Expert Systems With Applications 235 (2024) 121054

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Supplementing domain knowledge to BERT with semi-structured information of documents☆

Jing Chen, Zhihua Wei *, Jiaqi Wang, Rui Wang, Chuanyang Gong, Hongyun Zhang, Duoqian Miao

School of Computer Science and Technology, Tongji University, Shanghai, China

ARTICLE INFO

Keywords:
BERT
Domain adaptation
Semi-structured information
Biomedical question answering
Pre-trained language model

ABSTRACT

Domain adaptation is a good way to boost BERT's performance on domain-specific natural language processing (NLP) tasks. Common domain adaptation methods, however, can be deficient in capturing domain knowledge. Meanwhile, the context fragmentation inherent in Transformer-based models also hinders the acquisition of domain knowledge. Considering the semi-structural characteristics of documents and its potential for alleviating these problems, we leverage the semi-structured information of documents to supplement domain knowledge to BERT. To this end, we propose a topic-based domain adaptation method, which enhances the capture of domain knowledge at various levels of text granularity. Specifically, topic masked language modeling is designed at the paragraph level for pre-training; topic subsection matching degree dataset is automatically constructed at the subsection level for intermediate fine-tuning. Experiments are conducted over four biomedical NLP tasks across six datasets. The results show that our method benefits BERT, RoBERTa, SpanBERT, BioBERT, and PubMedBERT in nearly all cases. And we see significant gains in two question answering (QA) tasks, especially customer health QA, the topic-related one, with an average accuracy improvement of 4.8%. Thus, the semi-structured information of documents can be exploited to make BERT capture domain knowledge more effectively.

1. Introduction

Transformer-based pre-trained language models (TPLM) obtain excellent results in natural language understanding (NLU) tasks (Devlin, Chang, Lee, & Toutanova, 2019; Joshi, Chen, Liu, Weld, Zettlemoyer, & Levy, 2020; Lan, Chen, Goodman, Gimpel, Sharma, & Soricut, 2020; Liu et al., 2019), among them the most representative one is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). The main feature of BERT is the "Pre-train + Fine-tune" paradigm. BERT conducts self-supervised learning on massive general domain text, to learn universal language representations (i.e., pre-training), then re-trains on small scale, labeled data to quickly adapt to target tasks (i.e., fine-tuning). Research shows BERT captures rich syntactic, semantic, and world knowledge during pre-training, and transfers these knowledge to specific tasks via fine-tuning (Rogers, Kovaleva, & Rumshisky, 2020). This knowledge-transfer fails when it comes to specific domains, biomedical domain for example (Lee et al., 2020). To close the gap between specific domain and general domain, several domain-customized BERT models are developed (Chalkidis, Fergadiotis, Malakasiotis, Aletras, & Androutsopoulos, 2020; Gu et al., 2021; Lee et al., 2020; Yang, Uy, & Huang, 2020). Researchers use in-domain text, either to further pre-train general-domain pre-trained models, or to pre-train language models from scratch. And new state-of-the-art results are observed in many domain-specific NLU tasks. However, such domain adaptation methods can be deficient in capturing the domain knowledge focused by domain experts (Kalyan, Rajasekharan, & Sangeetha, 2021).

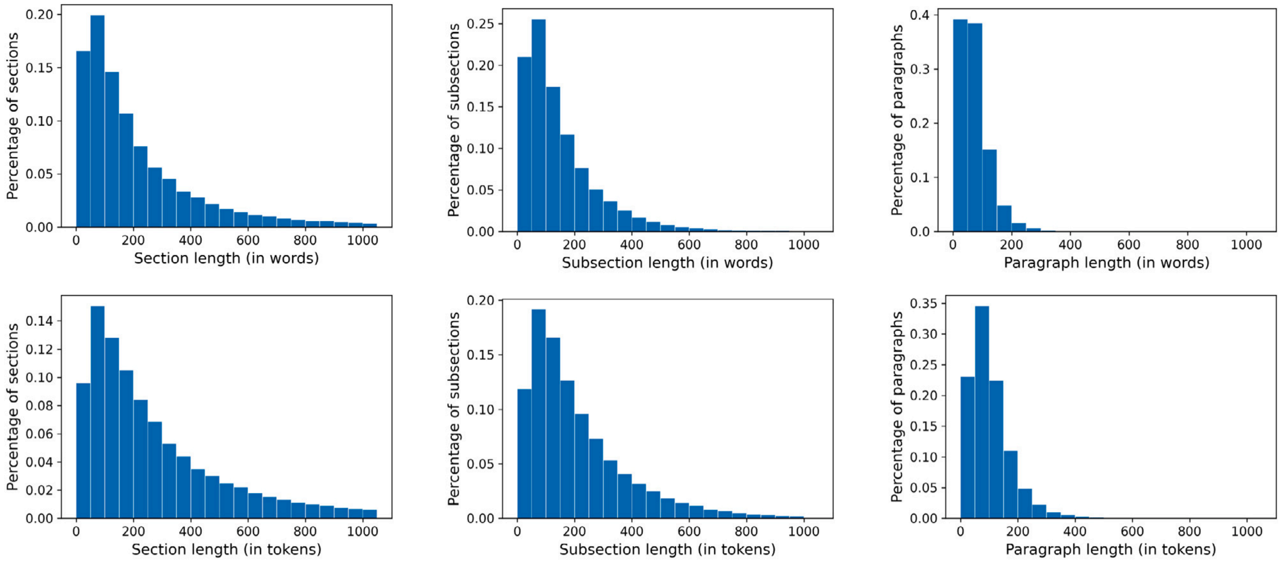
Domain knowledge is knowledge of a specific, specialized discipline or field, in contrast to general (or domain-independent) knowledge (Hjørland & Albrechtsen, 1995). For example, in clinical medicine, domain knowledge involves clinicians may diagnose, treat, and otherwise care for patients. So how do humans learn domain knowledge? We find the semi-structured information of documents, including heading and hierarchy, plays a significant role in this learning process. Hierarchy is the order in which the ideological content of an article is expressed, reflects the development stages of objective things or all aspects of contradictions. Hierarchy typically consists of several natural

☆ This document is the results of the research project funded by National Natural Science Foundation of China.

* Corresponding author.

E-mail addresses: chenjing_miss@tongji.edu.cn (J. Chen), zhihua_wei@tongji.edu.cn (Z. Wei), wangjq@tongji.edu.cn (J. Wang), rwang@tongji.edu.cn (R. Wang), gongchuanyang@tongji.edu.cn (C. Gong), zhanghongyun@tongji.edu.cn (H. Zhang), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.eswa.2023.121054
Received 3 January 2023; Received in revised form 25 July 2023; Accepted 25 July 2023
Available online 3 August 2023



(a) Sections

(b) Subsections

(c) Paragraphs

Fig. 1. Length distributions of sections, subsections and paragraphs. Note the subsections here are not natural ones, but "functional subsections" (for details, see Section 3.2.1). It is obvious the lengths vary widely between them, almost 40% of sections exceed 200 words, whereas the proportion is 25% for subsection, and only 3% for paragraphs.

paragraphs, and can be subdivided into section, subsection, with paragraph being the smallest. Heading is a brief statement that identifies the central argument of a specific content, which divides into the first-level heading (h1, i.e., the article title), the second-level heading (h2, i.e., the section title), the third-level heading (h3, i.e., the subsection title), etc. For example, in the Wikipedia article "COVID-19", *cause*, *diagnosis* and *treatment* in the table of contents are second-level headings and section divisions. Combining heading and hierarchy, we could learn various aspects of the disease, i.e., the domain knowledge interests clinicians and patients.

Regrettably, the semi-structured information has been seriously neglected in BERT's domain adaptation: (1) The pre-training data from Wikipedia only retains unstructured text passages, whereas headers and hierarchy information of the article are discarded (Devlin et al., 2019); (2) The training sequence forms either by sampling and concatenating two segments of text (Devlin et al., 2019), or by packing full sentences sampled contiguously from one or more documents (Joshi et al., 2020; Liu et al., 2019). In addition, BERT requires a fixed-length input sequence of up to 512 tokens, then the overlong ones are truncated without respecting the sentence or any other semantic boundary. Hence, the model lacks the necessary contextual information to predict the first few symbols of input sequence, leading to inefficient optimization and inferior performance (i.e., context fragmentation) (Dai, Yang, Yang, Carbonell, Le, & Salakhutdinov, 2020). Clearly, the negligence of the semi-structured information and the context fragmentation problem all hinder BERT to capture domain knowledge.

Some researchers try to resolve the two problems. Dai et al. (2020) propose a novel Transformer architecture – Transformer-XL, which addresses the context fragmentation problem with a segment-level recurrence mechanism. But until now, the Transformer-XL based pre-trained language model (PLM) is rare. He, Zhu, Zhang, Chen, and Caverlee (2020) introduce a disease knowledge infusion training procedure (diseaseBERT). It takes question-answer pairs as training sequences, where the question is constructed with a disease name (the article title) and an aspect name of the disease (the section title), and the answer is the whole section's content. Then pre-train BERT using a modified masked language modeling (MLM) (Devlin et al., 2019), which merely masks the title words. To our knowledge, DiseaseBERT is the first work to explicitly use the semi-structured information for BERT's domain adaptation, but there are still deficiencies: (1) Its usage of the semi-structured information is limited, the questions only cover two kinds of titles; (2) The answers are generally long, and simple chunking leads to contextual fragmentation. Fig. 1(a) and (c) illustrate this more intuitively, the lengths of almost 40% of sections exceed 200 words, whereas the proportion is only 3% for paragraphs, so taking paragraphs as answers are more helpful for alleviating contextual fragmentation.

In this paper, we propose topic-based domain adaptation (TDA), to enable BERT to better capture domain knowledge with the semi-structured information of documents. TDA emphasizes the intrinsic relation among heading, hierarchy, and domain knowledge, and allows BERT to capture the domain knowledge at various levels of text granularity. Specifically, at the paragraph level, we create topic-paragraph pairs as training sequences, where a paragraph's topic is derived by concatenating the headings of each level hierarchy the paragraph belongs to, i.e., h1, h2, h3, etc., with separators, for details, see Section 3.1.2. Then topic masked language modeling (TMLM) is designed, to selectively mask some heading elements in the topic part. It forces BERT to learn the semantic relationship between a paragraph and its topic, and thereby capture the domain knowledge embedded in paragraphs. At the subsection level, the paragraphs under the same topic are merged into a functional subsection, then topic-subsection pairs are available. On this basis, topic subsection matching degree (TSMD) dataset is automatically constructed, which is used for intermediate fine-tuning, to help target task via transfer learning. The overall framework of TDA is shown in Fig. 2.

The biomedical domain is taken as a case study to illustrate our TDA method, and it is evaluated on four disease-related tasks across six datasets. Experiments show that TDA can benefit BERT in various domain-specific tasks, especially customer health QA (CHQA). More than that, TDA can be easily drawn on to other domains. The data and code are available at Code Ocean.[1]

The rest of the paper is structured as follows: The related work is discussed in Section 2; The TDA method composed of TMLM and TSMD is detailly presented in Section 3; The datasets, baselines, implementation, as well as results and discussion are reported in Section 4; Finally, the conclusion and future work of this study are given in Section 5.

[1] https://codeocean.com/capsule/1721209/tree

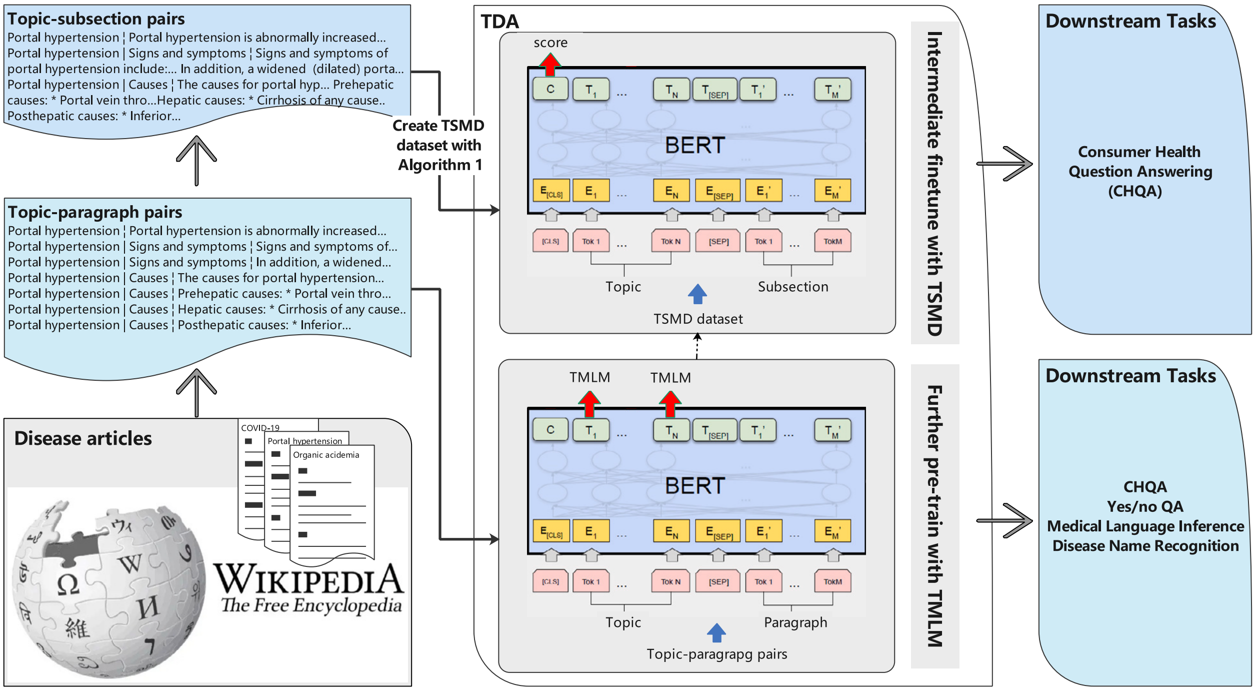

Topic-subsection pairs
Portal hypertension ¦ Portal hypertension is abnormally increased...
Portal hypertension | Signs and symptoms ¦ Signs and symptoms of portal hypertension include:... In addition, a widened (dilated) porta...
Portal hypertension | Causes ¦ The causes for portal hyp... Prehepatic causes: * Portal vein thro...Hepatic causes: * Cirrhosis of any cause.. Posthepatic causes: * Inferior...
Create TSMD dataset with Algorithm 1
Topic-paragraph pairs
Portal hypertension ¦ Portal hypertension is abnormally increased...
Portal hypertension | Signs and symptoms ¦ Signs and symptoms of...
Portal hypertension | Signs and symptoms ¦ In addition, a widened...
Portal hypertension | Causes ¦ The causes for portal hypertension...
Portal hypertension | Causes ¦ Prehepatic causes: * Portal vein thro...
Portal hypertension | Causes ¦ Hepatic causes: * Cirrhosis of any cause..
Portal hypertension | Causes ¦ Posthepatic causes: * Inferior...
Disease articles
COVID-19
Portal hypertension
Organic acidemia
WIKIPEDIA
The Free Encyclopedia
TDA
score
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
TokM
Topic
Subsection
TSMD dataset
Intermediate finetune with TSMD
TMLM
TMLM
Topic
Paragraph
Topic-paragrapg pairs
Further pre-train with TMLM
Downstream Tasks
Consumer Health Question Answering (CHQA)
Downstream Tasks
CHQA
Yes/no QA
Medical Language Inference
Disease Name Recognition

Fig. 2. The framework of TDA. TMLM: The Wikipedia disease-related articles are collected as in-domain text corpus. Then topic-paragraph pairs are created with the paragraph level semi-structured information. We further pre-train a general-domain BERT model with TMLM on the topic-paragraph pairs, before fine-tune it on various domain-specific tasks. TSMD: Topic-subsection pairs are transformed from Topic-paragraph pairs. On this basis, we create TSMD dataset with Algorithm 1. After intermediate fine-tuning BERT with TSMD, we can fine-tune it on the target CHQA task.

2. Related work

2.1. Domain knowledge enhanced PLM

The current TPLMs mainly train on massive unstructured data from Internet, the lack of industry related domain knowledge leads to their poor performance in domain-specific NLP tasks. The application of CBLUE (Zhang et al., 2022) shows the general TPLM is less effective than human in handling biomedical domain tasks. Thus, incorporating domain knowledge into TPLM is a research hotspot. The mainstream domain knowledge enhanced TPLM divides into two categories.

One category annotates the domain knowledge contained in text with weak supervision, and designs knowledge-driven pre-training tasks. Considering the features of in-domain text corpus, the general domain vocabulary can be extended with in-domain vocabulary (Poerner, Waltinger, & Schütze, 2020; Tai, Kung, Dong, Comiter, & Kuo, 2020; Yao, Huang, Wang, Dong, & Wei, 2021; Zhang et al., 2020), which allows PLMs to learn prior domain knowledge during pre-training and fine-tuning. Considering the features of downstream tasks, Gururangan et al. (2020) present task-adaptive pre-training – it involves further pre-training on task-related unlabeled instances; Gu, Zhang, Wang, Liu, and Sun (2020) propose a selective masking strategy, which enables language model to learn task-specific patterns during pre-training; Zhang et al. (2020) formulate synthetic tasks with the inherent structure in unlabeled data for intermediate fine-tuning. Our TMLM falls into this category, as a variant of MLM, it tasks the semi-structured information as weak supervision signal, and emphasizes its importance for domain knowledge learning.

Another category conducts joint pre-training on in-domain structured knowledge base and unstructured text. For the heterogeneous embedding space problem, K-BERT (Liu et al., 2020) first expands the original text into a tree structure using triples of knowledge graph (KG), and then compresses it back into a text sequence with soft-position and visible matrix; whereas, QA-GNN (Yasunaga, Ren, Bosselut, Liang, & Leskovec, 2021) connects the QA context and KG to form a joint graph, and mutually update their representations through graph neural networks; BERT-MK (He, Zhou, et al., 2020) designs a vanilla Transformer encoder based knowledge fusion module – K-Encoder, to extract entity knowledge and fuse heterogeneous information. To avoid knowledge forgetting, DAKI (Lu, Dou, & Nguyen, 2021) and MoP (Meng, Liu, Clark, Shareghi, & Collier, 2021) integrate domain knowledge via lightweight adapters, the former independently trained adapters for different sources of domain knowledge, and the latter partitioned a big KG into smaller sub-graphs and train their respective adapters.

2.2. Clever use of the semi-structured information

The concept of semi-structured information is rarely seen in the current research, but many datasets did use this concept knowingly or unknowingly, especially the cloze-style QA datasets. They were created with the semi-structured information in news articles (Hermann et al., 2015) or in books (Bajgar, Kadlec, & Kleindienst, 2016; Hill, Bordes, Chopra, & Weston, 2016) or in scientific literature (Kim et al., 2018; Pappas, Androutsopoulos, & Papageorgiou, 2018; Pappas, Stavropoulos, Androutsopoulos, & McDonald, 2020). These datasets are generally large in scale (ranging from 100 K to 16.4 M instances) and thus can be used for pre-training or as a task itself (Jin et al., 2022). However, insufficient use of the semi-structured information make them noisy. Besides, the non-cloze-style QA datasets, PubMedQA (Jin, Dhingra, Liu, Cohen, & Lu, 2019) and MedQuAD (Ben Abacha & Demner-Fushman, 2019), use the semi-structured information more accurately, which allows their data quality to be greatly enhanced. As a consumer health QA-like dataset, our TSMD is automatically created by emphasizing semantic integrity of the semi-structured information, and greatly benefit specific downstream task through intermediate fine-tuning.

In addition to the datasets, there are works incorporate the semi-structured information into pre-training corpus. HTLM (Aghajanyan et al., 2021), a hyper-text language model trained on a large-scale web crawl, designs prompts that incorporate the established semantics of HTML to better control for the desired model output. LinkBERT (Yasunaga, Leskovec, & Liang, 2022) creates the pre-training inputs by

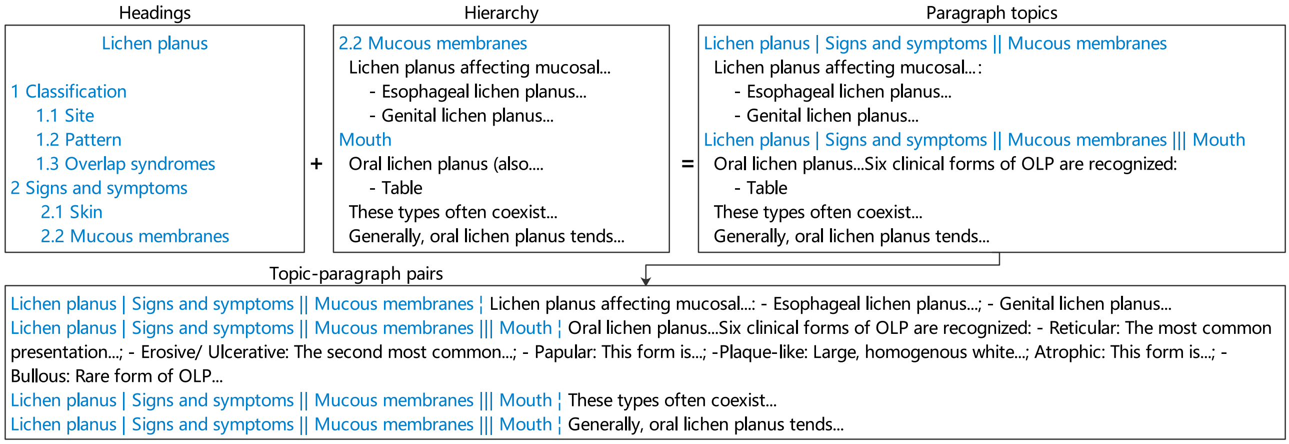

Headings
Lichen planus
1 Classification
1.1 Site
1.2 Pattern
1.3 Overlap syndromes
2 Signs and symptoms
2.1 Skin
2.2 Mucous membranes
+
Hierarchy
2.2 Mucous membranes
Lichen planus affecting mucosal...
- Esophageal lichen planus...
- Genital lichen planus...
Mouth
Oral lichen planus (also....
- Table
These types often coexist...
Generally, oral lichen planus tends...
=
Paragraph topics
Lichen planus | Signs and symptoms || Mucous membranes
Lichen planus affecting mucosal...:
- Esophageal lichen planus...
- Genital lichen planus...
Lichen planus | Signs and symptoms || Mucous membranes ||| Mouth
Oral lichen planus...Six clinical forms of OLP are recognized:
- Table
These types often coexist...
Generally, oral lichen planus tends...
Topic-paragraph pairs
Lichen planus | Signs and symptoms || Mucous membranes ¦ Lichen planus affecting mucosal...: - Esophageal lichen planus...; - Genital lichen planus...
Lichen planus | Signs and symptoms || Mucous membranes ||| Mouth ¦ Oral lichen planus...Six clinical forms of OLP are recognized: - Reticular: The most common presentation...; - Erosive/ Ulcerative: The second most common...; - Papular: This form is...; -Plaque-like: Large, homogenous white...; Atrophic: This form is...; - Bullous: Rare form of OLP...
Lichen planus | Signs and symptoms || Mucous membranes ||| Mouth ¦ These types often coexist...
Lichen planus | Signs and symptoms || Mucous membranes ||| Mouth ¦ Generally, oral lichen planus tends...

Fig. 3. Generation of topic-paragraph pairs. The focal step of the whole process is getting the topic of a paragraph. Considering the hierarchy of an article, a paragraph topic is formed by concatenating the headings of each level hierarchy the paragraph belongs to, with separators. Then pair it with the paragraph, a topic-paragraph pair is generated.

placing linked documents in the same context, to capture dependencies or knowledge that span across documents. HKLM (Zhu, Peng, Lyu, Hou, Li, & Xiao, 2023), a unified PLM for all forms of text, including unstructured text, semi-structured text, and well-structured text, it models the semi-structured text by proposing title matching training, which classify whether the title matches the paragraph. Inspired by this, our TDA, a novel domain adaptation framework, aims to make BERT capture more domain knowledge with better use of the semi-structured information. TMLM and TSMD are the two key technologies, which enable BERT to capture the domain knowledge embedded in paragraph and subsection respectively during multiple training phases.

3. Methods

In this section, topic-based domain adaptation (TDA) is presented detailly. We firstly introduce the pre-training task — topic masked language modeling (TMLM), that enables BERT to capture the domain knowledge contained in paragraph. And then describe the way of automatically building the dataset — topic subsection matching degree (TSMD), with the subsection level semi-structured information.

3.1. Topic masked language modeling

TMLM is the paragraph level TDA method, which consists of three main steps: (1) build in-domain text corpus; (2) construct topic-paragraph pairs; (3) propose a new language modeling task — TMLM. After pre-training BERT with TMLM on the topic-paragraph pairs we created, the domain knowledge contained in paragraphs is encoded. Next, we will discuss each step in more detail.

3.1.1. In-domain text corpus

Following He, Zhu, et al. (2020), we verify the effectiveness of TDA in the biomedical domain, and the disease-related articles from English Wikipedia are used as in-domain text source. To get as many articles as possible, we collect disease terms from two main branches of the Medical Subject Headings (MeSH) tree,[2] i.e. Diseases [C] and Mental Disorders [F03]. In addition, the Wikipedia page “Category: Lists of diseases”[3] serves as a supplement source of disease terms. After eliminating those duplicate or empty entries, 4930 disease-themed English Wikipedia articles are obtained.

To construct an in-domain text corpus that incorporates the semi-structured information of document, we retain the heading and hierarchy of articles in web crawler phase. During data cleaning, the texts irrelevant to the article topic, and those images, complicated tables, special characters that are hard to process for BERT are filtered out. After data pre-processing, the obtained in-domain text corpus is further organized into topic-paragraph pairs — the pre-training corpus with paragraph level semi-structured information.

Table 1
Statistics of the topic-paragraph pairs.

Number of articles	4930
Number of sections	30,432
Average length of sections (in words)	250.14
Average length of sections (in tokens)	359.14
Average length of section topics (in words)	3.26
Average length of section topics (in tokens)	6.79
Number of paragraphs	104,696
Average length of paragraphs (in words)	72.47
Average length of paragraphs (in tokens)	104.39
Average length of paragraph topics (in words)	4.27
Average length of paragraph topics (in tokens)	7.94

3.1.2. Topic-paragraph pairs

As discussed in Section 1, paragraphs are better candidates as answers compared to sections in terms of length, and thus we focus on the domain knowledge contained in paragraph. Generally speaking, paragraph title along with paragraph itself can depict paragraph level domain knowledge, but the following defects exist: (a) many paragraphs do not have a title; (b) the title of a paragraph alone cannot fully summarize its topic. Considering the hierarchy of an article, we concatenate the headings of each level hierarchy a paragraph belongs to, i.e., h1, h2, h3, etc., with separators to form the paragraph topic, then pair it with the paragraph, a topic-paragraph pair is generated. The whole process is shown in Fig. 3. Also worth noting is that when a table or a list appears as supplementary content of one paragraph, we do the following: (1) convert the table to a list, (2) convert the list to plain text, then concatenate it to the paragraph.

The statistics of the pre-training corpus are shown in Table 1. For fair comparison, we also get topic-section pairs by following the similar process depicted in Fig. 3, and obtain their statistics. As you can see, there are more heading elements in paragraph topics than that of section topics, and the average length of paragraphs are evidently shorter than that of sections, with less than one third of its length. All of which demonstrate the superiority of topic-paragraph pairs, whether in the high usage of the semi-structured information or the potential to reduce context fragmentation.

[2] https://meshb.nlm.nih.gov/treeView
[3] https://en.wikipedia.ahmu.cf/wiki/Category:Lists_of_diseases

[MASK] [MASK] ¦ Lichen planus (LP) is a chronic inflammatory and immune-mediated disease that affects the...
[MASK] [MASK] | Signs and symptoms || [MASK] [MASK] ||| Mouth ¦ These types often coexist in the same...
Lichen planus | [MASK] [MASK] [MASK] || Mucous membranes ||| [MASK] ¦ Generally, oral lichen planus tends...

Fig. 4. Examples of the topic masking strategy adopted by TMLM. For one training instance, odd term and even term heading elements in the topic part is masked with equal probability, and the words in the masked heading element are replaced by '[MASK]'.

3.1.3. Topic masking strategy

To make BERT capture the domain knowledge contained in topic-paragraph pairs, we propose a topic masking strategy for TMLM, which selectively masks some heading tokens in the topic part by an average 50% masking rate. Specifically, if there is only one heading element in the topic part, mask it. If the number of heading elements in the topic part exceeds 1, mask the odd and even heading elements with equal probability. Thereby the topic part can serve as a cloze-style question, and the paragraph is the target answer. The final training instances are shown in Fig. 4.

BERT predicts the masked heading elements in the same way as BERT predicts the randomly masked words in input sequences, i.e., the MLM task. And MLM for BERT just like human solves cloze test questions. We can guess the missing words from context, and BERT adopts a multi-layer, bidirectional Transformer encoder architecture, and the word representations obtained by MLM are jointly conditioned on both left and right context. So when BERT predicts the masked heading elements, it has to learn the semantic relationship between the paragraph and its topic. In this way, BERT captures the domain knowledge embedded in paragraphs during pre-training.

3.2. Topic subsection matching degree

TSMD is the subsection level TDA method, the dataset creation follows three steps: (1) provide theoretical basis; (2) construct topic-subsection pairs; (3) generate MEDIQA-like QA instances. Before fine-tuning BERT on MEDIQA-2019, we use TSMD for intermediate fine-tuning, to benefit target task via transfer learning. Next, we will flesh out these steps.

3.2.1. Theoretical basis

Gu et al. (2020) note that because of the high cost and long time-consuming of tagging data, insufficient supervised data is frequently a matter during BERT fine-tuning. This issue becomes particularly prominent in the specific domain, for data annotation here requires the intervention of domain expert. Thus, BERT shows poor performance in domain-specific tasks. However, intermediate fine-tuning on large, related dataset allows BERT to learn more domain-specific and task-specific patterns, which improves BERT performance on domain-specific target tasks (Kalyan et al., 2021).

The questions of CHQA typically raise by the general public on search engines, range from self-diagnosis to finding medications, and the hospitable netizens provide uneven answers. It is vitally important to provide accurate answers for such questions, because consumers are unable to judge the quality of medical contents. Thus rating and re-ranking the candidate answers to consumer health questions are the objectives of CHQA task. MEDIQA-2019 is the representative dataset (Abacha, Shivade, & Demner-Fushman, 2019), Xu, Liu, Li, Poon, and Gao (2019) cast it as a regression problem – a numerical score ranging from −2 to 2 is assigned to each QA instance, which effectively simplifies BERT's prediction process on this task.

Based on them, and considering the semantic integrity of topic at the subsection level, we create a MEDIQA-like dataset — TSMD, with the subsection level semi-structured information, and use it for intermediate fine-tuning.

3.2.2. Topic-subsection pairs

We convert the topic-paragraph pairs obtained in Section 3.1.2 to topic-subsection pairs. It is worth noting that the subsection here is a "functional subsection", which merges all paragraphs under the same topic. Then concatenate each "functional subsection" with its topic to get topic-subsection pairs (called subsection in the following content).

Fig. 1(b) shows the length distribution of subsections. Compare Fig. 1(b) with Fig. 1(c), it is evident that subsection usually has richer and fuller context about the topic. Therefore it is more favorable for creating MEDIQA-like QA instances with topic-subsection pairs, and rating them based on the matching degree between topic and subsection.

3.2.3. QA instance generation

We create QA instances with articles as units. First, the topic-subsection pairs are split by article, then to ensure a relatively balanced score distribution of QA instances, the articles with less than three topic-subsection pairs are filtered out. We take the remaining 4,619 articles as the collection of articles, denoted by $\mathbb{A}$. And two negative instances are generated for each positive one, to make TSMD have a similar data distribution with MEDIQA-2019.

Preparation: we randomly select an article $\mathcal{A}$ from $\mathbb{A}$, let $\mathbf{B}$ be the list of its topic-subsection pairs, and b be the randomly selected element from $\mathbf{B}$, to prepare for the positive instance. A non-b element $\bar{b}$ is randomly selected from $\mathbf{B}$, to prepare for one b-related negative instance. Besides, let $\mathcal{B}$ be the non-$\mathcal{A}$ article randomly selected from $\mathbb{A}$, and $\mathbf{C}$ be the list of its topic-subsection pairs. An element c is randomly selected from $\mathbf{C}$, to prepare for the other b-related negative instance.

The proportion of positive instances for an article should be set in advance. On one hand, we hope to get as many as QA instances from an article, on the other, it should leave choice space for the negative instances. Then the number of positive instances generated by $\mathcal{A}$ is:

$$n_p = len(\mathbf{B}) \times p_0 \tag{1}$$

where p_0 is the proportion of positive instances. To satisfy the above two conditions, we set p_0 as 0.4.

Note: To help you better understand the scoring mechanism about the matching degree between topic and subsection, we define two key topic-related concepts, the first filial (F1) topics, and the offspring topics. Here, we take the topic-subsection pairs of the Wikipedia article "COVID-19" as an example to illustrate the two concepts: the term "*COVID-19*" is seen as a maternal topic, and its F1 topics are the ones that contain and only contain its sub-level headings besides itself, such as *COVID-19* | *Etymology*, *COVID-19* | *Cause*, *COVID-19* | *Pathophysiology*, etc. While the topics descending from the root node – "*COVID-19*" are all its offspring topics, such as *COVID-19* | *Cause*, *COVID-19* | *Prevention* || *Vaccine* and *COVID-19* | *Mortality* || *Infection fatality rate* ||| *Estimates*.

Positive instance generated by b The topic of b is denoted as $topic_b$, and the subsection of b is denoted as $subsection_b$. We assume that the contribution of each F1 topic to the maternal topic is equal, if the number of $topic_b$'s F1 topics is t, then we can rate b by:

$$score_b = \begin{cases} 2/t & t > 0 \\ 2 & t = 0 \end{cases} \tag{2}$$

Negative instance generated by $\bar{b}$ We measure a topic's level by the level of its last heading element. The levels of $topic_b$ and $topic_{\bar{b}}$ are denoted as l_b and $l_{\bar{b}}$, respectively; the initial differentiation level of

$topic_b$ and $topic_{\bar{b}}$, i.e. the level of their first different heading element is denoted as l_d.

We combine $topic_b$ and $subsection_{\bar{b}}$ into a new instance b', and then rate b' according to the distance d between $topic_b$ and $topic_{\bar{b}}$, which can be further divided into 3 cases:

(1) if $l_b < l_{\bar{b}}$, and $topic_{\bar{b}}$ is an offspring topic of $topic_b$, then the distance d is:

$$d = l_{\bar{b}} - l_b \tag{3}$$

we assume that the F1 topics of $topic_b$ is t_1, and the number of topics at $l_{\bar{b}}$–level under the F1 topic it belongs to is dth power of 2, then we can rate b' by the following expression:

$$score_{b'} = (2/t_1)/2^d \tag{4}$$

(2) if $l_b > l_{\bar{b}}$, and $topic_b$ is an offspring topic of $topic_{\bar{b}}$, then the distance d is:

$$d = l_b - l_{\bar{b}} + 1 \tag{5}$$

when $topic_{\bar{b}}$ has t_2 F1 topics, similar to case (1) we rate b' by:

$$score_{b'} = (2/t_2)/2^d \tag{6}$$

(3) in addition to the above cases, the distance d between $topic_b$ and $topic_{\bar{b}}$ is:

$$d = \max(l_b, l_{\bar{b}}) - \min(l_b, l_{\bar{b}}) + 1 \tag{7}$$

and we can rate b' by:

$$score_{b'} = 2^d \times s_0 \tag{8}$$

s_0 is the score of unit-distance between topics under the same level, which can be measured by cosine similarity. We can obtain the embedding vector of a topic by word2vev (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). In this paper, the s_0 is set as −0.6, which is an average cosine similarity for 30 pairs topics.

Finally, to limit the scores within $[-2, 2]$, we constrain the scores via:

$$score_{b'} = \max(s_1, score_{b'}) \tag{9}$$

where s_1 is the minimum score set for b', we just need to make sure it falls in [−2, −1.75]. When we evaluate TSMD on downstream tasks, we find the effect of this value is kind of slight, here we set it as −1.95.

Negative instance generated by c We combine $topic_b$ and $subsection_c$ into a new instance b'', and we assume the topic-subsection pairs from different articles are independent of each other, then we rate b'' by:

$$score_{b''} = -2 \tag{10}$$

The complete procedure for automatically constructing the TSMD dataset is presented in Algorithm 1. Finally we obtained 32,695 TSMD instances.

Algorithm 1 TSMD construction procedure

```
Require: the articles collection, $\mathbb{A}$
Ensure: TSMD dataset
for each article $\mathcal{A}$ in $\mathbb{A}$ do
    Determine the number of positive instance $n_p$ by Eq. (1)
    Extract $n_p$ elements from List $\mathbf{B}$ based on simple random sampling method, to get List $\mathbf{B^1}$
    for each element $b$ in $\mathbf{B^1}$ do
        Rate $b$ by Eq. (2)
    end for
    Randomly sample $n_p$ elements from $set(\mathbf{B}) - set(\mathbf{B^1})$ to get List $\mathbf{B^2}$
    Let the elements from $\mathbf{B^1}$ and $\mathbf{B^2}$ be bijective, get the elements of List $\mathbf{B^3}$ by the way $b'$ generates
    for each element $b'$ in $\mathbf{B^3}$ do
        Rate $b'$ by Eq. (3)–Eq. (9)
    end for
    Randomly sample $n_p$ elements from the topic-subsection pair list $\mathbf{C}$ to get List $\mathbf{C^1}$
    Let the elements from $\mathbf{B^1}$ and $\mathbf{C^1}$ be bijective, get the elements of List $\mathbf{B^4}$ by the way $b''$ generates
    for $b''$ in $\mathbf{B^4}$ do
        Rate $b''$ by Eq. (10)
    end for
end for
```

4. Experiments

In this section, we will evaluate TDA over four disease-related tasks, i.e. CHQA, yes/no QA, medical language inference and disease name recognition. We expect CHQA, the topic-related task, will particularly benefit from it.

As previously mentioned, TMLM and TSMD are the key technologies of TDA. To systematically assess it, we design three experimental modes; (1) TPLM + TMLM, which conducts continual pre-training of an existing TPLM on the topic-paragraph pairs constructed in Section 3.1.2 with TMLM, can be used for all downstream tasks; (2) PLM + TSMD, which conducts intermediate fine-tuning of an existing TPLM on TSMD before fine-tuning on the target CHQA task; (3) PLM + TMLM + TSMD, which investigates the effect of using TMLM and TSMD sequentially before fine-tuning on CHQA. The rest of the tasks can hardly benefit from TSMD, for their low correlation.

4.1. Downstream tasks

Consumer Health Question Answering MEDIQA-2019 (Ben Abacha & Demner-Fushman, 2019) and TRECQA-2017 (Abacha, Agichtein, Pinter, & Demner-Fushman, 2017) are the two typical datasets of this task. Originally, a Reference Score (1 to 10) and a Reference Rank (4: Excellent, 3: Correct but Incomplete, 2: Related, 1: Incorrect) are assigned to each CHQA pair. Later, Xu et al. (2019) cast this task as a regression problem to predict the score, which greatly simplifies the task.

Yes/no QA We consider PubMedQA (Jin et al., 2019) for this task, it is collected from PubMed abstracts that use binary questions as titles (e.g.: Can vitamin C prevent complex regional pain syndrome in patients with wrist fractures?) and have structured abstracts. The task is to answer such research questions with yes/no/maybe using the corresponding abstracts (the conclusive parts are cropped) as contexts.

Medical Language Inference MEDNLI (Romanov & Shivade, 2018) is a clinical natural language inference (NLI) dataset, where a description about a patient from MIMIC-III clinical notes is seen as the premise, and clinicians generate three descriptions of it as hypotheses: a true one (entailment), a false one (contradiction), and one that might be true (neutral). It is clearly a multi-classification problem.

Disease Name Recognition NCBI (Doğan, Leaman, & Lu, 2014) and BC5CDR (Wei et al., 2016) are the datasets chose for the named entity recognition (NER) task, they are developed by medical experts annotating diseases mentioned in the collections of PubMed titles and abstracts. Peng, Yan, and Lu (2019) cast this task as a classification task to label tokens in sentences with B, I, or O.

As is shown in Table 2, the six datasets are small in size (ranging from 500 to 10,000 instances), with only hundreds (MEDIQA-2019 and TRECQA-2017) or even tens (PubMedQA) of Dev instances for the two QA tasks. Sellam et al. (2022) note the model's performance may vary for the multiple sources of randomness in experiments, i.e. the randomness due to the pre-training seed, the fine-tuning seed, and the finite test data. The main idea is to use the average behavior over seeds as a means of summarizing expected behavior in an ideal world with infinite samples. Thus following Gu et al. (2021), we report the average scores from ten runs for MEDIQA-2019, TRECQA-2017 and PubMedQA, and five runs for the other datasets.

Table 2
Summary of task datasets.

Datasets	Train	Dev	Test
MEDIQA-2019	1701	234	1107
TRECQA-2017	1969	234	839
PubMedQA	450	50	500
MEDNLI	11,232	1395	1422
BC5CDR-disease	4182	4244	4424
NCBI-disease	5145	787	960

Table 3
Hyperparameters for TPLM + TMLM.

Hyperparameter	Value
Max sequence length	384
Batch size	24
Learning rate	1e−5
Train epochs	5
Optimizer	AdamW
GPU	1 NVIDIA V100 GPU

4.2. Baselines

We take BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and SpanBERT (Joshi et al., 2020) the three general domain TPLMs, and BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021) the two biomedical domain TPLMs as the mian baselines. For fair comparison and carbon reduction, their base models from HuggingFace Transformers[4] are used in our experiments.

BERT A Transformer-based bidirectional language representation model. Hailed as a milestone, it set new states of the art on 11 NLU tasks, it has been a basic tool in NLU now. We find that the cased version is slightly better than the uncased version in preliminary experiments and therefore bert-base-cased is selected in our study.

RoBERTa A robustly optimized BERT, it made the following improvements in training method: (1) removing the next sentence prediction (NSP) task, each input is packed with full sentences sampled contiguously from one or more documents; (2) dynamic masking, it generated the masking pattern every time it feed a sequence to the model; (3) text encoding, it used a byte-level BPE vocabulary of 50 K subword units. We use roberta-base in experiments.

SpanBERT It differs from BERT by: (1) removing the NSP task; (2) span masking, it masks contiguous random spans of tokens; (3) span boundary objective, which used the representations of the tokens at the span's boundary to predict the span. SpanBERT outperforms BERT in nearly all tasks, and spanbert-base-cased is used in our experiments.

BioBERT As the first biomedical-domain TPLM, Bio-BERT initializes with Google BERT pre-trained with the general-domain text and inherits its vocabulary, then further pre-trains on the biomedical-domain text. Again, biobert-base-cased-v1.1 is selected.

PubMedBERT It generates an in-domain vocabulary and pre-trains BERT from scratch with purely the biomedical-domain text. PubMedBERT is the first to show "for domains with abundant unlabeled text, pretraining language models from scratch results in substantial gains over continual pretraining of general-domain language models" (Gu et al., 2021). It only trained uncased models, so PubMedBERT-Base-Uncased-abstract-fulltext is selected.

In addition, **diseaseBERT** (He, Zhu, et al., 2020) and **DAKI-BERT** (Lu et al., 2021) that encoded domain knowledge stored in multiple sources via adapters (Houlsby et al., 2019) are relatively new domain adaptation methods, and they are taken as supplement baselines. It should be noted that continual pre-training with BERT's vanilla MLM and SpanBERT's span masking on the topic-paragraph pairs constructed in Section 3.1 is included as part of the ablation study.

4.3. Implementation

We initialize a language model with the pre-trained parameters of a baseline model, then adopt the proper mode of TDA before fine-tuning on the downstream tasks. We use AdamW (Loshchilov & Hutter, 2017) to update the model parameters across the entire experiment, including further training with TMLM, intermediate fine-tuning on TSMD and, fine-tuning on the downstream tasks. We set $max\_seq\_length = 384$ for TMLM pre-training, TSMD intermediate fine-tuning, and all QA tasks. The rest hyperparameters are inherited from diseaseBERT (He et al. 2020). We list the main hyperparameters of TMLM in Table 3.

For TPLM + TMLM, our pre-training corpus (52 MB) is about 2.5 times bigger than that of diseaseBERT (He, Zhu, et al., 2020), then a longer training time is needed. When the mode is performed on one NVIDIA V100 GPU, it takes about 80 min to complete one training epoch, and just 1–5 epochs are enough to enhance PLMs' better performance on the four downstream tasks. For TPLM + TSMD, owning to the smaller size of TSMD dataset (35MB), intermediate fine-tuning on TSMD is faster (about 30 min). And it takes no more than 10 epochs to reach its best performance.

4.4. Results

Fig. 5 shows the impact of the input max_seq_length length on PubMedBERT + TMLM. And We take the performance of PubMedBERT + TMLM on MEDIQA-2019 to illustrate it. We can see the performance continually improves with max_seq_length, but when it reaches $max\_seq\_length = 384$, the rapid rising slows down. The figure explains to some extent the reason why we set $max\_seq\_length = 384$ is that: (1) The lengths (in tokens) of topic-paragraph pairs used by TMLM are generally within 384 (the ratio is 99.43); (2) In contrast with the full length (512 tokens), the smaller max_seq_length can accelerate training, and nearly no performance degrade.

Table 4 shows the performance of consumer health QA tasks. Predictably, the topic-related task benefits a lot from TDA. The part of the best results, TPLM + TSMD and TPLM + TMLM + TSMD, almost equal shares. TPLM + TSMD increases the accuracy by 5% for MEDIQA-2019 and 3.56% for TRECQA-2017 on average. TPLM + TMLM + TSMD increases the accuracy by 5.44% for MEDIQA-2019 and 4.12% for TRECQA-2017 on average. The results suggest: (1) The three experimental modes all work, although TPLM + TMLM is less prominent in this task, but it has a more widely application, which can be shown in later part of the paper; (2) TPLM + TMLM + TSMD is slightly better than TPLM + TSMD, which suggests continual pre-training on the two tasks is more effective than on just one; (3) Our TSMD constructed with the semi-structured information and the CHQA datasets embody the same essence, thereby can help it via transfer learning; (4) Intermediate fine-tuning is more effective in capturing task-specific domain knowledge than continual pre-training with TMLM. Overall, the excellent performance of TDA confirms our predictions that the semi-structured information does have the potential to help TPLM learn more domain knowledge.

The results of the Yes/no QA task are shown in Table 5. TMLM helps TPLMs do better on PubMedQA, attaining a performance boost of 3.32% absolute on average. We conjecture that it rewards the learning of semantic relationship between the paragraph and its topic, which involves domain knowledge and logical reasoning ability needed by the Yes/no QA task. As for the results of MEDNLI, we see from Table 5 that TMLM makes limited improvement, by 0.9% on average. And different from the previous experiments, the biomedical TPLMs benefit more from TMLM than the general TPLMs, which can be explained by the

[4] https://huggingface.co/models

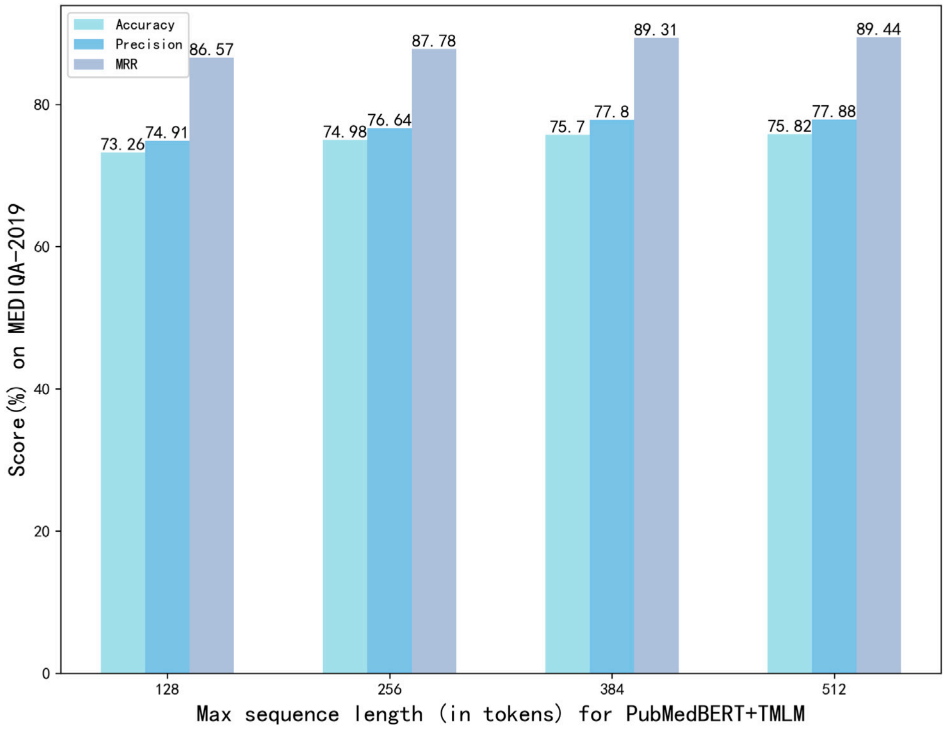


Fig. 5. Impact of the input *max_seq_length* length on PubMedBERT + TMLM. The performance continually improves with *max_seq_length*, but when it reaches *max_seq_length* = 384, the rapid rising slows down.

Table 4
Experimental results on consumer health QA task.

Datasets	MEDIQA-2019			TRCEQA-2017		
Metrics(%)	Accuracy	MRR	Precision	Accuracy	MRR	Precision
BERT (Devlin et al., 2019)	67.75	79.28	72.22	79.02	52.48	62.12
BERT + TMLM	71.91	83.56	75.84	81.05	**53.13**	64.96
BERT + TSMD	**73.98**	83.22	78.29	81.41	51.76	66.5
BERT + TMLM + TSMD	73.62	**86.22**	**79.91**	**81.88**	51.28	**67.10**
RoBERTa (Liu et al., 2019)	70.1	83.74	70.67	76.16	**43.59**	57.8
RoBERTa + TMLM	71.43	81.22	73.53	78.22	41.2	60.77
RoBERTa + TSMD	**76.15**	**89.06**	**77.72**	**81.13**	43.27	**71.28**
RoBERTa + TMLM + TSMD	75.79	87.02	77.05	80.57	43.44	67.19
SpanBERT (Liu et al., 2019)	66.31	83.33	67.4	69.01	47.41	45.13
SpanBERT + TMLM	70.64	85.83	73.43	76.28	**47.44**	58.72
SpanBERT + TSMD	71.54	85	72.92	77.71	45.03	58.44
SpanBERT + TMLM + TSMD	**73.8**	**85.89**	**76.34**	**80.43**	45.99	**69.62**
BioBERT (Lee et al., 2020)	71.54	84.44	73.67	80.45	50.96	65.29
BioBERT + TMLM	74.43	88.72	76.53	80.81	**55.72**	64.05
BioBERT + TSMD	**75.79**	89.53	**79.88**	**81.41**	55.53	65.13
BioBERT + TMLM + TSMD	74.8	**89.56**	79.36	81.17	54.65	**66.36**
PubMedBERT (Gu et al., 2021)	72.9	84	77.25	80.45	52.24	62.96
PubMedBERT + TMLM	75.7	89.11	77.84	81.76	**54.65**	**67.3**
PubMedBERT + TSMD	76.24	86.34	**83.56**	81.29	54.33	66.22
PubMedBERT + TMLM + TSMD	**77.78**	**92.22**	81.71	**81.88**	54.11	67.28
diseaseBERT (He, Zhu, et al., 2020)	66.40	83.33	68.94	75.33	**56.41**	54.01
DAKI-BERT (Lu et al., 2021)	69.47	**85.06**	70.17	77.95	54.65	58.27
diseaseBioBERT (He, Zhu, et al., 2020)	72.09	**87.78**	74.40	78.43	**54.76**	58.45
DAKI-BioBERT (Lu et al., 2021)	72.54	87.33	77.46	78.55	54.17	59.04

knowledge needed by MEDNLI is distinguished from those captured by TMLM, more logical reasoning ability instead of more domain knowledge is needed by this task. Making TPLM to incorporate more logical reasoning ability will be our future direction. Table 5 also shows the performance on NER task. For BC5CDR, the accuracy of the models equipped with TMLM increases by 0.45% on average, and 0.58% for NCBI. Although NER-related task is not covered in TDA, it still works, which probably owing to that TMLM forces PLMs remember the disease terms during pre-training.

Fig. 6 shows the impact of train epoch of PubMedBERT + TSMD on MEDIQA-2019. We can see that it takes just two train epochs to reach the best performance, on both accuracy and precision. And MRR shows

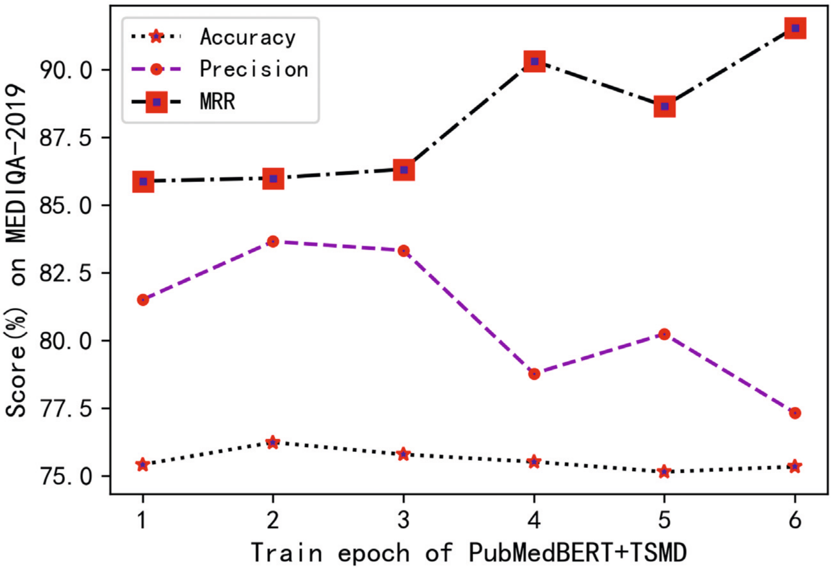


Fig. 6. Impact of train epoch of PubMedBERT + TSMD on MEDIQA-2019. It takes just two train epochs to reach the best performance, on both accuracy and precision.

Table 5
Experimental results on Yes/no QA, NLI and NER tasks.

Tasks Datasets Metrics(%)	Yes/no QA PubMedQA Accuracy	NLI MEDNLI Accuracy	NER BC5CDR F1	NER NCBI F1
BERT (Devlin et al., 2019)	55.4	78.83	83.28	85.56
BERT + TMLM	**58.2**	**79.82**	**84.23**	**86.52**
RoBERTa (Liu et al., 2019)	55.6	82.49	83.47	87.01
RoBERTa + TMLM	**57.4**	**83.54**	**83.7**	**87.63**
SpanBERT (Joshi et al., 2020)	55.2	80.66	84.18	87.13
SpanBERT + TMLM	**58.8**	**80.8**	**84.42**	**88.32**
BioBERT (Lee et al., 2020)	60.2	82.77	85.58	87.70
BioBERT + TMLM	**61.4**	**84.04**	**86.13**	**87.91**
PubMedBERT (Gu et al., 2021)	55.8	83.76	87.82	88.3
PubMedBERT + TMLM	**63**	**84.6**	**87.89**	**88.83**
diseaseBERT (He, Zhu, et al., 2020)	56.6	77.29	**83.47**	**86.81**
DAKI-BERT (Lu et al., 2021)	**57.1**	**77.85**	83.43	85.67
diseaseBioBERT (He, Zhu, et al., 2020)	60.7	82.21	**86.52**	87.14
DAKI-BioBERT (Lu et al., 2021)	**61.2**	**83.41**	86.51	**89.01**

a contrary growth trend compared to the former two metrics. Now we know the knowledge transfer from TSMD to MEDIQA-2019 is efficient, which in turn demonstrate our TSMD is indeed a MEDIQA-like dataset.

4.5. Ablation study

In the above experiments, the masking strategies used by these TPLMs can be divided into two classes: Vanilla MLM used by BERT, RoBERTa, BioBERT and PubMedBERT; span masking used by SpanBERT. And we have shown no matter what masking strategy the TPLM used during pre-training, they all benefited from our selective masking strategy during continual pre-training. Now we want to know: (1) whether our selective masking is superior than the two vanilla masking strategies, thus we conduct continual pre-training BERT with vanilla MLM and span masking on the in-domain text corpus obtained in Section 3.1, for a fair comparison. (2) whether the Hyperparameter we set for our selective masking is optimal, thus we compare the effect of different making rates for the heading elements and random masking or not. We report the ablation study on MEDIQA-2019 and PubMedQA, the results are shown in Table 6. Similar results are observed on the remaining tasks, but omitted here due to the space limitation.

Table 6 shows the superiority of our selective masking: (1) For random masking, increasing the heading elements masking rate in a certain range is suitable for this task, and 50% works best; (2) Our selective masking (Default) works better than the randomized one (50% random heading masking), which on one hand shows masking heading elements

Table 6
Ablation study on MEDIQA-2019 and PubMedQA.

Datasets	MEDIQA-2019			PubMedQA
Metrics(%)	Accuracy	MRR	Precision	Accuracy
Default (selective masking)	**71.91**	**83.56**	**75.84**	**58.2**
75% random heading masking	70.55	83.28	71.65	56
50% random heading masking	71.18	82.61	73.82	57.2
30% random heading masking	70.73	82.94	74.31	56.4
15% random heading masking	70.46	82.84	73.33	55.8
BERT + Vanilla MLM	69.74	79.0	74.23	57
BERT + Vanilla span masking	68.85	83.34	73.02	57.8

enables TPLM to capture more domain knowledge than the randomized one, and on the other reveals the semi-structured information offers more possibilities for domain knowledge learning; (3) For BERT, TMLM is a more effective masking strategy in capturing domain knowledge than the two vanilla the masking strategies, especially in the CHQA task, we owning this to the domain knowledge captured by TMLM is just what CHQA needs.

5. Conclusion and future work

In this paper, we show the importance of semi-structured information of documents for BERT models capturing domain knowledge. Firstly, we realize the value of semi-structured information in human learning domain knowledge and design a novel pre-training corpus construction method, which incorporates the semi-structured information well. Secondly, we propose TDA, which enhances the capture of domain knowledge at various levels of text granularity. As key technologies of TDA, TMLM and TSMD enable BERT to capture the domain knowledge embedded in paragraph and subsection respectively during pre-training and intermediate fine-tuning. The experimental results show the effectiveness of TDA on four biomedical domain tasks, and a significant improvement is observed in the QA tasks, especially the topic-related one — CHQA. The last that must be emphasized is that our TDA can be easily applied to other domain, even the general domain. Because of the experimental conditions and equipment limitations, the further validation of TDA on the other domain can be served as part of the future research work.

Although our TDA works well from the experimental results, it still has room for improvement: (1) Hyperparameters optimization. While it usually requires huge computation, a small-scale grid search is necessary; (2) TDA implementation. The three modes of TDA all belong to the continual learning, with the drawback that it may forget the previously learned knowledge, and our experiments also confirm this. Fortunately, sequential multi-task learning from ERNIE 2.0 (Sun et al., 2020) and adapters (Houlsby et al., 2019; Lu et al., 2021; Wang et al., 2021) are both effective solutions to the problem. (3) Integration of TMLM and MLM. Our TMLM and BERT's MLM are basically the same. However, we use them in different phases, which would increase computation and lead to the same problem of continual learning. Therefore, integrating the two into one training procedure is a valuable question to be studied. These limitations can be directions for future work.

CRediT authorship contribution statement

Jing Chen: Conceptualization of this study, Methodology, Software, Writing – original draft. **Zhihua Wei:** Project administration, Supervision, Conceptualization, Writing – review & editing. **Jiaqi Wang:** Data curation, Software, Writing – original draft. **Rui Wang:** Methodology, Software, Investigation, Writing – review& editing. **Chuanyang Gong:** Investigation. **Hongyun Zhang:** Supervision. **Duoqian Miao:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

The work is partially supported by the National Natural Science Foundation of China (No. 61976160, 61906137, 61976158, 62076184, 62076182) and Shanghai Science and Technology Plan Project (No. 21DZ1204800) and Technology research plan project of the Ministry of Public and Security (Grant No. 2020JSYJD01).

References

Abacha, A. B., Agichtein, E., Pinter, Y., & Demner-Fushman, D. (2017). Overview of the medical question answering task at TREC 2017 LiveQA. In *TREC* (pp. 1–12).

Abacha, A. B., Shivade, C., & Demner-Fushman, D. (2019). Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP workshop and shared task* (pp. 370–379).

Aghajanyan, A., Okhonko, D., Lewis, M., Joshi, M., Xu, H., Ghosh, G., et al. (2021). HTLM: Hyper-text pre-training and prompting of language models. In *International conference on learning representations*.

Bajgar, O., Kadlec, R., & Kleindienst, J. (2016). Embracing data abundance: Booktest dataset for reading comprehension. arXiv preprint arXiv:1610.00956.

Ben Abacha, A., & Demner-Fushman, D. (2019). A question-entailment approach to question answering. *BMC Bioinformatics*, *20*(1), 1–23.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). http://dx.doi.org/10.18653/v1/2020.findings-emnlp.261, URL: https://aclanthology.org/2020.findings-emnlp.261.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2020). Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 2978–2988).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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* (pp. 4171–4186).

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, *47*, 1–10.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1–23.

Gu, Y., Zhang, Z., Wang, X., Liu, Z., & Sun, M. (2020). Train no evil: Selective masking for task-guided pre-training. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 6966–6974). URL: https://aclanthology.org/2020.emnlp-main.566.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., et al. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 8342–8360). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.740, Online. URL: https://aclanthology.org/2020.acl-main.740.

He, B., Zhou, D., Xiao, J., Jiang, X., Liu, Q., Yuan, N. J., et al. (2020). BERT-MK: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2281–2290). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.207, Online. URL: https://aclanthology.org/2020.findings-emnlp.207.

He, Y., Zhu, Z., Zhang, Y., Chen, Q., & Caverlee, J. (2020). Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 4604–4614). URL: https://aclanthology.org/2020.emnlp-main.372.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., et al. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, *28*.

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016). The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of 4th international conference on learning representations*.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American society for information science*, *46*(6), 400–425.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., et al. (2019). Parameter-efficient transfer learning for NLP. In *International conference on machine learning* (pp. 2790–2799).

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). 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* (pp. 2567–2577). URL: https://aclanthology.org/D19-1259.

Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., et al. (2022). Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*, *55*(2), 1–36.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, *8*, 64–77.

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). AMMU: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, Article 103982.

Kim, S., Park, D., Choi, Y., Lee, K., Kim, B., Jeon, M., et al. (2018). A pilot study of biomedical text comprehension using an attention-based deep neural reader: Design and experimental analysis. *JMIR Medical Informatics*, *6*(1), Article e8751.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite bert for self-supervised learning of language representations. In *Proceedings of international conference on learning representations*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., et al. (2020). K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2901–2908).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lu, Q., Dou, D., & Nguyen, T. H. (2021). Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3855–3865).

Meng, Z., Liu, F., Clark, T., Shareghi, E., & Collier, N. (2021). Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 4672–4681). Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.383, Online. URL: https://aclanthology.org/2021.emnlp-main.383.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Pappas, D., Androutsopoulos, I., & Papageorgiou, H. (2018). BioRead: A new dataset for biomedical reading comprehension. In *Proceedings of the eleventh international conference on language resources and evaluation*.

Pappas, D., Stavropoulos, P., Androutsopoulos, I., & McDonald, R. (2020). BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing* (pp. 140–149).

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP workshop and shared task* (pp. 58–65). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-5006, URL: https://aclanthology.org/W19-5006.

Poerner, N., Waltinger, U., & Schütze, H. (2020). Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and Covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1482–1490). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.134, Online. URL: https://aclanthology.org/2020.findings-emnlp.134.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. http://dx.doi.org/10.1162/tacl_a_00349, URL: https://aclanthology.org/2020.tacl-1.54.

Romanov, A., & Shivade, C. (2018). Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1586–1596). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1187, URL: https://aclanthology.org/D18-1187.

Sellam, T., Yadlowsky, S., Tenney, I., Wei, J., Saphra, N., D'Amour, A., et al. (2022). The MultiBERTs: BERT reproductions for robustness analysis. In *International conference on learning representations*. URL: https://openreview.net/pdf?id=K0E_F0gFDgA.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., et al. (2020). Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8968–8975).

Tai, W., Kung, H., Dong, X. L., Comiter, M., & Kuo, C.-F. (2020). exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1433–1439).

Wang, R., Tang, D., Duan, N., Wei, Z., Huang, X., Ji, J., et al. (2021). K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1405–1418). URL: https://aclanthology.org/2021.findings-acl.121.

Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., et al. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative v chemical-disease relation (CDR) task. *Database*, *2016*.

Xu, Y., Liu, X., Li, C., Poon, H., & Gao, J. (2019). Doubletransfer at mediqa 2019: Multi-source transfer learning for natural language understanding in the medical domain. In *Proceedings of the 18th BioNLP workshop and shared task*. URL: https://aclanthology.org/W19-5042.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.

Yao, Y., Huang, S., Wang, W., Dong, L., & Wei, F. (2021). Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 460–470).

Yasunaga, M., Leskovec, J., & Liang, P. (2022). LinkBERT: Pretraining language models with document links. In *Association for computational linguistics*.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 535–546). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.45, Online. URL: https://aclanthology.org/2021.naacl-main.45.

Zhang, N., Chen, M., Bi, Z., Liang, X., Li, L., Shang, X., et al. (2022). CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 7888–7915). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.544, URL: https://aclanthology.org/2022.acl-long.544.

Zhang, R., Gangi Reddy, R., Sultan, M. A., Castelli, V., Ferritto, A., Florian, R., et al. (2020). Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 5461–5468). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.440, Online. URL: https://aclanthology.org/2020.emnlp-main.440.

Zhu, H., Peng, H., Lyu, Z., Hou, L., Li, J., & Xiao, J. (2023). Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation. *Expert Systems with Applications*, *215*, Article 119369. http://dx.doi.org/10.1016/j.eswa.2022.119369, URL: https://www.sciencedirect.com/science/article/pii/S0957417422023879.

Zhihua Wei is currently a Professor at Tongji University. She received a Ph.D. degree pattern recognition and intelligent system from Tongji University in China, a Ph.D. degree in Information from Lyon2 University in France, and B.S. and M.S. degrees both in Computer Science from Tongji University in China. Her current research interests include machine learning, natural language processing and speech processing. she is a member of the granular computing and knowledge discovery Professional Committee of Chinese Artificial Intelligence Society, and a member of the natural understanding professional committee of Chinese Artificial Intelligence Society.