ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240674

TabMedBERT: A Tabular Knowledge Enhanced Biomedical Pretrained Language Model

Xu Yan[a,1], **Lei Geng**[a,2], **Ziqiang Cao**[a,*], **Juntao Li**[a], **Wenjie Li**[b], **Sujian Li**[c], **Xinjie Zhou**[d], **Yang Yang**[d] and **Jun Zhang**[e]

[a]Institute of Artificial Intelligence, Soochow University
[b]The Hong Kong Polytechnic University
[c]Peking University
[d]Pharmcube Inc.
[e]Changping Laboratory

Abstract. Most existing biomedical language models are trained on plain text with general learning goals such as random word infilling, failing to capture the knowledge in the biomedical corpus sufficiently. Since biomedical articles usually contain many tables summarising the main entities and their relations, in the paper, we propose a **Tab**ular knowledge enhanced bio**Med**ical pretrained language model, called TabMedBERT. Specifically, we align entities between table cells, and article text spans with pre-defined rules. Then we add two table-related self-supervised tasks to integrate tabular knowledge into the language model: Entity Infilling (EI) and Table Cloze Test (TCT). While EI masks tokens within aligned entities in the article, TCT converts aligned entities in the table layout into a cloze text by erasing one entity and prompts the model to extract the appropriate span to fill in the blank. Experimental results demonstrate that TabMedBERT surpasses all competing language models without adding additional parameters, establishing a new state-of-the-art performance of 85.59% (+1.29%) on the BLURB biomedical NLP benchmark and 7 additional information extraction datasets. Moreover, the model architecture for TCT provides a straightforward solution to revise information extraction with paired entities.

1 Introduction

In recent years, biomedical natural language processing (NLP) tasks have moved a big step forward with the flourishing of biomedical pretrained language models (PLMs). For example, BioBERT [19] and PubMedBERT [12] pretrain on the PubMed[3] article abstracts or full texts. The primary training objective is masked language modeling (MLM) [8]. However, random masking does not consider biomedical entities, precisely the core of understanding medical articles. Some researchers [14, 37, 36] try to inject external knowledge into biomedical PLMs. Still, this knowledge may lack relevance to the article, and it is costly to maintain the timeliness of a biomedical knowledge base.

Biomedical articles often feature comprehensive tables that summarize and elucidate the research findings. These tables highlight the key entities and their relationships within the articles, ensuring that the extracted knowledge aligns perfectly with the corresponding content. Motivated by this, we propose TabMedBERT, a **Tab**ular knowledge enhanced bio**Med**ical pretrained language model. Our model incorporates table-aligned articles for secondary pretraining, building upon existing biomedical PLMs while maintaining compatibility with various models and without introducing additional parameters. We collect tables from various sources, including parsed PDF articles and clinical trials. We combine rules and model-based methods to align entities between table cells and corresponding spans in the articles. To incorporate the knowledge of aligned entities and their relationships into the language model, we introduce two table-related training tasks: Entity Infilling (EI) and Table Cloze Test (TCT). In the EI task, tokens within aligned entities in the article are masked and then recovered. In the TCT task, aligned entities in the table layout are converted into a cloze text by erasing one entity. The language

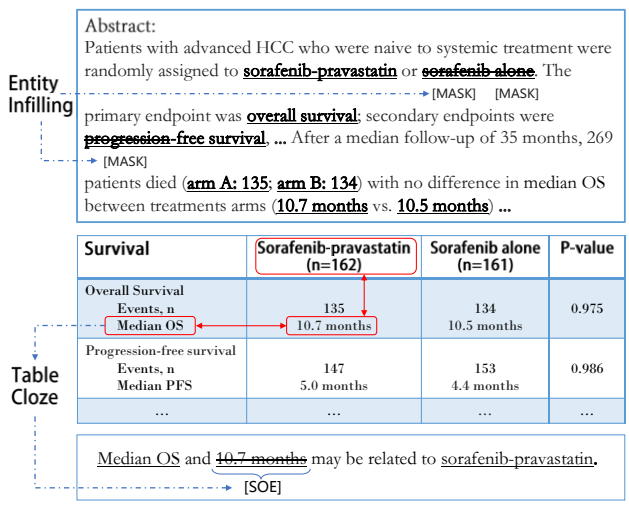

Figure 1. Examples of alignment between article text and table. The underlines represent entities in the article that align with the content of table cells. The entity infilling task involves masking tokens within aligned entities. The "red boxes" represent cells with layout correspondence. In pretraining, we leverage this correspondence to build a table cloze task by templates and erase one entity with a special [SOE] token.

* Corresponding Author. Email: zqcao@suda.edu.cn
[1] Equal contribution.
[2] Equal contribution.
[3] https://pubmed.ncbi.nlm.nih.gov/

model is then prompted to extract the appropriate article span to fill in the blank. Figure 1 provides an example of these tasks. The EI task encourages the model to grasp the fundamental concepts in the article, while TCT facilitates learning entity relations described in tables. Additionally, we have observed that biomedical articles often involve paired entities, such as endpoints and experimental results. In such cases, our TCT task serves as a direct solution to identifying the missing components of paired entities, thereby assisting information extraction (IE) tasks.

We conducted experiments on various biomedical tasks, including the BLURB biomedical NLP benchmark and many other IE tasks. The results demonstrate that our TabMedBERT outperforms all other language models, achieving a remarkable new state-of-the-art performance of 85.59% (+1.29%) on the BLURB score[4]. Notably, our TCT task introduces a novel and general approach for language models to consistently revise paired entities in biomedical articles, leading to an impressive increase of over 7%.

The contributions of our work are summarized as follows:

- We propose a groundbreaking approach that aligns tabular knowledge with articles to enhance the training of biomedical PLMs.
- We introduce two table-related self-supervised tasks: Entity Infilling (EI) and Table Cloze Test (TCT) to improve information extraction tasks by revising paired entities.
- Notably, our model achieves state-of-the-art performance on the BLURB biomedical benchmarks and various other IE tasks without any increase in parameters.

2 Related Work

2.1 Biomedical Pretrained Language Models

In recent years, the advent of pre-trained language models like BERT [8] has brought about a revolution in various downstream NLP tasks, including biomedical research. For example, SciBERT [4] incorporates science papers for training. PubMedBERT [12] extends BioBERT [19] by constructing a domain-specific vocabulary and training on PubMed abstracts. Several subsequent studies [16, 1] are further expanded to clinical medicine. Some generative models are also emerging, such as BioGPT [23]. These models only consider the plain text itself, which becomes a bottleneck in model performance. Notably, BioLinkBERT [36] incorporates the connection between biomedical documents and achieves remarkable performance on document-level biomedical tasks.

2.2 Knowledge-Enhanced Pretrained Language Models

Instead of focusing on internal plain text, researchers have explored ways to utilize external knowledge. For instance, WKLM [35] leverages entities from Wikipedia. Some researchers have focused on improving the model structure to incorporate knowledge. K-ADAPTER [33] and KnowBERT [27] introduce additional model structures beyond the middle layer of the PLMs to encode multiple KGs. In the biomedical domain, KeBioLM [37] incorporate various biomedical knowledge from the Unified Medical Language System[5]. While these methods rely on existing knowledge bases or modify the model structure, our TabMedBERT effectively incorporates overlooked but informative tabular data, surpassing previous approaches without additional parameters.

[4] The results will be submitted soon: https://microsoft.github.io/BLURB
[5] https://www.nlm.nih.gov/research/umls/index.html

3 Table Collection and Alignment

In this section, we first collect massive tables linked to biomedical articles (§ 3.1). Then, we combine rules and model-based methods to align table cells and corresponding articles (§ 3.2).

3.1 Table Acquisition

We collect tables from three sources: parsed PubMed PDF articles, the PubTables-1M dataset [29], and clinical trials sourced from NCT[6]. We download many PDF documents on PubMed and use the OCR tool[7] to extract their tables. Furthermore, we leverage the tables from the PubTables-1M dataset, which comprises nearly one million tables associated with PubMed articles. Medical literature often includes pertinent clinical trials registered in NCT as supporting evidence. We establish a connection between the tables in the experimental records and the articles by linking them through the NCT number mentioned on the PubMed website. In Table 1, we present the statistics of the above resources.

Table 1. Statistics of table sources. The number of NCT articles represents the number of clinical trial records.

Source	Articles	Total Tables	Avg. Tables Per Article
NCT	55813	167439	3.1
PubTables-1M	401733	949250	2.36
PubMed PDF	52373	223632	4.27

3.2 Table Entity Alignment

We combine rules and model-based methods to align entities[8] in table cells with the corresponding spans in the article. The rules can be outlined as follows: (i) Given a cell and its related article, we convert each non-stopword and non-punctuation word in the cell to the same part of speech and identify their corresponding positions in the article. (ii) Next, we list potential spans in the article, compare them with the cell, and select the one with the highest word overlap proportion. For example, in the filtered cell "Sequenced genome falciparum" and the span "Sequenced genome plasmodium falciparum", the proportion is 0.75.

Rule-based methods are generally proficient at resolving the majority of matching challenges. However, in scenarios where crafting specific rules proves to be a complex task, we resort to employing ChatGPT [9] to assess a subset of PubMed PDFs directly. This approach is advantageous due to its convenience and its resilience to subpar OCR outcomes. Looking ahead, we intend to delve further into model-based techniques, which may involve training models to deduce the semantic relationships between entity terms in the representation space.

4 Method

TabMedBERT is a self-supervised pretraining approach incorporating internal knowledge from tables in biomedical articles. An overview of our model is depicted in Figure 2. We employ table-related tasks to learn from tabular knowledge: Entity Infilling (EI)

[6] https://clinicaltrials.gov/
[7] https://toscode.gitee.com/paddlepaddle/PaddleOCR
[8] Here, we refer to a cell as either an entity or its attribute for convenience.
[9] https://chat.openai.com/

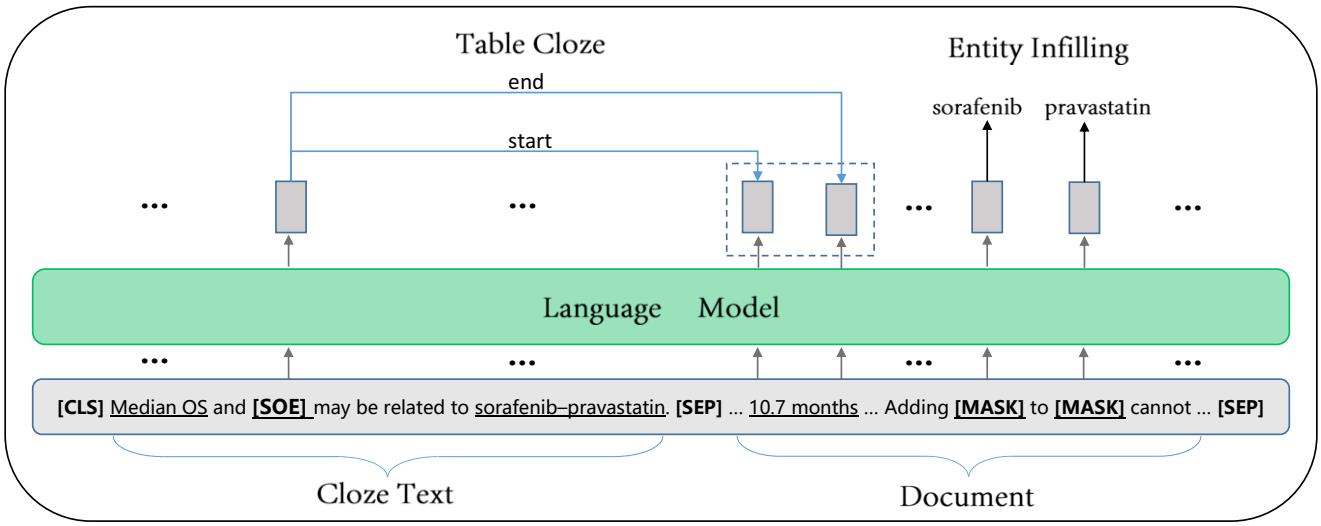

Figure 2. Overview of our TabMedBERT. Following Figure 1, the underlines represent aligned entities in the document, and then we convert aligned entities with layout correspondence into cloze text. To incorporate the tabular knowledge into LM pretraining, we design two table-related training objectives: Entity Infilling (EI) preferably masks and recovers tokens within aligned entities in the document (§ 4.1). Table Cloze Test (TCT) drives the LM to extract the correct span in the document to fill the cloze part (§ 4.2).

and Table Cloze Test (TCT). In this approach, we convert a biomedical document, comprising both the document text and its associated tables, into a structured text format as follows:

$$X = \text{[CLS]}\ X_C\ \text{[SEP]}\ X_E\ \text{[SEP]}, \tag{1}$$

X_C represents the table cloze text described in Section 4.2, and X_E denotes the document text with masked entities, as mentioned in Section 4.1. [CLS] and [SEP] are special tokens commonly used in masked language models. The input instance is then encoded using a language model encoder:

$$h = Encoder(X), \tag{2}$$

where $h = (h_1, h_2, ..., h_m)$ represents a sequence of contextualized representations for all input tokens. The tasks of EI and TCT are trained based on these representations.

4.1 Entity Infilling

Based on table alignment, we identify a subset of tokens that represent entities in the article. In biomedical information extraction, specialized terms abound, making it challenging to identify key medical entities and their relationships. We argue that entities in tables mirror medical facts in abstracts, like drugs and diseases, thus warranting prioritized masking, a proven and efficient approach. Following the established methods for masking and recovering key entities [21], we introduce the Entity Infilling task (EI) to mask tokens from the identified subset and recover their original values. The selection of candidate tokens to mask follows these strategies:

- We prioritize considering the entire entity in the entity subset as candidates. However, the number of tokens selected from the same entity should not exceed 3.
- Tokens belonging to the entity that coincides with the answer in the cloze test task (§ 4.2) must be excluded.
- If the number of candidate tokens is insufficient, we randomly select tokens from the article to supplement.

Similar to the normal MLM task [8], we guarantee that the proportion of candidate tokens for masking accounts for 15% in X_E; of those, 80% are replaced with [MASK], 10% with a random token, and 10% are kept unchanged. After masking, we denote the document text as X_E. The training of the EI task involves optimizing the following loss function:

$$L_{EI} = -\sum_i \log p(x_i|h_i), \tag{3}$$

where x_i represents the original token at each masked position in X_E, and h_i denotes its corresponding representation.

4.2 Table Cloze Test

To incorporate the relationship between aligned entities into the pretraining process, we introduce the Table Cloze Test (TCT) task. This task converts the aligned entities into cloze templates based on their extracted relationships from the table layout. By randomly removing one entity from the cloze template, the TCT task prompts the language model to identify the most suitable entity span in the document to fill in the blank. This approach enables the language model to simultaneously learn about the essential entities in the article and their relationships. We will first explain how we construct the cloze text and then provide an overview of the model architecture used for extracting the entity span.

4.2.1 Cloze Construction

Table 2. Templates designed for the table cloze test task. Ent1/2/3 are the placeholders that entities will replace.

Entities	Templates
2	Ent1 is associated with Ent2.
3	Ent1 and Ent2 may be related to Ent3.

By analyzing the layout of cells, we can extract relationships between aligned entities in the article. We consider the following two cases for determining these relationships:

- Single Column or Row: When a heading cell is present in a single column or row, it is directly related to other cells in that column or row. For instance, in Figure 1, Sorafenib alone is directly related to 134.
- Cells with Row and Column Headings: A cell with row and column heading cells is associated with specific entities. In boxes of Figure 1, Median OS and 10.7 months are associated with Sorafenib-pravastatin.

It is important to highlight that in medical clinical trials, the intervention methods, measurement indicators, and resulting measurements all play vital roles in accurately describing medical events. For instance, in specific tasks like PICO [26] and relation extraction, spans such as "Median OS" and "10.7 months" hold significant importance as key pieces of information extracted from medical texts. We gather aligned entities that meet either of the above cases. Then, we create relations with 2 or 3 entities and design corresponding templates, as shown in Table 2. The placeholders in the templates are replaced with the actual entities from the tables. During pretraining, we randomly erase at most one entity in the cloze template. As depicted in Figure 2, we introduce a new special token [SOE] to replace the erased entity, representing the **S**pan **O**f **E**ntity. We denote the modified cloze text as X_C. In practice, each input instance can contain up to 5 cloze templates in X_C.

4.2.2 Attention-based Pointer Network

To learn the Table Cloze Test (TCT), we utilize an attention-based pointer network inspired by the works of attention mechanisms [31] and pointer networks [32]:

$$Q = W_q \cdot q, \tag{4}$$

$$K = W_k \cdot k, \tag{5}$$

$$f(q,k) = \frac{QK^T}{\sqrt{d}}. \tag{6}$$

Here, W_q and W_k are learnable parameters, and d is the dimension of q, k. We compute scores for each token in X using separate functions depending on whether the token corresponds to the start or end of the entity span replaced by [SOE]:

$$u_i^{SOE} = f^{start}(h_{SOE}, h_i), \tag{7}$$

$$v_i^{SOE} = f^{end}(h_{SOE}, h_i), \tag{8}$$

where h_{SOE} represents the representation of [SOE], and f^{start} and f^{end} are functions with their own independent parameters. The probabilities for the start and end positions are computed as follows:

$$\mathbf{p}^{start} = softmax(u_1^{SOE}, ..., u_m^{SOE}), \tag{9}$$

$$\mathbf{p}^{end} = softmax(v_1^{SOE}, ..., v_m^{SOE}), \tag{10}$$

where both $p^{start}, p^{end} \in R^{m\times 1}$. We optimize this task by minimizing the cross-entropy loss between the predicted start and end indices (s, e) and the ground truth start and end indices (s, e) for each [SOE]:

$$L_{TCT} = -\sum_{i\in D}\{\log p(s_i|\mathbf{p}_i^{start}) + \log p(e_i|\mathbf{p}_i^{end})\}, \tag{11}$$

where D is the set of all [SOE] tokens in X_C. In addition, we randomly sample aligned entities that do not have corresponding relations to construct negative cloze text. This further enhances the model's ability to discern relationships. The pointer is uniformly directed to the [CLS] token for these negative samples. We empirically set the proportion of negative samples to 30%.

To summarize, the complete pretraining loss of TabMedBERT is a combination of the losses from the EI and TCT tasks:

$$L = L_{EI} + L_{TCT}. \tag{12}$$

4.3 Revising Paired Entity Recognition

Entities in biomedical articles often appear in pairs, such as drug combinations, endpoints, and results. However, the co-occurrence between these entities is often overlooked by existing models. To address this, we propose a revision method using TabMedBERT to enhance paired entity recognition. For paired entities, if an entity is missed during NER, we revise it using the following steps: (i) After processing the article with the NER model, we convert the predicted part of paired entities into cloze text using a designed template. (ii) The cloze text and the article are inputted similarly to the Table Cloze Test (TCT) task to identify the missing entities. (iii) The revised entities are combined with the previous NER results to generate the final outcomes. For example, in the BC5CDR$_r$ dataset, each sentence usually contains two paired entity types: "Chemical" and "Disease". If the NER model only predicts "Chemical", we form a cloze text for revision like "[Chemical] is associated with [SOE]." The same process applies to vice versa.

Similarly, our approach can also be applied to revise relation extraction. Suppose a RE model fails to predict the relationship between two entities. In that case, we utilize the same method as entity revision to check if we can construct a cloze with one entity and find the other. We revise the RE predictions accordingly if the two entities can be connected. In the practical workflow of biomedical text data annotation, annotators are entrusted with identifying medical entities within paragraphs. To uphold annotation accuracy, the paired entity revision model is utilized to detect any potential gaps in annotations, thereby offering prompts to annotators for necessary corrections. Our experiments demonstrate that TabMedBERT exhibits excellent revision capability in zero-shot and fine-tuning scenarios (§ 5.4.3).

5 Experiments

5.1 Datasets

Pretraining We gather 509,919 PubMed abstracts linked to tables and ensure through PMID that there is no overlap between the training abstracts and the downstream evaluation datasets. On average, each abstract contained 31.5 aligned entities for the Entity Infilling (EI) task and 3.75 relations for the Table Cloze Test (TCT) task. To care for shorter sentence-level tasks, we included two additional relation extraction (RE) datasets, namely BioRel [34] and TBGA [24], in the pretraining process. We utilize gold entities and relations from these datasets to construct inputs similar to the abstract.

BLURB Benchmark consists of five named entity recognition (NER) tasks [20, 9, 10, 7], a PICO (population, intervention, comparison, and outcome) extraction task [26], three relation extraction (RE) tasks [18, 15, 5], a sentence similarity (SS) task [30], a document classification (DC) task [2], and two questions answering (QA) tasks [17, 25], as shown in Table 5. We follow the same fine-tuning method and evaluation metric used by BioLinkBERT.

Information Extraction We conduct experiments on various biomedical NER and RE tasks to verify the effectiveness of our method, including datasets published in the popular BLURB benchmark: On NER, BC5-chem&BC5-disease [20], NCBI-disease [9], BC2GM [10], JNLPBA [7] belong to the BLURB, while BC5CDR [20], ADE [13], CHR [28], BioRED [22] are not. On RE, ChemProt [18], DDI [15], GAD [5] all belong to BLURB, while AIMed [6], BC5CDR [20], ADE [13] are not. We conducted two kinds of RE experiments on BC5CDR and ADE, using gold entities or prediction results of a NER model as input. We make statistics on the above datasets and list the results in Table 3 and 4.

Table 3. Statistics of NER datasets. "Classes" means the number of entity categories in the dataset. The dataset with special mark † belongs to BLURB.

NER	Train	Dev	Test	Classes
BC5-chem†	4560	4581	4797	1
BC5-disease†	4560	4581	4797	1
NCBI-disease†	5424	923	940	1
BC2GM†	12500	2500	5000	1
JNLPBA†	16807	1739	3856	1
BC5CDR	500	500	500	2
ADE	3845	-	427	2
CHR	7298	1182	3614	1
BioRED	400	100	100	6

Table 4. Statistics of RE datasets. "Classes" means the category number of relations in the dataset.

RE	Train	Dev	Test	Classes
ChemProt†	18035	11268	15745	5
DDI†	25296	2496	5716	4
GAD†	4261	535	534	1
AIMed	5251	-	583	1
BC5CDR	500	500	500	1
ADE	3845	-	427	1

Paired Entity Revision We construct a dataset containing paired entities based on the BC5CDR dataset to evaluate our model's revision ability according to the following principles: (1) Divide the dataset by sentence, and filter out all cross-sentence relation pairs; (2) Keep the entity pairs with relations in the sentence and filter out sentences without relations. After that, 1085/1176/1144 sentences in train/dev/test are qualified. We call this dataset BC5CDR$_r$.

5.2 Baselines

We compare different types of pretrained models according to the performance on the BLURB benchmark. We first choose SciBERT [3] and PubMedBERT [11], trained on plain text as baselines. We also involve knowledge-enhanced models: BioLinkBERT [36] and KeBioLM [37]. For generative models, in addition to BioGPT [23] and ChatGPT, we include two additional biomedical models: PMC-LLaMA [10] and BioMedGPT [11].

[10] https://huggingface.co/axiong/PMC_LLaMA_13B
[11] https://huggingface.co/PharMolix/BioMedGPT-LM-7B

5.3 Implementation Details

Pretraining In our pretraining process, we initialize the encoder with the parameters officially released by BioLinkBERT$_{base}$ (110M params) for the base model and randomly initialize the parameters of the pointer network. We use a peak learning rate 5e-5, batch size 512, and train for 648000 steps. We warm up the learning rate in the first 10% steps. The base model was pretrained on 8 NVIDIA RTX A5000 GPUs for half a week. For the large model, we initialize parameters from BioLinkBERT$_{large}$ (340M params), following the same procedure as the base but appropriately increasing the learning rate to 6e-5 and the batch size to 1024. Training took 12 days on 8 NVIDIA RTX A5000 GPUs with automatic mixed precision.

Fine-tuning Our downstream tasks are primarily NER and RE tasks. In our NER model, we couple the encoder portion of the language model with conditional random fields (CRF). We set the learning rate to 3e-5, utilize a batch size of 16, and impose a maximum input sentence length of 1024. The training process runs for a maximum of 100 epochs, with early stopping triggered after 10 epochs. For Relation Extraction, we adopt the method proposed by PURE [38]. The learning rate for the downstream RE model is set to 2e-5. We configure the sentence-level RE task with 10 training epochs, a maximum input sentence length of 256, and a batch size of 32. Regarding document-level RE tasks, training continues for 5 epochs, the maximum input sentence length is set to 1024, and the batch size is 8. In evaluating our models on the BLURB benchmark datasets, we extensively employ the official fine-tuning and evaluation code [12], making only minor adjustments to a limited number of hyperparameters.

Paired Entity Revision We use the same model framework for the TCT task in the revision phase. The model parameters are initialized with our trained TabMedBERT$_{base}$ and fine-tuned with BC5CDR$_r$. We set the learning rate of the revision model to 2e-5, the batch size to 16, and the limit of the input sentence maximum length to 512. The train epochs are set to 15.

5.4 Experimental Results

5.4.1 BLURB Benchmark

Table 5 presents the results obtained on the BLURB dataset. TabMedBERT achieves a new state-of-the-art (SOTA) performance on the BLURB leaderboard, with an impressive absolute value of 85.59%. In particular, TabMedBERT$_{large}$ outperforms the previous SOTA model, BioLinkBERT$_{large}$, in nearly all task categories, exhibiting an average performance improvement of +1.29%. It also surpasses PubmedBERT$_{large}$ by a significant margin. At the base level, our TabMedBERT$_{base}$ also far surpasses other models, reaching 84.14%.

TabMedBERT excels in information extraction tasks, particularly in NER, RE, and PICO, with an average improvement of +1.65%. This is due to the fact that our table-related tasks reinforce the model's understanding of entities and their relationships. Additionally, our model achieves promising results in Question Answering (QA) and Document Classification (DC) tasks, even without specific document-level training like BioLinkBERT$_{large}$. This may be attributed to the Table Cloze Test (TCT) task, which requires the model to identify relevant spans from the overall context, thereby enhancing its performance on document-level tasks.

[12] https://github.com/michiyasunaga/LinkBERT.git

Table 5. Performance on BLURB benchmark. TabMedBERT attains improvement on all tasks, establishing a new state of the art on BLURB. Gains are notably large on information extraction tasks such as NER, PICO, and RE.

	SciBERT	KBioLM	PubMed-BERT$_{base}$	BioLink-BERT$_{base}$	TabMed-BERT$_{base}$	PubMed-BERT$_{large}$	BioLink-BERT$_{large}$	TabMed-BERT$_{large}$
NER								
BC5-chem	92.36	93.17	93.33	93.75	**93.80**	93.23	94.04	**94.67**
BC5-disease	84.16	85.98	85.62	86.10	**86.16**	85.77	**86.39**	86.25
NCBI-disease	86.88	88.40	87.82	88.18	**88.51**	88.25	88.76	**89.73**
BC2GM	82.6	84.19	84.52	**84.90**	84.82	84.72	85.18	**87.54**
JNLPBA	78.67	79.04	80.06	79.03	**82.64**	79.44	80.06	**82.43**
PICO								
EBM PICO	72.90	73.56	73.38	73.97	**74.25**	73.61	74.19	**76.32**
RE								
ChemProt	75.02	77.51	77.24	**77.57**	77.51	78.77	79.98	**80.22**
DDI	82.32	82.56	82.36	82.72	**82.84**	82.39	83.35	**83.83**
GAD	81.76	80.97	82.34	84.39	**85.04**	83.57	84.90	**85.61**
SS								
BIOSSES	89.51	51.33	92.30	**93.25**	93.10	92.73	**93.63**	92.78
DC								
HoC	83.43	83.75	82.32	84.35	**84.97**	82.57	84.87	**85.99**
QA								
PubMedQA	59.60	52.80	55.84	70.20	**64.80**	67.38	**72.18**	71.70
BioASQ	80.71	73.57	87.56	91.43	**95.42**	93.36	94.82	**95.71**
BLURB score	80.76	77.48	81.10	83.39	**84.14**	82.86	84.30	**85.59**

Table 6. Comparison of results on NER. BLURB-NER represents the average score of NER tasks in BLURB. AVG is the mean F1 on all NER datasets.

NER	ChatGPT	SciBERT	KeBioLM	PubMed-BERT$_{base}$	BioLink-BERT$_{base}$	TabMed-BERT$_{base}$	BioLink-BERT$_{large}$	TabMed-BERT$_{large}$
BLURB-NER	-	84.93	86.24	86.27	86.39	**87.18**	86.88	**88.12**
BC5CDR	54.80	88.37	90.07	89.37	89.56	**90.18**	90.06	**90.42**
ADE	71.79	89.94	90.33	90.22	90.12	**90.80**	90.83	**91.38**
BioRED	64.51	88.47	90.15	90.74	90.87	**91.13**	89.13	**91.81**
AVG	63.70	86.42	87.77	87.27	87.50	**88.50**	88.05	**89.27**

5.4.2 *Information Extraction*

Tables 6 and 7 provide a comprehensive overview of the performance of various models on information extraction tasks. Our TabMedBERT model consistently outperforms other models across different model sizes, demonstrating the significant advantage of our training approach in information extraction. When comparing our base model to its initialized counterpart, BioLinkBERT$_{base}$, we observe improvements of +1.0% and +1.01% in averaged NER and RE, respectively. Similarly, for TabMedBERT$_{large}$ compared to BioLinkBERT$_{large}$, these improvements are +1.22% and +0.7%. These results underscore the effectiveness of our table-based training tasks in boosting model performance.

As shown in Table 9, generative models demonstrate subpar performance in information extraction tasks and require significantly more parameters to achieve comparable results. This highlights the ongoing preference for understanding-based models for such tasks. The results show that generative models perform much worse than discriminative models in IE tasks and require significantly more space and time. However, generative models excel in question-answering (QA) tasks and are effective in document classification without needing complex prompts. In summary, TabMedBERT should be the preferred choice for IE tasks.

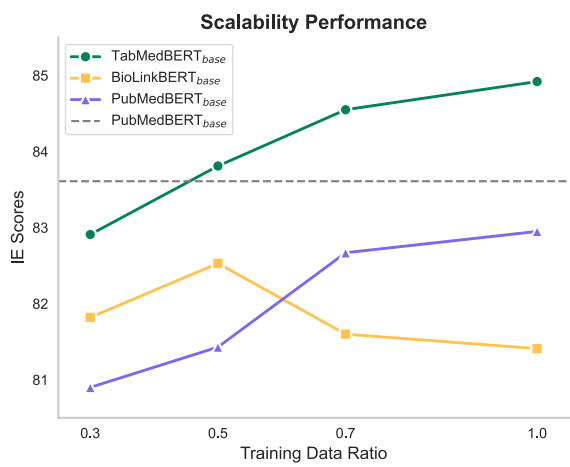

Figure 3. The scalability performance of the model. The IE score represents the performance of IE tasks. Squares and triangles represent the use of plain text without introducing table-related tasks.

5.4.3 *Paired Entity Revision*

The findings are detailed in Table 8. Our model showcases notable enhancements across all Relation Extraction (RE) models, even in

Table 7. Comparison of experimental results on RE. BLURB-RE represents the average score of RE tasks in BLURB. AVG is the mean F1 on all RE datasets. Results of BC5CDR and ADE mean to predict relations on entities predicted by the NER model in Table 6.

RE	ChatGPT	SciBERT	KeBioLM	PubMed-BERT$_{\texttt{base}}$	BioLink-BERT$_{\texttt{base}}$	TabMed-BERT$_{\texttt{base}}$	BioLink-BERT$_{\texttt{large}}$	TabMed-BERT$_{\texttt{large}}$
BLURB-RE	-	79.70	80.34	80.64	81.56	**81.80**	82.74	**83.22**
AIMed	-	88.51	87.50	88.41	85.96	**89.66**	87.89	**89.87**
BC5CDR	32.01	57.24	61.35	60.55	61.24	**62.30**	62.42	**62.56**
ADE	59.90	81.52	83.37	81.62	82.30	**83.60**	83.47	**84.08**
AVG	45.95	77.72	78.87	78.75	79.03	**80.16**	80.33	**81.02**

Table 8. Comparison of results before and after revision of our model on BC5CDR$_r$. F1 measures the results here.

	Revision Type	SciBERT	BioLinkBERT	KeBioLM	PubMedBERT	TabMedBERT
NER	**before**	89.26	90.72	90.85	91.19	91.51
	zero-shot	88.90	90.41	90.68	90.96	91.18
	fine-tuned	**89.82**	**90.88**	**91.16**	**91.48**	**91.79**
RE	**before**	73.65	74.24	77.10	76.78	77.86
	zero-shot	78.87	79.65	81.38	81.86	82.21
	fine-tuned	**80.04**	**81.37**	**82.30**	**83.20**	**83.71**

Table 9. Comparison of large generative LMs in effectiveness and efficiency. † indicates fine-tuning, others are zero-shot.

	BC5CDR(RE)	ADE(NER)	HoC(DC)	PubMedQA	Params	Time Cost
TabMedBERT†	**62.5**	**91.3**	**86.0**	71.7	**340M**	1×
BioGPT†	44.9	83.1	85.1	**78.2**	355M	$\geq$ 4×
PMC-LLaMA	51.3	79.5	78.4	77.9	13B	$\geq$ 10×
BioMedGPT	49.2	68.0	75.7	74.9	7B	$\geq$ 10×
ChatGPT	32.0	71.8	73.1	63.9	175B	>>>

scenarios where zero-shot learning is applied. TabMedBERT exhibits its effectiveness in refining the outcomes of Named Entity Recognition (NER) and RE tasks through fine-tuning. Especially in the RE task, there is a substantial average improvement of over 6% observed across five models. This improvement can be attributed to the complexities involved in error propagation within the RE task. On one hand, numerous entity pairs within sentences are inadequately identified. Conversely, the relationships between entity pairs are often inaccurately predicted. Our approach, employing the table cloze test task, adeptly tackles these challenges by refining predictions that involve missing paired entities. This simultaneous refinement significantly contributes to the observed performance improvements.

5.5 Scalability Study

As portrayed in Figure 3, the proficiency of our TabMedBERT model shows a consistent upward trend as we augment the volume of training data. Conversely, there is a noticeable decline in its performance when it relies solely on textual data, excluding tasks related to tables. This discrepancy underscores a crucial aspect: the effectiveness of our approach lies not solely in the richness of the dataset but rather in the intricacies of our training methodology. It is noteworthy that as training progresses, the performance of LinkBERT exhibits a gradual deterioration. This decline can predominantly be attributed to the disruption in knowledge assimilation arising from the absence of document linking, a vital component of training with textual data.

5.6 Ablation Study

In Table 10, substituting EI with a random masked MLM leads to a decrease of 2.0% in the mean F1 score for the NER task and 1.73% for the RE task. Significantly, the absence of the Entity Infilling (EI) task notably affects the NER tasks more than other aspects. This is chiefly due to the EI task's contribution of a substantial amount of biomedical entity information to the model. Additionally, excluding the cloze task diminishes the model's capacity to grasp relationship information between entities, leading to a noteworthy 2.4% decrease in performance on downstream RE tasks. These observations highlight the critical significance of both tasks for the model's overall efficacy.

Table 10. Ablation experiment of TabMedBERT on NER and RE tasks. "w/o EI" replaces biomedical entities masking with random masking. "w/o TCT" refers to removing the table cloze task.

	TabMedBERT	**w/o EI**	**w/o TCT**
NER-AVG	**88.50**	86.51	87.23
RE-AVG	**80.16**	78.43	77.75

6 Conclusion

This paper presents TabMedBERT, an innovative biomedical pretrained language model enriched with tabular knowledge. TabMedBERT incorporates two novel self-supervised tasks related to tables, namely Entity Infilling (EI) and Table Cloze Test (TCT), enhancing its understanding of structured data. Our evaluation across various biomedical tasks demonstrates TabMedBERT's state-of-the-art performance compared to other language models. Particularly noteworthy is the model's architecture for TCT, offering a straightforward solution for refining information extraction involving paired entities. Looking ahead, we envision several avenues for advancing our research. Firstly, we intend to explore generative approaches to harness tabular knowledge more effectively. Additionally, we are actively pursuing the extension of this model's applicability to multilingual scenarios.

Acknowledgements

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62106165) and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. Publicly Available Clinical BERT Embeddings, June 2019. URL http://arxiv.org/abs/1904.03323. arXiv:1904.03323 [cs].

[2] S. Baker, I. Silins, Y. Guo, I. Ali, J. Högberg, U. Stenius, and A. Korhonen. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440, Feb. 2016. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btv585. URL https://academic.oup.com/bioinformatics/article/32/3/432/1743783.

[3] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[4] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A Pretrained Language Model for Scientific Text, Sept. 2019. URL http://arxiv.org/abs/1903.10676. arXiv:1903.10676 [cs].

[5] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):1–17, 2015.

[6] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial intelligence in medicine*, 33(2):139–155, 2005.

[7] N. Collier and J.-D. Kim. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, 2004.

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

[9] R. I. Doğan, R. Leaman, and Z. Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

[10] J. Du, Q. Chen, Y. Peng, Y. Xiang, C. Tao, and Z. Lu. Ml-net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11):1279–1285, 2019.

[11] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

[12] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, Jan. 2022. ISSN 2691-1957, 2637-8051. doi: 10.1145/3458754. URL http://arxiv.org/abs/2007.15779. arXiv:2007.15779 [cs].

[13] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012.

[14] B. He, D. Zhou, J. Xiao, Q. Liu, N. J. Yuan, T. Xu, et al. Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147*, 2019.

[15] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, and T. Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.

[16] K. Huang, J. Altosaar, and R. Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, Nov. 2020. URL http://arxiv.org/abs/1904.05342. arXiv:1904.05342 [cs].

[17] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu. PubMedQA: A Dataset for Biomedical Research Question Answering. 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 2567–2577, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://www.aclweb.org/anthology/D19-1259.

[18] M. Krallinger, O. Rabal, S. A. Akhondi, M. P. Pérez, J. Santamaría, G. P. Rodríguez, G. Tsatsaronis, A. Intxaurrondo, J. A. López, U. Nandal, et al. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146, 2017.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682, Sept. 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btz682. URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz682/5566506.

[20] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

[21] C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bionlp-1.21. URL https://aclanthology.org/2021.bionlp-1.21.

[22] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, and Z. Lu. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23 (5):bbac282, 2022.

[23] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, 23(6):bbac409, Nov. 2022. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbac409. URL http://arxiv.org/abs/2210.10341. arXiv:2210.10341 [cs].

[24] S. Marchesin and G. Silvello. Tbga: a large-scale gene-disease association dataset for biomedical relation extraction. *BMC bioinformatics*, 23 (1):1–16, 2022.

[25] A. Nentidis, K. Bougiatiotis, A. Krithara, and G. Paliouras. Results of the seventh edition of the BioASQ Challenge. volume 1168, pages 553–568. 2020. doi: 10.1007/978-3-030-43887-6_51. URL http://arxiv.org/abs/2006.09174. arXiv:2006.09174 [cs].

[26] B. Nye, J. J. Li, R. Patel, Y. Yang, I. J. Marshall, A. Nenkova, and B. C. Wallace. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature, June 2018. URL http://arxiv.org/abs/1806.04185. arXiv:1806.04185 [cs].

[27] M. E. Peters, M. Neumann, R. L. Logan IV, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*, 2019.

[28] S. K. Sahu, F. Christopoulou, M. Miwa, and S. Ananiadou. Inter-sentence relation extraction with document-level graph convolutional neural network. *arXiv preprint arXiv:1906.04684*, 2019.

[29] B. Smock, R. Pesala, and R. Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. pages 4634–4642, June 2022.

[30] G. Soğancıoğlu, H. Öztürk, and A. Özgür. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, July 2017. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btx238. URL https://academic.oup.com/bioinformatics/article/33/14/i49/3953954.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[32] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer Networks, Jan. 2017. URL http://arxiv.org/abs/1506.03134. arXiv:1506.03134 [cs, stat].

[33] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

[34] R. Xing, J. Luo, and T. Song. Biorel: towards large-scale biomedical relation extraction. *BMC bioinformatics*, 21(16):1–13, 2020.

[35] W. Xiong, J. Du, W. Y. Wang, and V. Stoyanov. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *arXiv preprint arXiv:1912.09637*, 2019.

[36] M. Yasunaga, J. Leskovec, and P. Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022.

[37] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang. Improving biomedical pretrained language models with knowledge. *arXiv preprint arXiv:2104.10344*, 2021.

[38] Z. Zhong and D. Chen. A frustratingly easy approach for entity and relation extraction. In *North American Association for Computational Linguistics (NAACL)*, 2021.