better performance. For the *chid* task, four "[MASK]" tokens are used to replace the "#idiom" placeholders in the texts. The pattern and all idiom candidates are appended to the end of the text and the model is forced to predict the correct idiom in the masked position. For the *cluewsc* task, we replace the target possessive pronouns or adjectives with "[MASK]" tokens. The model should learn to predict the marked entity if the co-reference is true, and "[UNK]" tokens otherwise.

Table 5. Patterns for all FewCLUE tasks. English translations are also provided.

Task	Pattern
eprstmt	[MASK]满意，[TEXT] [MASK] satisfied, [TEXT]
csldcp	这篇论文阐述了[MASK][MASK]主题。[TEXT] This paper explains the [MASK][MASK] topic. [TEXT]
tnews	以下新闻的主题是[MASK][MASK]。[TEXT] The topic of the following news is [MASK][MASK]. [TEXT]
iflytek	作为一款[MASK][MASK]应用，[TEXT] As an application of [MASK][MASK], [TEXT]
ocnli	"[TEXT1]"和"[TEXT2]"的关系是[MASK][MASK]。 The relations between "[TEXT1]" and "[TEXT2]" is [MASK][MASK].
bustm	"[TEXT1]"和"[TEXT2]"意思[MASK]同。 The meanings of "[TEXT1]" and "[TEXT2]" are [MASK].
chid	[TEXT][MASK][MASK][MASK][MASK][TEXT] 候选词：[WORD1],[WORD2],[WORD3]... [TEXT][MASK][MASK][MASK][MASK][TEXT] Candidate words: [WORD1],[WORD2],[WORD3]...
csl	[TEXT]，关键词：[WORD1],[WORD2],[WORD3]...，答案: [MASK] [TEXT], Keywords: [WORD1],[WORD2],[WORD3]...，Answer: [MASK]
cluewsc	[TEXT][MASK][MASK][MASK][TEXT]

Our results (i.e., our final submissions) over the private test set are shown in Table 4, together with the performance of models submitted by the second and the third best candidates. Overall, our model KEBERT$_{large}$ outperforms the second best candidate in 5 out of 9 tasks. The large model KEBERT$_{xlarge}$ achieves the state-of-the-art performance in the FewCLUE benchmark, outperforming the second candidate by 1.6 point.

Table 6. Comparison of KEBERT with baseline algorithms over the public test set. The performance of baselines is reported by the CLUE committee.

Task	eprstmt	csldcp	tnews	iflytek	ocnli	bustm	chid	csl	cluewsc	**Avg.**
Fine-tuning [12]	63.2	35.7	49.3	32.8	33.5	55.5	15.7	50.0	49.6	42.8
PET [20]	87.2	56.9	51.2	35.1	43.9	64.0	61.3	55.0	50.0	56.1
P-tuning [11]	88.5	44.4	48.2	32.0	35.0	65.4	57.6	50.0	51.0	52.5
EFL [24]	85.6	46.7	53.5	44.0	67.5	67.6	28.2	61.6	54.2	56.5
KEBERT$_{large}$	88.72	58.14	61.79	49.96	71.05	61.95	73.86	67.35	**64.45**	66.36
KEBERT$_{xlarge}$	**89.27**	**60.18**	**64.30**	**51.01**	**72.32**	**67.63**	**73.89**	**67.92**	61.86	**67.60**

We also compare KEBERT against several baseline algorithms for few-shot learning, including RoBERTa fine-tuning [12], PET [20], P-tuning [11] and EFL [24]. The results over the public test set are presented in Table 6. As seen, the improvement gained by KEBERT is also consistent across all the tasks.

We further conduct an ablation study to test the effectiveness of the Fuzzy-PET algorithm, with results shown in Table 7. We compare the results of Fuzzy-PET against the vanilla PET algorithm [20]. The results show that our proposed Fuzzy-PET algorithm can both reduce the human effort to select best verbalizers and achieve higher performance in few-shot tasks.

Table 7. Ablation study of Fuzzy-PET over the public test set on KEBERT$_{large}$.

Task	eprstmt	csl	bustm
Vanilla PET [20]	87.87	66.92	60.70
Fuzzy-PET	**88.66**	**67.35**	**61.95**

5 Concluding Remarks

In this paper, we present our solution to the FewCLUE benchmark, which achieves the first place among all teams. Specifically, we propose the large-scale knowledge-enhanced pre-trained model named KEBERT to digest the relation triples from KGs, and the Fuzzy-PET algorithm for few-shot learning, together with continual pre-training techniques. The ablation studies have shown the importance of different integral parts of our solution. In the future, we will release more technical details to public and apply our approach to more scenarios and NLP tasks.

References

1. Bao, H., et al.: UniLMv2: pseudo-masked language models for unified language model pre-training. In: ICML, vol. 119, pp. 642–652 (2020)
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. Brown, T.B., et al.: Language models are few-shot learners. In: NeurIPS (2020)
4. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. CoRR abs/1604.06174 (2016)
5. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G.: Pre-training with whole word masking for Chinese BERT. CoRR abs/1906.08101 (2019)
6. Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R.: Transformer-XL: attentive language models beyond a fixed-length context. In: ACL, pp. 2978–2988 (2019)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp. 4171–4186 (2019)
8. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. CoRR abs/2012.15723 (2020)
9. Jacob, B., et al.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: CVPR, pp. 2704–2713 (2018)
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. In: ICLR (2020)
11. Liu, X., et al.: GPT understands, too. CoRR abs/2103.10385 (2021)
12. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019)
13. Micikevicius, P., et al.: Mixed precision training. CoRR abs/1710.03740 (2017)
14. Peters, M.E., et al.: Knowledge enhanced contextual word representations. In: EMNLP, pp. 43–54 (2019)
15. Phang, J., Févry, T., Bowman, S.R.: Sentence encoders on stilts: supplementary training on intermediate labeled-data tasks. CoRR abs/1811.01088 (2018)
16. Qiu, M., et al.: EasyTransfer - a simple and scalable deep transfer learning platform for NLP applications. CIKM 2021 (2020). https://arxiv.org/abs/2011.09463
17. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: a survey. CoRR abs/2003.08271 (2020)
18. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020)
19. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters. In: SIGKDD, pp. 3505–3506. ACM (2020)
20. Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In: EACL, pp. 255–269 (2021)
21. Shin, T., Razeghi, Y., IV, R.L.L., Wallace, E., Singh, S.: AutoPrompt: eliciting knowledge from language models with automatically generated prompts. In: EMNLP, pp. 4222–4235 (2020)
22. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
23. Wang, A., et al.: SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In: NeurIPS, pp. 3261–3275 (2019)
24. Wang, S., Fang, H., Khabsa, M., Mao, H., Ma, H.: Entailment as few-shot learner. CoRR abs/2104.14690 (2021)

25. Wang, W., et al.: StructBERT: incorporating language structures into pre-training for deep language understanding. In: ICLR (2020)
26. Wang, X., Gao, T., Zhu, Z., Liu, Z., Li, J., Tang, J.: KEPLER: a unified model for knowledge embedding and pre-trained language representation. CoRR abs/1911.06136 (2019)
27. Xu, L., et al.: CLUE: a Chinese language understanding evaluation benchmark. In: COLING, pp. 4762–4772 (2020)
28. Xu, L., Zhang, X., Dong, Q.: CLUECorpus 2020: a large-scale Chinese corpus for pre-training language model. CoRR abs/2003.01355 (2020)
29. 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)
30. Yin, W., Rajani, N.F., Radev, D.R., Socher, R., Xiong, C.: Universal natural language processing with limited annotations: try few-shot textual entailment as a start. In: EMNLP, pp. 8229–8239 (2020)
31. Zhang, D., et al.: E-BERT: a phrase and product knowledge enhanced language model for e-commerce. CoRR abs/2009.02835 (2020)
32. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. In: ACL, pp. 1441–1451 (2019)