A Domain Knowledge Enhanced Pre-Trained Language Model for Vertical Search: Case Study on Medicinal Products

Kesong Liu* and **Jianhui Jiang** and **Feifei Lyu**
Alibaba Group
{kesong.lks,huidong.jjh,lvfeifei.lff}@alibaba-inc.com

Abstract

We present a biomedical knowledge enhanced pre-trained language model for medicinal product vertical search. Following ELECTRA's replaced token detection (RTD) pre-training, we leverage biomedical entity masking (EM) strategy to learn better contextual word representations. Furthermore, we propose a novel pre-training task, *product attribute prediction* (PAP), to inject product knowledge into the pre-trained language model efficiently by leveraging medicinal product databases directly. By sharing the parameters of PAP's transformer encoder with that of RTD's main transformer, these two pre-training tasks are jointly learned. Experiments demonstrate the effectiveness of PAP task for pre-trained language model on medicinal product vertical search scenario, which includes query-title relevance, query intent classification, and named entity recognition in query.

1 Introduction

Pre-trained language models (PLMs) have significantly improved the performance of various natural language processing (NLP) tasks in the recent years. It is now a common practice to adapt the pretrain-then-finetune approach in NLP. PLMs such as BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) capturing word meaning through self-supervised learning from large corpus have shown a significant improvement on various text mining tasks. In the biomedical domain, many BERT variants such as BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021) and BioMedBERT (Chakraborty et al., 2020), follow either continual pre-training or pre-training from scratch approach using domain specific corpora to further improve the model performance.

Our motivation is to apply powerful pre-trained language models on medicinal product vertical search engine to solve query understanding and query title relevance tasks. Most of the models trained on either the general corpus or the medical literature corpus lack medicinal product knowledge. For the product vertical search scenario, medicinal product information is usually stored in structured relational database tables, and traditional pre-trained language models focus on natural language text in the form of sentences without considering the semantic relationship modeling of structured text in the product information tables. From the perspective of users' search habits, in addition to searching directly for the medicine name, users also often search for disease, symptoms and other important proudct attribute words to find the medicine. Therefore, we propose a novel pre-training task called *product attribute prediction* (PAP). Although we could carefully craft a medicinal product knowledge graph from product databases and then try to inject the extracted explicit knowledge graph into the pre-trained language model, we argue that the procedure of building product knowledge graph is laborious but avoidable by training the language model directly on product structural information.

In this paper, we propose a novel ELECTRA(Clark et al., 2020) based biomedical knowledge enhanced pre-trained language model. It consists of two pre-training tasks: replaced token detection and product attribute prediction. Our approach is inspired by ELECTRA and TransE(Bordes et al., 2013) methods but distinguish itself in two prominent ways. Firstly, we use entity masking strategy for biomedical text instead of only masking random tokens. Text spans of biomedical named entities are masked dynamically before each training iteration. We let the model predict whether these terms are replaced to incorporate biomedical domain knowledge. Secondly, we utilize medicinal product textual information instead of node identifiers in product knowledge graphs, to further bring rich medicinal product knowledge into the

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 1014–1023
October 12–17, 2022.

pre-trained language model. The triples (product title, attribute name, attribute values), which can be easily drawn from medicinal product databases, are all encoded and then used in the contrastive loss of PAP pre-training task to capture product knowledge.

Our main contributions can be summarized as follows: 1) We augment ELECTRA's replaced token detection pre-training task by leveraging biomedical entities masking (EM) to learn better contextual word representation; 2) We propose a novel pre-training task, product attribute prediction (PAP), which can inject medicinal product knowledge into the pre-trained language model by exploiting medicinal product databases directly. The proposed pre-training task is also applicable to vertical search scenarios for products in general, not limited to medicinal products; and 3) We have demonstrated the effectiveness of PAP pre-training task for PLMs in medicinal product vertical search scenario.[1]

2 Related Work

2.1 Pre-trained Language Model

Recently pre-trained language models have dominated many NLP tasks by pre-training on a large corpus of text followed by fine-tuning on a specific task. ELMo (Peters et al., 2018) learns the contextual representations based on a bidirectional language model (biLM) with forward and backward LSTM layers. GPT (Radford et al., 2018) as an effective pre-trained generative model predicts the next token based on the left-hand side context by adapting the transformer. GPT-2 (Radford et al., 2019) brings task information to the pre-training process and adopt the model to zero-shot tasks. GPT-3 (Brown et al., 2020) further improves task-agnostic, few-shot performance and produce human-like texts. BERT (Devlin et al., 2019) presents a bi-directional LM to predict the masked tokens and demonstrates strong performance on a wide range of NLP benchmarks. RoBERTa (Liu et al., 2019) shows that more careful parameter tuning on more data can benefit PLMs. ALBERT (Lan et al., 2019) uses weight sharing and embedding factorization to reduce memory consumption and improve training speed. XLNet (Yang et al., 2019) as a permutation language model predicts masked tokens in a permuted order in a auto-regressive way. T5 (Raffel et al., 2020) and BART (Lewis et al., 2020a) adopts denoising sequence-to-sequence pre-training method. ELECTRA (Clark et al., 2020) introduce a more sample-efficient pre-training task called replaced token detection, which is replacing some tokens with plausible alternatives and predicting whether each token was replaced or not. MacBERT (Cui et al., 2020) adopts MLM as correction (Mac) and achieve state-of-the-art performances on several Chinese NLP tasks.

2.2 Knowledge Enhanced and Domain Specific Pre-trained Language Model

ERNIE (Zhang et al., 2019b) utilize pre-processed knowledge embeddings of entity mentions in text. KnowBert (Peters et al., 2019) uses retrieved relevant entity embeddings and word-to-entity attention to update contextual word representations. K-ADAPTER (Wang et al., 2021b) integrates knowledge into PLM with neural adapters. E-BERT (Poerner et al., 2020) adds aligned entity embeddings into BERT without additional pre-training. Joint representation learning of words and entities (Zhang et al., 2019a, 2021) leverage external Knowledge Graphs. BioBERT (Lee et al., 2020) is pre-trained on PubMed and PubMed Central articles. SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021) and Bio-LM (Lewis et al., 2020b) have shown that pre-training from scratch with domain-specific data can improve performance. KeBioLM (Yuan et al., 2021) and UmlsBERT (Michalopoulos et al., 2021) leverage UMLS knowledge bases during the pre-training. Domain specific pre-training (Wang et al., 2021d) has also been employed for biomedical literature search problems. Moreover, (Wang et al., 2021a) gives a systematic survey for biomedical domain PLMs.

For Chinese medical text mining, MC-BERT (Zhang et al., 2020) introduces a conceptualized representation learning approach for Chinese biomedical corpora and a Chinese Biomedical Language Understanding Evaluation benchmark (ChineseBLUE). EMBERT (Cai et al., 2021) is an entity-level knowledge-enhanced pre-trained language model, which leverages several distinct self-supervised tasks. BioHanBERT (Wang et al., 2021c), as a hanzi-aware PLM, utilizes component-level internal semantic information of Chinese characters to enhance the semantics of Chinese biomedical concepts and terminologies.

[1] Our code and models are publicly available on Github: https://github.com/liuks/ep_plm

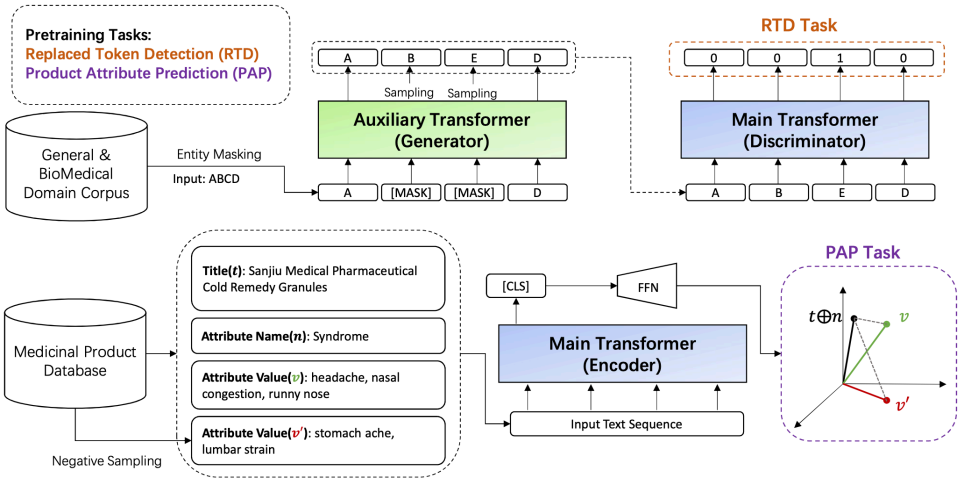

Figure 1: The overview of our pre-training model architecture. The auxiliary transformer is pre-trained by masked language model with biomedical entity masking. The corrupted text is then used as the main transformer's input in replaced token detection task. The same main transformer encodes medicinal product title, attribute name, attribute value and phrase level negative sampled attribute value, respectively. These encoded embeddings are then used in product attribute prediction task.

3 Method

To augment the pre-trained language model with biomedical domain knowledge, we follow ELECTRA's replaced token detection (RTD) pre-training task on general domain and biomedical specific corpora, and further introduce dynamic biomedical entity word masking to learn better contextual word representations. Although it is possible to use other pre-training tasks such as the masked token prediction in BERT, we chose the ELECTRA's pre-training task due to the sample efficiency of the RTD task and the comparable performance of ELECTRA-Base and BERT-large models(Clark et al., 2020). For better application to the vertical search scenario of medicinal product search, it is desirable that the pre-trained language model contains product-related knowledge. Considering that important product attributes such as drug names, diseases and symptoms are common search terms and the structured information of products also contains these terms, so we propose the product attribute prediction pre-training task to model the semantic relationships of product attributes.

As shown in Figure 1, our approach consists of two pre-training tasks: replaced token detection and product attribute prediction.

3.1 Replaced Token Detection with Biomedical Entity Masking

Given an input text sequence $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ with m tokens, a text span sequence $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ with n span units is produced by applying Chinese word segment and biomedical name entity recognition. These text span units are then randomly replaced with an equal length `[MASK]` tokens to create $\mathbf{x}^{\text{mask}}$ with about 15% tokens masked out, e.g. selecting s_j and replacing every token x_i for $x_i \in s_j$ with the `[MASK]` token.

The masked sequence $\mathbf{x}^{\text{mask}}$ is then input to the auxiliary transformer (generator) to produce a corrupt sequence $\mathbf{x}^{\text{corrupt}}$ by sampling new tokens according to $\hat{x}_i \sim p_G(x_i|x^{\text{mask}})$ for i in all masked positions. Comparing the original text and the corrupt text gives the supervisory signal label sequence $\mathbf{L} = (l_1, l_2, \ldots, l_m)$, where $l_i = \mathbb{1}(x_i, x_i^{\text{corrupt}})$. The main transformer (discriminator) predicts whether tokens are replaced or not. For replaced token detection task, the loss function is:

$$L_{\text{RTD}} = -\sum_{\mathbf{x}} \Big(\log p_G(x^{\text{corrupt}}|x^{\text{mask}}) + \lambda \log p_D(x^{\text{corrupt}}, \mathrm{x}) \Big) \quad (1)$$

where p_G is generating token probability for masked-out positions in the generator network; p_D is replaced probability for all position tokens in the discriminator network; λ is the hyperparameter for balancing these two network losses.

Different from whole word masking with WordPiece in BERT, we not only randomly mask whole words but also biomedical entities including drug name, chemical name, disease, syndrome, efficacy words and so on, which can explicitly inject

biomedical domain knowledge to the pre-trained language model.

3.2 Product Attribute Prediction Task

Let t, n, v be a medicinal product title, attribute name and attribute value respectively. We encode the text description of product tile t, attribute name n and attribute value v using the same main transformer (encoder) E to obtain text representations. Inspired by the TransE (Bordes et al., 2013) and RotateE (Sun et al., 2019) models, we define the distance function of the triples according to Equation (2) and PAP loss function according to Equation (3) to push the projection of concatenated title and attribute name representation $E(t \oplus n)$ near that of attribute value representation $E(v)$ but far away from negative sampled attribute value representation $E(v')$. The projection denoted as f is implemented as a feed forward neural network layer.

$$d(t, n, v) = \|f(E(t \oplus n)) - f(E(v))\| \quad (2)$$

$$\begin{aligned} L_{\text{PAP}} = &- \log \sigma(\gamma - d(t, n, v)) \\ &- \sum_{v'} \frac{1}{k} \log \sigma(d(t, n, v') - \gamma) \end{aligned} \quad (3)$$

where $\gamma > 0$ is a margin hyperparameter; σ is the sigmoid function; v' denotes the negative randomly sampled attribute value; k is the number of negative values v' in the summation, and is chosen to be twice the number of the positive attribute values. It is worth noting that the representation of $E(t \oplus n)$ can be broadcast across all corresponding positive attribute value v and negative ones v' to accelerate computing in the implementation, because a medicinal product usually contains only one title and multiple attribute names, each of which corresponds multiple attribute values.

For example, given a medicinal product with title (*Sanjiu Medical & Pharmaceutical Cold Remedy Granules*), we firstly draw the attribute name (*syndrome*) and corresponding attribute values (*headache, fever, nasal congestion and runny nose*) from product database. Then we randomly sample negative values (*stomachache, lumbar strain*) for the given syndrome attribute. Intuitively, we would like the name of the medicine to be semantically closer to the corresponding indication of the medicine and more semantically distant from the random chosen indications. In addition, most queries in the medicinal product vertical search usually come from product titles and attribute values, so it is beneficial to model the semantic relationships between these terms using the PAP pre-training task.

Thanks to PAP contrastive loss, the semantic relations of product attributes and the original medicinal product title, which usually contains brand, drug name, ingredients, etc., is explicitly learned. Thus, we can inject medicinal product knowledge into the pre-trained language model.

3.3 Multi-task Pre-training

We pre-train the model from scratch using general and biomedical domain corpora for RTD task and medicinal product datasets for PAP task, see Equation (4).

$$L = L_{\text{RTD}} + L_{\text{PAP}} \quad (4)$$

We train the model parameters by repeatedly switching back and forth between RTD and PAP tasks. The hyper parameter ρ denotes the probability of selecting the PAP task training batch at each gradient descent iteration. The overall training procedure is shown in Algorithm 1.

Algorithm 1 Overall Training Procedure

```
Initialize model parameters randomly.
Mark biomedical entity boundaries for the general and biomedical domain corpora.
Collect triples (product title, attribute name, attribute value) from medicinal product database.
while needing more training steps do
    Select a training batch from RTD and PAP tasks randomly
    if the training batch is from RTD task then
        Sample tokens for randomly masked word spans using the auxiliary transformer.
        Calculate RDT task loss using the main transformer according to Equation (1) and update model parameters.
    else                                                    ▷ for PAP task
        Sample negative attribute values randomly for each product title and attribute name pair.
        Calculate PAP task loss using the same main transformer according to Equation (2) and (3) and update model parameters.
```

4 Experiments and Results

To demonstrate the effectiveness of the PAP pre-training task, we train the pre-trained language

Dataset	#Sen	#Tok
Wikipedia	0.5M	0.4B
News Articles	1M	2B
Package Insert of Drugs†	8K	0.9M
Medical Encyclopedia†	9K	1M
Biomedical Community QA†	39M	8B

Table 1: Statistics of pre-training corpus. Datasets with dagger symbols indicate that they are from the biomedical domain. Other datasets are from general domain.

Medicinal Product dataset	#
Product Titles	29K
Product Attribute Categories	2
Product Attribute Values	191K
Product Attribute Values per Title	6.5

Table 2: Statistics of medicinal product dataset

Dataset	Train	Dev	Test
QTRel-easy	9,303	1,000	1,000
QTRel-hard	8,941	1,000	1,000
QIC	34,929	13,234	13,234
QNER	135,411	14,047	14,047

Table 3: Statistics of PAP evaluation benchmark for medicinal product search.

model from scratch, first on the general corpus and biomedical domain corpus shown in Table 1 using only the RTD task with biomedical entity masking, and then adding the PAP task on the medicinal product dataset shown in Table 2. The reason that we use pre-training from scratch strategy instead of parameter initialization from other existing PLMs to continue pre-training is to exclude the influence caused by different datasets.

Therefore, we first train a base-size model (ELECTRA+EM) on the general and biomedical domain datasets in Table 1 as a strong baseline and compare it with other Chinese PLM models. Then we train another base-size model (ELECTRA+EM+PAP) on the common datasets in Table 1 and the medicinal product dataset in Table 2 and do ablation experiments to verify the effectiveness of the PAP pre-training task.

4.1 Pre-training Datasets

We collect general and biomedical domain specific Chinese corpus, as shown in Table 1. The general domain corpus consists of the Chinese Wikipedia dataset and Chinese News Articles dataset, which are publicly available from NLP Chinese Corpus (Xu, 2019). The biomedical domain specific Chinese corpus consists of Package Insert of Drugs, Medical Encyclopedia and Biomedical Community QA, which are from Shenma Search Engine[2]. As shown in Table 2, We construct (product title, attribute name, attribute value) dataset from vertical search medicinal product database.

4.2 Evaluation Datasets

For the ELECTRA+EM model, we use Chinese Biomedical Language Understanding Evaluation benchmark (ChineseBLUE) (Zhang et al., 2020) to demonstrate the benefits of its in-domain pre-training. The benchmark contains a variety of NLP tasks: cEHRNER and cMedQANER are two named entity recognition tasks; cMedQQ is a paraphrase identification task to determine whether two sentences have the same meaning; cMedQA and cMedQNLI are two question answering tasks that can be approximated as a ranking of candidate answer sentences based on their similarity; cMedIR is a ranking task that retrieves the most relevant documents for a given search query; cMedIC is an intent classification task that assigns three types of labels to query terms with no intention, weak intention, and firm intention; cMedTC is a text classification task that assigns multiple labels to biomedical texts. Further details about these datasets can be found in (Zhang et al., 2020).

For the ELECTRA+EM+PAP model, we construct a benchmark containing four NLP tasks from our medicinal product vertical search scenario to validate the advantage of PAP pre-training task compared to ELECTRA+EM model. As shown in Table 3, the benchmark contains four datasets: QTRel-easy and QTRel-hard are two query-title relevance tasks, where the "hard" part in dataset name means that none of the query terms appear in the corresponding medicinal product title and the "easy" dataset does not have this constraint; QIC is a query intention classification task which assigns 22 different types of labels to queries, including medicine name, disease, symptom, inquiry, etc. QNER is a named entity recognition in query task with 28 total entity types, such as brands, main ingredients of drugs, dosage forms, etc.

[2] http://m.sm.cn/

4.2.1 Parameter Settings

To compare with ELECTRA and other baseline models, we leverage the same model settings of the transformer as ELECTRA. Both ELECTRA+EM and ELECTRA+EM+PAP use the base version of ELECTRA, which contains 12 layers, 12 self-attention heads, and 768-dimensional of hidden size for the discriminator network and 1/3 generator size.

For ELECTRA+EM, we set the initial learning rate as 2e-4, batch size as 128, maximum sequence length as 128, training steps as 420M. For ELECTRA+EM+PAP, we use the same learning rate and maximum sequence length, but set batch size as 16, training steps as 3400M. For the optimizer, we use the same setting with ELECTRA, both in pre-training and fine-tuning steps.

4.3 Results

4.3.1 ELECTRA+EM

For the ELECTRA+EM model, we compare it with several typical Chinese general and biomedical domain PLM baselines, namely BERT-Base[3], ELECTRA-Base[4] (Cui et al., 2020), MC-BERT[5] (Zhang et al., 2020), EMBERT (Cai et al., 2021) and BioHanBERT (Wang et al., 2021c).

For BERT-Base, ELECTRA-Base and our ELECTRA+EM, we run the finetuning 5 times for each downstream task and report the results in *average/maximum* metric format. For MC-BERT, EMBERT and BioHanBERT, we directly cite the results from the corresponding papers. The evaluation metric of all ChineseBLUE datasets is F1 score except the cMedIR dataset whose metric is PAIR score.

As shown in Table 4, ELECTRA+EM achieves comparable performance compared to other baseline and state-of-the-art methods. The comparison also demonstrates the benefits of in-domain pre-training from scratch. We can therefore use ELECTRA+EM as a very strong baseline model to verify the effectiveness of PAP pre-training task.

4.3.2 ELECTRA+EM+PAP

We also run the finetuning 5 times for each downstream task and report the results in *average/maximum* metric format for PAP evaluation benchmark.

As shown in Tables 5, ELECTRA+EM+PAP outperforms ELECTRA-Base significantly on all four tasks. There may be two reasons for these large improvements. Firstly, medicinal product titles and queries in the benchmark usually consist of brand, disease, symptom words and phrases, which are also abundant in PAP pre-training task datasets. Secondly, PAP pre-training task leverages the semantic relation of product titles and attributes to obtain better word representations.

Hyperparamter Since ELECTRA+EM+PAP adopts a multi-task pre-training framework, it is necessary to tune the hyperparameter ρ, the probability of selecting PAP task while training. We search for the best ρ out of $[1\%, 5\%, 10\%, 15\%]$. For the margin hyperparameter γ in PAP loss, we search the best out of $[2, 4, 6, 8]$. We find the combination of $\gamma = 4, \rho = 5\%$ works best. Since results are more insensitive to the hyperparameter γ, we fix $\gamma = 4$ and then plot the effect of the hyperparameter ρ on results, as is shown in Figure 2.

Ablation Study As shown in Table 5, the method without PAP pre-training task, ELECTRA+EM, has worse performance than ELECTRA+EM+PAP. This is reasonable because the product knowledge learned by PAP pre-training task is beneficial for medicinal product search.

It is interesting that on average the performance degradation in terms of F1 score is much larger for QTRel-hard task than QTRel-easy, QIC and QNER tasks. We hypothesize that biomedical entity semantics, which plays a crucial role in QIC and QNER tasks, can be largely captured by biomedical entity masking in RTD task. The PAP pre-training task may be more beneficial for product related concept complex interaction understanding task, such as QTRel-hard dataset.

Case Study As shown in Table 6, We compare ELECTRA+EM+PAP and ELECTRA-Base on QTRel-hard task for a given query "中耳炎" ("tympanitis"). The first two column scores are the predicted probabilities that the given query-title pair is relevant. The label +/- indicates whether the query-title pair is relevant or not in real. Base on their package inserts, the indications of cefixime and azithromycin tablets include tympanitis, while mosapride citrate is not suitable for treating tympanitis. ELECTRA+EM+PAP scores align *more*

[3] https://github.com/google-research/bert

[4] https://github.com/ymcui/Chinese-ELECTRA

[5] https://github.com/alibaba-research/ChineseBLUE

Model	cEHRNER	cMedQANER	cMedQQ	cMedQA
MC-BERT	90.0	88.1	87.5	82.3
BioHanBERT(10K)	90.51	-	86.46	96.53
BioHanBERT(20K)	91.67	-	87.14	96.37
BioHanBERT(30K)	91.83	-	86.26	96.36
BioHanBERT(40K)	90.44	-	87.18	96.49
BioHanBERT(50K)	90.91	-	87.86	**96.65**
EMBERT♣	-	84.49	87.59	75.10
EMBERT♠	-	85.02	**88.06**	75.32
BERT-Base	90.19/90.50	85.05/85.29	87.06/87.43	96.03/96.09
ELECTRA-Base	91.63/92.17	86.37/86.81	87.27/87.46	95.60/95.90
ELECTRA+EM	**92.10/92.85**	**88.18/88.53**	87.56/87.89	96.55/96.78
Model	cMedQNLI	cMedIR	cMedIC	cMedTC
MC-BERT	95.5	2.04	87.5	82.1
BioHanBERT(10K)	95.86	-	90.48	81.78
BioHanBERT(20K)	95.59	-	**96.43**	83.67
BioHanBERT(30K)	95.72	-	83.33	83.00
BioHanBERT(40K)	95.50	-	90.48	82.72
BioHanBERT(50K)	95.78	-	86.90	83.06
EMBERT♣	96.50	-	-	-
EMBERT♠	96.59	-	-	-
BERT-Base	96.05/96.11	3.03/3.07	92.43/92.89	82.99/83.78
ELECTRA-Base	95.42/95.55	3.41/3.49	90.61/92.31	83.43/83.78
ELECTRA+EM	**96.66/96.78**	**3.64/3.71**	92.48/93.26	**83.73/84.00**

Table 4: Experimental results on ChineseBLUE test datasets. For the BioHanBERT model, the number in parentheses indicates the number of steps in the training step. For the EMBERT model, ♣ and ♠ indicate it is initialized by BERT-Base and MC-BERT, respectively.

Model	QTRel-easy	QTRel-hard	QIC	QNER
ELECTRA-Base	98.20/98.51	81.90/82.80	81.37/82.06	82.53/82.78
ELECTRA+EM+PAP	**98.67/98.95**	**84.88/85.86**	**86.90/87.26**	**88.28/88.72**
ELECTRA+EM	98.44/98.66	82.90/83.46	86.58/87.11	88.03/88.35
Average Drop w/o PAP	-0.23	-1.98	-0.32	-0.25

Table 5: Experimental results on PAP evaluation benchmark for medicinal product search.

ELECTRA -base	ELECTRA +EM+PAP	Label	Product Title
0.62	0.88	+	999 头孢克肟片 0.1g*7片/盒 999 Cefixime Tablets 0.1g*7tablets/box
0.49	0.87	+	999 阿奇霉素片 0.25g*6片/盒 999 Azithromycin Tablets 0.25g*6tablets/box
0.93	0.42	-	信谊 美唯宁 枸橼酸莫沙必利胶囊 5mg*24粒 SINE MeiWeiNing Mosapride Citrate Capsules 5mg*24

Table 6: Examples of query-title relevance scores on QTRel-hard task for the query “tympanitis”.

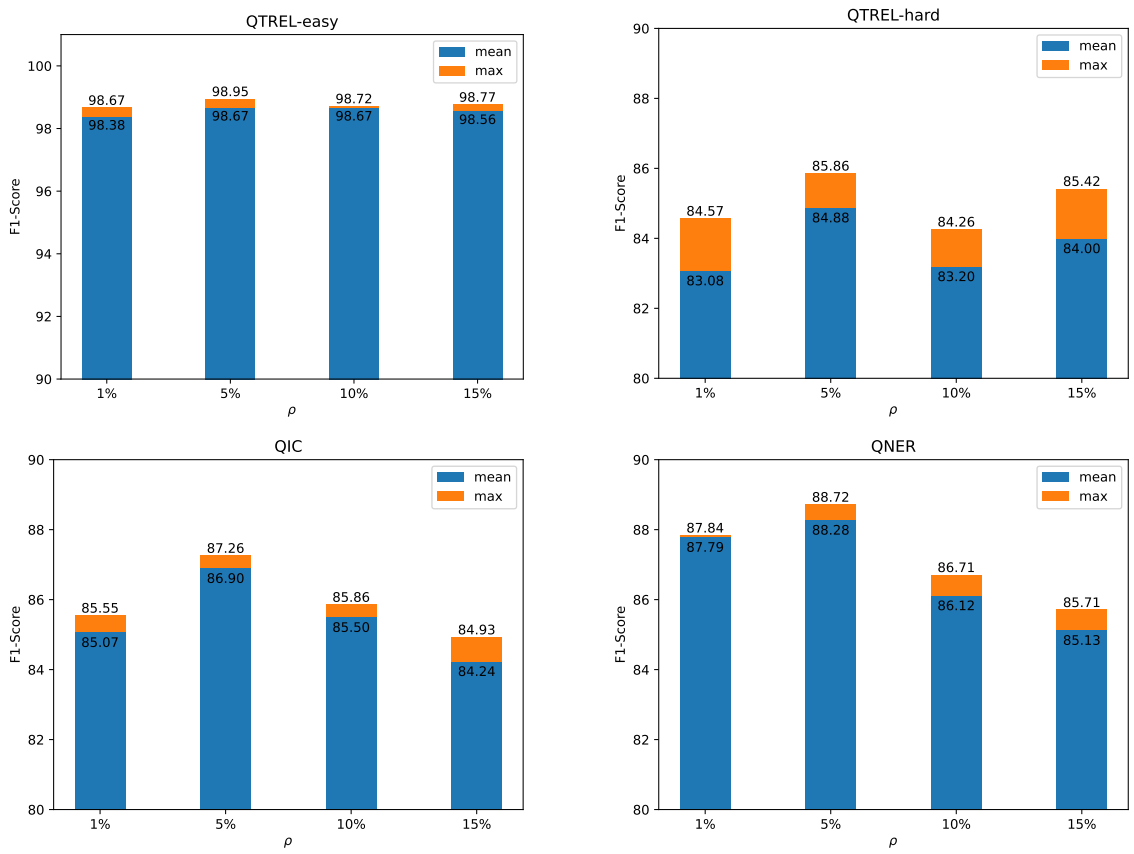

Figure 2: Hyperparameter tuning for ρ (when $\gamma = 4$) on PAP evaluation benchmark for medicinal product search.

closely with labels than ELECTRA-Base. For the first example, ELECTRA+EM+PAP is more confident about the relevance of "cefixime tablets" and "tympanitis" than ELECTRA-Base. When a threshold probability of 0.5 is applied, ELECTRA+EM+PAP succeeds in classifying the last two examples, while ELECTRA-Base fails.

5 Discussion

As mentioned in Section 3, other self-supervised learning tasks for language models could be learned together with the PAP pre-training task. We only explore the combination of ELECTRA's RTD and PAP for joint training. Due to limited computational resources, we could not push the model to larger sizes. These may prevent the full potential of PAP pre-training task from being unleashed.

For vertical search applications, PAP pre-training task is also applicable to other product searches. For example, based on movie structured information such as movie title, genre and story type, it is feasible to use PAP task to model the semantic relationships between these movie attributes. In the study of product knowledge enhanced language models, comparing the way PAP uses product structured information with the graph embedding approach based on product knowledge graphs may be an interesting research problem for the future.

6 Conclusion

In this article, we propose a biomedical knowledge enhanced pre-trained language model for medicinal product vertical search. We improve ELECTRA's replaced token detection pre-training task with biomedical entity masking (EM). Then we present a novel pre-training task, product attribute prediction (PAP), to incorporate medicinal product knowledge into the PLM. We train ELECTRA+EM and ELECTRA+EM+PAP two biomedical knowledge enhanced pre-trained language models to demonstrate the effectiveness of PAP pre-training task for medicinal product vertical search. Our work may shed some light on combining the powerful pre-trained language models with product knowledge for vertical search scenarios.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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)*, pages 3615–3620.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *In NIPS*, pages 2787–2795.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Zerui Cai, Taolin Zhang, Chengyu Wang, and Xiaofeng He. 2021. Embert: A pre-trained language model for chinese medical text mining. In *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I*, page 242–257, Berlin, Heidelberg. Springer-Verlag.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. Biomedbert: A pre-trained biomedical language model for qa and ir. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020b. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. 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)*, pages 43–54.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-bert: Efficient-yet-effective entity embeddings for bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 803–818.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021a. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021b. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Xiaosu Wang, Yun Xiong, Hao Niu, Jingwen Yue, Yangyong Zhu, and Philip S. Yu. 2021c. Biohanbert: A hanzi-aware pre-trained language model for chinese biomedical text mining. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1415–1420.

Yu Wang, Jinchao Li, Tristan Naumann, Chenyan Xiong, Hao Cheng, Robert Tinn, Cliff Wong, Naoto Usuyama, Richard Rogahn, Zhihong Shen, et al. 2021d. Domain-specific pretraining for vertical search: Case study on biomedical literature. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3717–3725.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online. Association for Computational Linguistics.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.

Ningyu Zhang, Zhanlin Sun, Shumin Deng, Jiaoyan Chen, and Huajun Chen. 2019a. Improving few-shot text classification via pretrained language representations. *arXiv preprint arXiv:1908.08788*.

Taolin Zhang, Chengyu Wang, Minghui Qiu, Bite Yang, Zerui Cai, Xiaofeng He, and Jun Huang. 2021. Knowledge-empowered representation learning for Chinese medical reading comprehension: Task, model and resources. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2237–2249, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.