SUPPLEMENTING DOMAIN KNOWLEDGE TO BERT WITH SEMI-STRUCTURED INFORMATION OF DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

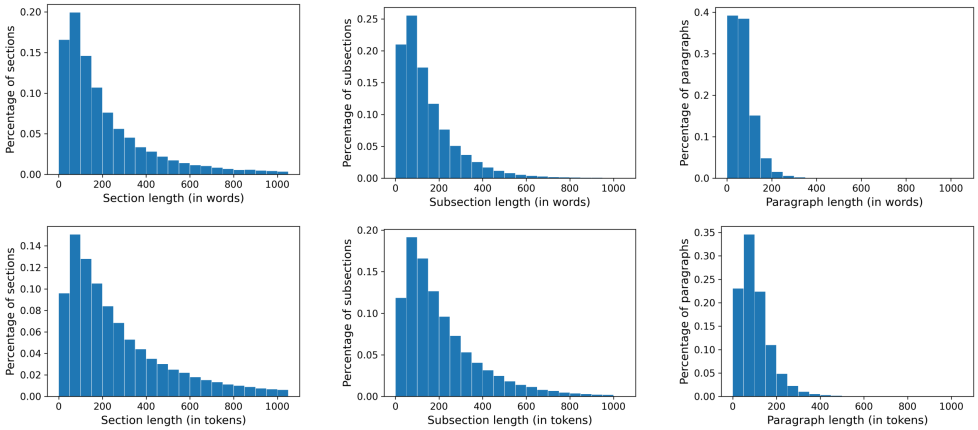
Adapting BERT on in-domain text corpus is a good way to boost its performance on domain-specific natural language processing (NLP) tasks. Common domain adaptation methods, however, can be deficient in capturing domain knowledge. Meanwhile, the context fragmentation inherent in Transformer-based models also hinders the acquisition of domain knowledge. Given the semi-structural characteristics of documents and their potential for alleviating these problems, we leverage 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 model 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 three biomedical NLP tasks across five datasets, and the results validate the importance of the previously overlooked semi-structured information for domain adaptation. Our method benefits BERT, RoBERTa, BioBERT, and PubMedBERT in nearly all cases and yield significant gains over the topic-related task, question answering, with an average accuracy improvement of 4.

1 INTRODUCTION

BERT-like models (Devlin et al., 2019; Liu et al., 2019; Joshi et al., 2020; Lan et al., 2020) learn rich syntactic, semantic, and world knowledge in general domain text during pre-training (Rogers et al., 2020), and can subsequently be applied to target domain tasks via fine-tuning. To close the gap between target domain and general domain, several domain-customized BERT models (Lee et al., 2020; Gu et al., 2021; Yang et al., 2020; Chalkidis et al., 2020) have been released, most of which are achieved by continual pre-training of general-domain language models or pre-training language models from scratch over in-domain text – and new state-of-the-art results are observed in many domain-specific natural language processing (NLP) tasks. However, such domain adaptation methods can be deficient in capturing the domain knowledge focused by domain experts, while they are skilled in learning universal in-domain language representations (Kalyan et al., 2021).

The semi-structured information, i.e. heading and hierarchy of documents plays a significant role when we learn domain knowledge. Heading is a brief statement that identifies the central argument of the article or section, which divides into first-level heading (h1, i.e., the article title), second-level heading (h2, i.e., the section title), third-level heading (h3, i.e., the subsection title), etc. Hierarchy is the order in which the ideological content of an article is expressed, reflecting the development stages of objective things or all aspects of contradictions. It often consists of multiple paragraphs, which can be subdivided into section, subsection, with paragraph being the smallest hierarchy. Take "COVID-19" on Wikipedia as an example, cause, diagnosis and treatment in the table of contents are second-level headings and section divisions, combining heading and hierarchy helps readers learn all aspects of the disease, i.e., the domain knowledge interests doctors and patients.

Regrettably, the semi-structured information has been seriously neglected: 1) the pre-training data from Wikipedia only retains text passages, whereas headers, lists, and tables are ignored (Devlin et al., 2019; Liu et al., 2019); 2) for BERT models, either two segments of text are sampled and



(a) Sections (b) Subsections (c) Paragraphs

Figure 1: The length distributions of sections, subsections and paragraphs.

then concatenated as a training input sequence (Devlin et al., 2019) or the input is packed with full sentences sampled contiguously from one or more documents (Liu et al., 2019; Joshi et al., 2020). In addition, Transformer-based models require a fixed-length input sequence of up to 512 tokens, so chunking is needed for long texts, and thereby the context fragmentation problem seems inevitable (Dai et al., 2020). All the above factors hinder models to learn domain knowledge better.

Some researchers try to resolve the above problems. Dai et al. (2020) proposes Transformer-XL based on a segment-level recurrence mechanism to improve vanilla Transformer with the context fragmentation problem. So far the pre-trained language models (PLMs) based on Transformer-XL are relatively few (Yang et al., 2019). DiseaseBERT (He et al., 2020) introduces a disease knowledge infusion training procedure, where the training sequences are question-answer pairs, with question constructed using disease (article title) and aspect (section title) and answer being the whole section; and masked language model (MLM) is merely active to the title tokens in training sequences. However, there are some drawbacks: 1) its usage of the semi-structured information is limited, the questions only covers two kinds of titles; 2) sections are generally long, and simple chunking still leads to contextual fragmentation. Figure 1 (a) and (c) illustrate this more intuitively that 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 likely to alleviate contextual fragmentation.

In this paper, we propose topic-based domain adaptation (TDA), which enables BERT to better learn domain knowledge with the semi-structured information of documents. This method emphasizes the intrinsic relation among heading, hierarchy, and domain knowledge, and enhances the capture of domain knowledge at various levels of text granularity. Specifically, at the paragraph level, we create topic-paragraph pairs as training sequences, then theme masked language model (TMLM) is designed, which selectively masks some headings in the topic part to force 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, theme subsection matching degree (TSMD) dataset is automatically constructed, which is used for intermediate fine-tuning, to help target task via transfer learning.

We evaluate our TDA on three tasks in biomedical domain, including consumer health question answering, medical language inference, and disease name recognition. And we implement TDA in three modes, i.e., PLM+TMLM, PLM+TSMD, and PLM+TMLM+TSMD. The results show that (1) TMLM benefits the BERT models in all tasks, especially on the QA tasks. For example, the accuracy of BERT on the MEDIQA-2019 is improved from 67.75% to 71.91%; (2) BERT models intermediate fine-tuned on TSMD gain more performance improvements than that of TMLM on QA task, with an average accuracy improvement of 5.4; (3) The performances of training with TMLM and TSMD sequentially fall somewhere in between most of the time. TDA can be easily drawn on to other domains. The code and dataset will be publicly available.

2 ADDITIONAL RELATED WORK

Domain knowledge enhanced PLM. Many studies have shown continual pre-training and domain-specific pre-training are effective domain adaptation methods, recently, some researchers improved them based on domain features. Considering the features of in-domain text corpus, general domain vocabulary can be extended with in-domain vocabulary (Poerner et al., 2020; Tai et al., 2020; Yao et al., 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) presents task-adaptive pre-training – it involves further pre-training on task-related unlabelled instances, Gu et al. (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 using the inherent structure in unlabeled data for intermediate fine-tuning.

Clever use of the semi-structured information. Quite a few cloze-style QA datasets are automatically created using part of the semi-structured information. They were created using the semi-structured information from either news articles (Hermann et al., 2015) or books (Hill et al., 2016; Bajgar et al., 2016) or scientific literature (Pappas et al., 2018; 2020; Kim et al., 2018). These datasets are usually vast and thus can be used for pretraining (Dhingra et al., 2018) or as tasks (Hermann et al., 2015; Pappas et al., 2020; Kim et al., 2018). However, they are generally noisy due to the limited use of semi-structured information. Besides, the non-cloze-style QA datasets, PubMedQA (Jin et al., 2019) and MedQuAD (Ben Abacha & Demner-Fushman, 2019), use the semi-structured information more accurately, which allows their data quality to be greatly enhanced.

Inspired by these studies, our TDA aims to make BERT capture more domain knowledge with better use of the semi-structured information. The novel domain adaptation framework involves TMLM and TSMD the two key technologies, which enable BERT to capture domain knowledge embedded in paragraph and subsection respectively during multiple training phases.

3 TOPIC MASKED LANGUAGE MODEL

In this section, we introduce a new pre-training task, topic masked language model (TMLM), to enable PLMs to capture the domain knowledge contained in paragraph. It consists of two steps: 1) construct a pre-training corpus with the paragraph level semi-structured information; 2) propose a topic masking strategy. In the following, we will discuss each step in more detail.

3.1 PRE-TRAINING CORPUS

Following He et al. (2020), we verify the effectiveness of TDA in the biomedical domain, and the disease-related articles in English Wikipedia are used as in-domain text source. To get as many articles as possible, we collect disease terms from two main branches, Diseases [C] and Mental Disorders [F03], of the Medical Subject Headings (MeSH) tree[1]. Besides,the Wikipedia page "Category: Lists of diseases" [2] serves as a supplement source of disease terms. After eliminating those duplicate or empty entries, 4,930 disease-themed English Wikipedia articles are obtained.

To construct an in-domain text corpus with the semi-structured information, we retain the heading and hierarchy of articles in web crawler phase. During data cleaning, the texts that are irrelevant to the topic of the article and the images, complicated tables, and special characters that are hard to process by PLMs are filtered out to reduce data noise. The in-domain text corpus we get is further organized into the paragraph level pre-training corpus – topic-passage pairs.

As discussed in §1, paragraphs are better as answers 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 its topic, thereby a topic-passage pair is generated. The whole process is shown in Figure 2. When a table or a list appears supplementary

[1] https://meshb.nlm.nih.gov/treeView

[2] https://en.wikipedia.ahmu.cf/wiki/Category:Lists_of_diseases

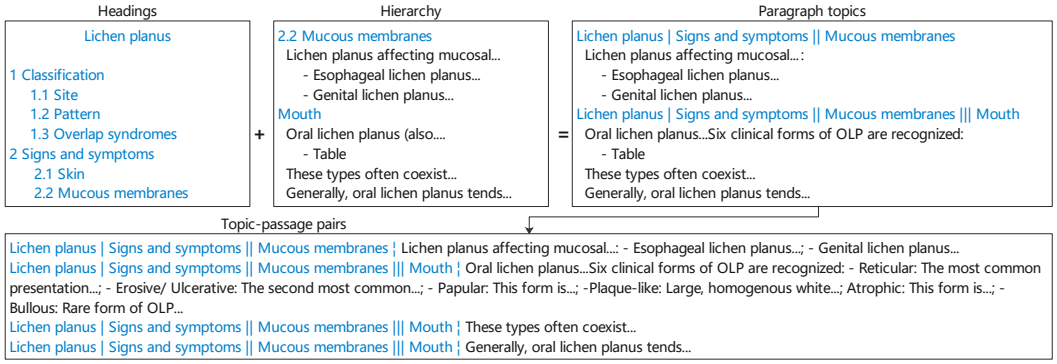


Figure 2: Topic-passage Pairs Generation

to one paragraph, we do the following: 1) convert the table to a list, 2) convert the list to plain text, and add it into the paragraph.

The statistics of the pre-training corpus are shown in Table 1. For fair comparison, we also get the section level in-domain text corpus, and generate topic-section pairs following the similar process in Figure 2, and obtain their statistics. As you can see from this table, 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 our paragraph level pre-training corpus in the high usage of the semi-structured information and the potential to reduce context fragmentation.

Table 1: Statistics of the pre-training corpus

Number of articles	4,930
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 passages	104,696
Average length of passages (in words)	72.47
Average length of passages (in tokens)	104.39
Average length of passage topics (in words)	4.27
Average length of passage topics (in tokens)	7.94

3.2 TOPIC MASKING

To help BERT capture the domain knowledge contained in paragraph, we propose a topic masking strategy, 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, mask it; if the number of heading elements in the topic exceeds 1, mask the odd and even heading elements with equal probability. Thereby the topic part can serve as a question, and the passage is the target answer. The final training instances are shown in Figure 3.

Figure 3: Examples of topic masking strategy

In this way, TMLM forces BERT to learn the semantic relationship between the paragraph and its topic during pre-training, thereby capture the domain knowledge embedded in paragraphs.

4 THEME SUBSECTION MATCHING DEGREE DATASET

Gu et al. (2020) note that insufficient supervised data is frequently a matter during fine-tuning, especially for specific domains, which results in PLM's poorer performance in domain-specific tasks. However, intermediate fine-tuning on large, related datasets allows PLMs to learn more domain-specific and task-specific patterns, which improves the performance on small target datasets (Kalyan et al., 2021). Inspired by this theory, we again use the subsection level semi-structured information to automatically create a large dataset - theme subsection matching degree (TSMD), which is used for intermediate fine-tuning to help the target task, consumer health QA (CHQA), via transfer learning.

4.1 KNOWLEDGE ARTICLES

The objective of the CHQA task like MEDIQA-2019 (Abacha et al., 2019) and TRECQA-2017 (Abacha et al., 2017) is to rate and re-rank candidate answers to consumer health questions. Xu et al. (2019) cast this task as a regression problem where numerical scores ranging from -2 to 2 are assigned to QA instances. Inspired by it, we reorganize the pre-training corpus obtained in §3.1 by merging all paragraphs under the same topic into a "functional subsection" (called subsection in the following content), then the new topic-subsection pairs are available. On this basis, we will generate MEDIQA-2019-like QA instances at the article level.

First, the topic-subsection pairs are split by article, and then to ensure a 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 $\mathbb{A}$. Figure 1 (b) shows the length distribution of subsections. Compare Figure 1(b) with Figure 1(c), it is evident that subsection usually has richer and fuller context about the topic, which is more favorable for us rating topic-subsection pairs based on the matching degree of a topic and its subsection.

4.2 QUESTION-ANSWER PAIRS

In this section, the generation of QA instances is depicted in detail. We generate two negative instances for each positive instance to ensure the diversity of negative instances.

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

When we generate QA instances at the article level, the number of positive instances should be set in advance to leave a degree of choice for the negative instances, then for article $\mathcal{A}$, we have:

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

where p_0 is the proportion of positive instances, we empirically set it as 0.4.

4.2.1 POSITIVE INSTANCE GENERATED BY b

Note: we define two key topic-related concepts, the first filial (F1) topics, and the offspring topics. Here, we take the topic-subsection pairs within 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*.

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:

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

Algorithm 1 TSMD construction procedure

Input: the articles collection, $\mathbb{A}$
Output: TSMD dataset
for each article $\mathcal{A}$ in $\mathbb{A}$ **do**
  Determine the number of positive instance n_p by Equ.(1)
  Randomly sample n_p elements from its topic-subsection pair list $\mathbf{B}$ to get List $\mathbf{B^1}$
  for each element b in $\mathbf{B^1}$ **do**
    Rate b by Equ.(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 Equ.(3) - Equ.(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 Equ.(10)
  end for
end for

4.2.2 NEGATIVE INSTANCE GENERATED BY $\overline{b}$

Note: we measure a topic's level by the level of its last heading element. The levels of $topic_b$ and $topic_{\overline{b}}$ are denoted as l_b and $l_{\overline{b}}$, respectively; the initial differentiation level of $topic_b$ and $topic_{\overline{b}}$ i.e., the level of their first different heading element is denoted as l_d.

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

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

$$d = l_{\overline{b}} - l_d \tag{3}$$

we assume that there are t_1 F1 topics of $topic_b$, and the number of $l_{\overline{b}}$–level topics is d-th power of 2, then we rate b' by the following expression:

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

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

$$d = l_b - l_d + 1 \tag{5}$$

when $topic_{\overline{b}}$ 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 $topic_{\overline{b}}$ is:

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

and we can rate b' by:

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

here we empirically set s_0 as -0.6.

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', here we set it as -1.95.

4.2.3 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.

5 EXPERIMENTS

In this section, we evaluate topic-based domain adaptation (TDA) over three tasks in biomedical domain. As you know, theme masked language model (TMLM) and theme subsection matching degree (TSMD) dataset are the key contributions of TDA. In order to evaluate TDA more scientifically, we design three modes: 1) PLM+TMLM, which conducts continual pre-training of a general-domain PLM on the pre-training corpus constructed in §3.1 with TMLM; 2) PLM+TSMD, which conducts intermediate fine-tuning of a general-domain PLM on TSMD before fine-tuning on the target consumer health QA (CHQA) task; 3) PLM+TMLM+TSMD, which investigates the effect of using TMLM and TSMD sequentially before fine-tuning on CHQA.

Table 2: 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
BioBERT (Lee et al., 2020)	71.54	84.44	73.67	79.26	50.96	62.01
BioBERT+TMLM	74.43	88.72	76.53	80.21	52.24	63.07
BioBERT+TSMD	**75.79**	89.53	**79.88**	**80.69**	51.57	**67.36**
BioBERT+TMLM+TSMD	74.8	**89.56**	79.36	80.45	**55.93**	62.85
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 et al., 2020)	66.40	83.33	68.94	75.33	**56.41**	54.01
Our diseaseBERT	**70.46**	76.77	**74.95**	**79.98**	53.96	**63.05**
DAKI-BERT (Lu et al., 2021)	69.47	**85.06**	70.17	77.95	54.65	58.27
diseaseBioBERT (He et al., 2020)	72.09	**87.78**	74.40	78.43	**54.76**	58.45
Our diseaseBioBERT	**73.26**	87.22	**77.82**	**79.74**	51.44	**62.66**
DAKI-BioBERT (Lu et al., 2021)	72.54	87.33	77.46	78.55	54.17	59.04

5.1 DOWNSTREAM TASKS

The three tasks are disease-related, including CHQA, medical language inference, and disease name recognition. We expect the topic-related task, CHQA, will particularly benefit from our TDA.

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

Medical Language Inference. MEDNLI (Romanov & Shivade, 2018) is a clinical 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 et al., 2014) and BC5CDR (Wei et al., 2016) are the datasets of NER task, they are developed by medical experts annotating diseases mentioned in the

collections of PubMed titles and abstracts. And the task is cast as a classification task to label tokens in sentences with B, I, or O (Peng et al., 2019).

It is notable that the five datasets are small in size (ranging from 1,000 to 10,000 instances), with only hundreds of Dev instances in the QA datasets, and model's performance may vary for the multiple sources of randomness in experiments (Sellam et al., 2022), especially for small datasets like MEDIQA-2019, TRECQA-2017. Following Gu et al. (2021), we report the average scores from ten runs for MEDIQA-2019 and TRECQA-2017 and five runs for other datasets.

5.2 BASELINES

We take BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) the two general domain PLMs, and BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021) the two biomedical domain PLMs as the mian baselines. For fair comparison and carbon reduction, their base models including bert-base-cased, roberta-base, biobert-base-cased-v1.1 and PubMedBERT-base-uncased-abstract-fulltext from HuggingFace Transformers are used in our experiments.

In addition, diseaseBERT series (He et al., 2020), our diseaseBERT series[3] and DAKI-BERT series which encode domain knowledge stored in multiple sources via adapters (Lu et al., 2021) 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 on the in-domain text corpus constructed in §3.1 is included as part of the ablation study.

5.3 IMPLEMENTATION

We initialize a language model with the pre-trained parameters of a baseline model, then adopt each mode of TDA on the PLM befor fine-tuning on the downstream tasks. We directly inherit the hyper-parameters of diseaseBERT (He et al., 2020) except that we set $max\_seq\_length = 512$ for TMLM training procedure, TSMD intermediate fine-tuning, and CHQA task. For PLM+TMLM, our pre-training corpus (52MB) is about 2.5 times bigger than that of He et al. (2020), which would consume longer training time. When the mode is performed on one NVIDIA V100 GPU, it takes about 80 minutes to complete one training epoch, and just 2-5 epochs are enough to enhance PLMs' better performance on the three downstream tasks. For PLM+TSMD, intermediate fine-tuning on TSMD is faster for its smaller size (35MB). It takes no more than 10 epochs to reach its best performance.

5.4 RESULTS

Table 2 shows the performance of consumer health QA tasks. Predictably, the topic-related task benefits a lot from TDA. Our best implementation of TDA – PLM+TSMD increases the accuracy by 5.38% for MEDIQA-2019 and 2.67% for TRECQA-2017 on average. PLM+TSMD surpasses PLM+TMLM+TSMD, which suggests 1) our TSMD constructed with the semi-structured information is similar with CHQA task thereby can help it via transfer learning; 2) intermediate fine-tuning is more effective in capturing task-specific domain knowledge than continual pre-training. Overall, the excellent performance of TDA confirms our predictions that the semi-structured information based method does make PLMs learn more domian knowledge. And the results of diseaseBERT and our diseaseBioBERT suggest longer training sequence suits consumer health QA task.

As is shown in Table 3, TMLM helps PLMs do better on MEDNLI. It rewards the learning of semantic relationship between the paragraph and its topic, which involves logical reasoning ability needed by MEDNLI task. Similarly, a longer training sequence is better for NLI task.

Table 3 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.

Ablation Study. To investigate the effect of different masking strategies, we conduct an ablation study on MEDIQA-2019, the results are shown in Table 4. The main differences between the strate-

[3]The reimplemention of diseaseBERT by setting the maximal sequence length as 512 both for disease knowledge infusion procedure and CHQA task

Table 3: Experimental results on NLI and NER tasks

Tasks	NLI	NER	
Datasets Metrics(%)	MEDNLI Accuracy	BC5CDR F1	NCBI F1
BERT (Devlin et al., 2019)	78.13	83.28	85.56
BERT+TMLM	**79.82**	**84.23**	**86.52**
RoBERTa (Liu et al., 2019)	82.49	83.47	87.01
RoBERTa+TMLM	**83.54**	**83.7**	**87.63**
BioBERT (Lee et al., 2020)	82.77	85.58	87.70
BioBERT+TMLM	**84.04**	**86.13**	**87.91**
PubMedBERT (Gu et al., 2021)	83.76	87.82	88.3
PubMedBERT+TMLM	**84.6**	**87.89**	**88.83**
diseaseBERT (He et al., 2020)	77.29	83.47	**86.81**
Our diseaseBERT	**78.76**	**83.73**	86.64
DAKI-BERT(Lu et al., 2021)	77.85	83.43	85.67
diseaseBioBERT (He et al., 2020)	82.21	86.52	87.14
Our diseaseBioBERT	82.63	**86.57**	87.57
DAKI-BioBERT (Lu et al., 2021)	**83.41**	86.51	**89.01**

gies are the masking rate of heading elements and if randomly masking or not. The results suggest: 1) increasing the masking rate in a certain range is suitable for this task; 2) TMLM (Default) works better than the randomized one (50% random heading masking); 3) For PLMs, TMLM is a more effective masking strategy in capturing domain knowledge than the vanilla MLM.

Table 4: Ablation study on MEDIQA-2019

Metrics(%)	Accuracy	MRR	Precision
Default (selective masking)	**71.91**	**83.56**	**75.84**
75% random heading masking	70.55	83.28	71.65
50% random heading masking	71.18	82.61	73.82
30% random heading masking	70.73	82.94	74.31
15% random heading masking	70.46	82.84	73.33
Vanilla MLM (Devlin et al., 2019)	69.74	79.0	74.23

6 CONCLUSION

In this paper, we show the importance of semi-structured information of documents to enhance the domain knowledge of PLMs. Firstly, we realize the value of the semi-structured information in human learning domain knowledge and design a general pre-training corpus construction method, which could incorporate the semi-structured information well. Secondly, we find the inner link between topic, paragraph, and domain knowledge and propose TDA, which enables PLM to capture the domain knowledge embedded in paragraph and subsection respectively during pre-training and fine-tuning. The experimental results show the effectiveness of TDA on the three biomedical domain tasks, and a significant improvement is observed in the topic-related task, CHQA. The last that must be emphasized is that our TDA is not domain-specific and can be easily applied to various domains, even the general domain.

REFERENCES

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pp. 1–12, 2017.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 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, 2019.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956*, 2016.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23, 2019.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, November 2020. doi: 10.18653/v1/2020.findings-emnlp.261. URL https://aclanthology.org/2020.findings-emnlp.261.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 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, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

Bhuwan Dhingra, Danish Pruthi, and Dheeraj Rajagopal. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 582–587, 2018. URL https://aclanthology.org/N18-2092.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

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

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 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, 2020. URL https://aclanthology.org/2020.emnlp-main.566.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 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, 2020. URL https://aclanthology.org/2020.emnlp-main.372.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of 4th International Conference on Learning Representations*, 2016.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua 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*, pp. 2567–2577, 2019. URL https://aclanthology.org/D19-1259.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, pp. 103982, 2021.

Seongsoon Kim, Donghyeon Park, Yonghwa Choi, Kyubum Lee, Byounggun Kim, Minji Jeon, Jihye Kim, Aik Choon Tan, Jaewoo Kang, et al. A pilot study of biomedical text comprehension using an attention-based deep neural reader: Design and experimental analysis. *JMIR medical informatics*, 6(1):e8751, 2018.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learning of language representations. In *Proceedings of International Conference on Learning Representations*, 2020.

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

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

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 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, 2021.

Dimitris Pappas, Ion Androutsopoulos, and Harris Papageorgiou. Bioread: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. Biomrc: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pp. 140–149, 2020.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 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, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5006. URL https://aclanthology.org/W19-5006.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.134. URL https://aclanthology.org/2020.findings-emnlp.134.

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

Alexey Romanov and Chaitanya Shivade. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1187. URL https://aclanthology.org/D18-1187.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=K0E_F0gFDgA.

Wen Tai, HT Kung, Xin Luna Dong, Marcus Comiter, and Chang-Fu Kuo. 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, 2020.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016, 2016.

Yichong Xu, Xiaodong Liu, Chunyuan Li, Hoifung Poon, and Jianfeng Gao. 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*, 2019. URL https://aclanthology.org/W19-5042.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 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, 2021.

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5461–5468, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.440. URL https://aclanthology.org/2020.emnlp-main.440.