Knowledge-Empowered Representation Learning for Chinese Medical Reading Comprehension: Task, Model and Resources

Taolin Zhang[1,2]*, **Chengyu Wang**[2]*, **Minghui Qiu**[2], **Bite Yang**[3]
Zerui Cai[4], **Xiaofeng He**[4]†, **Jun Huang**[2]
[1] School of Software Engineering, East China Normal University [2] Alibaba Group
[3] DXY [4] School of Computer Science and Technology, East China Normal University
zhangtl0519@gmail.com, zrcai_flow@126.com, yangbt@dxy.cn
{chengyu.wcy, minghui.qmh, huangjun.hj}@alibaba-inc.com
hexf@cs.ecnu.edu.cn

Abstract

Machine Reading Comprehension (MRC) aims to extract answers to questions given a passage, which has been widely studied recently especially in open domains. However, few efforts have been made on closed-domain MRC, mainly due to the lack of large-scale training data. In this paper, we introduce a multi-target MRC task for the medical domain, whose goal is to predict answers to medical questions and the corresponding support sentences from medical information sources simultaneously, in order to ensure the high reliability of medical knowledge serving. A high-quality dataset (more than 18k samples) is manually constructed for the purpose, named Multi-task **C**hinese **Med**ical **MRC** dataset (CMedMRC), with detailed analysis conducted. We further propose a **C**hinese **med**ical **BERT** model for the task (CMedBERT), which fuses medical knowledge into pre-trained language models by the dynamic fusion mechanism of heterogeneous features and the multi-task learning strategy. Experiments show that CMedBERT consistently outperforms strong baselines by fusing context-aware and knowledge-aware token representations.[1]

1 Introduction

Machine Reading Comprehension (MRC) has become a popular task in NLP, aiming to understand a given passage and answer the relevant questions. With the wide availability of MRC datasets (Rajpurkar et al., 2016; He et al., 2018; Cui et al., 2019) and deep learning models (Yu et al., 2018; Ding et al., 2019) (including pre-trained language models such as BERT (Devlin et al., 2019)), significant progress has been made.

Despite the success, a majority of MRC research has focused on open domains. For specific domains, however, the construction of high-quality MRC datasets, together with the design of corresponding models is considerably deficient (Welbl et al., 2017, 2018). The causes behind this phenomenon are threefold. Take the medical domain as an example. i) Data annotators are required to have medical backgrounds with high standards. Hence, simple crowd-sourcing (Rajpurkar et al., 2016; Cui et al., 2019) often leads to poor annotation results. ii) Due to the domain sensitivity, people are more concerned about the reliability of the information sources where the answers are extracted, and the explainability of the answers themselves (Lee et al., 2014; Dalmer, 2017). This is fundamentally different from the task requirements of open-domain MRC. iii) From the perspective of model learning, it is difficult for pre-trained language models to understand the meaning of the questions and passages containing a lot of specialized terms (Chen et al., 2016; Bauer et al., 2018). Without the help of domain knowledge, state-of-the-art models can perform poorly. As shown in Figure 1, BERT (Devlin et al., 2019) and MC-BERT (Zhang et al., 2020) only predict part of the correct answer, i.e., "torso" and "buttocks", instead of generating the complete answer to the medical question.

In this paper, we present a comprehensive study on Chinese medical MRC, including i) how the task is formulated, ii) the construction of the Chinese medical dataset and iii) the MRC model with rich medical knowledge injected. To meet the requirements of medical MRC, we aim to predict both the answer spans to a medical question, and the support sentence from the passage, indicating the source of the answer. The support sentences provide abundant evidence for users to learn medical knowledge, and for medical professionals to assess the trustworthiness of model output results.

*T. Zhang and C. Wang contributed equally to this work.
†Corresponding author.

[1]The code and dataset will be available at https://github.com/MatNLP/CMedMRC

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2237–2249
August 1–6, 2021. ©2021 Association for Computational Linguistics

Multi-task Chinese Medical MRC

Passage: 〈1〉叠瓦癣是一种特殊的体癣, 主要由同心性毛癣菌或称叠瓦癣菌引起…<14>躯干和臀部多见，时久可扩延于四肢，甚至口唇、甲沟及头皮<15>但掌跖多不受累，也不侵犯毛发 (<1>Imbricate tinea is a special kind of body ringworm, mainly caused by the concentric trichoderma or imbricated versicolor bacterium… <14>It often occurs in the torso and buttocks. The disease can spread to the extremities after a long time. It can even spread to the lips, nail groove and scalp. <15>But the metatarsal is not affected and also do not damage the hair...)

Question: 叠瓦癣可发生在什么部位？

(What part of body can imbricate tinea happen?)

Support sentence: 14

BERT_base Answer: 躯干和臀部 (Torso and buttocks)

MC-BERT Answer: 躯干和臀部多见，时久可扩延于四肢 (It often occurs in the torso and buttocks. The disease can spread to the extrmities after a long time.)

CMedBERT Answer: 躯干和臀部多见，时久可扩延于四肢，甚至口唇、甲沟及头皮 (It often occurs in the torso and buttocks. The disease can spread to the extremities after a long time. It can even spread to the lips, nail groove and scalp.)

Medical Entities: 1. 四肢瘫 (quadriplegia) 2. 皮 (skin) 3. 甲沟炎 (paronychia) 4. 甲沟 (nail groove)

Figure 1: Dataset example. MC-BERT and BERT can only predict part of the correct answer. With medical knowledge fused, our CMedBERT model can extract the complete answer. Contents in brackets refer to English translations.

For the dataset, we construct a highly-quality Chinese medical MRC dataset, named the Multi-task Chinese Medical MRC dataset (CMedMRC). It contains 18,153 <*question, passage, answer, support sentence*> quads. Based on the analysis of CMedMRC, we summarize four special challenges for Chinese medical MRC, including *long-tail terminologies, synonym terminology, terminology combination* and *paraphrasing*. In addition, we find that comprehensive skills are required for MRC models to answer medical questions correctly. For answer extraction in CMedMRC, direct *token matching* is required for answering 31% of the questions, *co-reference resolution* for 11%, *multi-sentence reasoning* for 18% and *implicit causality* for 22%. In addition, the answers to the remaining questions (16%) are extremely difficult to extract without rich medical background knowledge.

To address the medical MRC task, we propose the multi-task dynamic heterogeneous fusion network (CMedBERT) based on the MC-BERT (Zhang et al., 2020) model and a Chinese medical knowledge base (see Appendix). The technical contributions of CMedBERT are twofold:

- **Heterogeneous Feature Fusion**: We mimic humans' approach of reading comprehension (Wang et al., 1999) by learning attentively aggregated representations of multiple entities in the passage. Different from the knowledge fusion method used by KBLSTM (Yang and Mitchell, 2017) and KT-NET (Yang et al., 2019), we propose a two-level attention and a gated-loop mechanism to replace the knowledge sentinel, so that rich knowledge representations can be better integrated into the model.
- **Multi-task Learning**: Parameters of CMedBERT are dynamically learned by capturing the relationships between the two tasks via multi-task learning. We regard the semantic similarities between support sentences and answers to questions as the task similarities.

We compare CMedBERT against six strong baselines. For answer prediction, compared to the strongest competitor, the EM (Exact Match) and F1 scores are increased by +3.88% and +1.46%, respectively. Meanwhile, the support sentence prediction task result is increased by a large margin, i.e., +7.81% of EM and +4.07% of F1.

2 Related Work

MRC Datasets and Models. Due to the popularity of the MRC task, there exist many types of MRC datasets, such as span-extraction (Rajpurkar et al., 2016; Yang et al., 2018), multiple choices (Richardson et al., 2013; Lai et al., 2017), cloze-style (Hermann et al., 2015), cross-lingual (Jing et al., 2019; Yuan et al., 2020). For specific domains, however, the number of publicly available MRC datasets remains few, including SciQ (Welbl et al., 2017), Quasar-S (Dhingra et al., 2017) and Biology (Berant et al., 2014). CLiCR (Suster and Daelemans, 2018) is a cloze-style single-task English medical MRC dataset. However, it contains a relatively small variety of medical questions, automatically generated from clinical case reports. Recently, (Li et al., 2020) propose a multi-choice Chinese medical QA dataset, retrieving text snippets as the passage and the task only chooses an existing correct option from candidate set. Our work specifically focuses on the fine-grained medical MRC tasks and deep domain knowledge reasoning, with a manually constructed high-quality dataset released.

The model architecture of MRC mostly takes advantage of neural networks to learn token representations of passages and questions jointly (Qiu et al., 2019a; Liu et al., 2019). The interaction between questions and passages is modeled based on attention mechanisms. The rapid development of

Challenges	Characteristics	Example
Lexical-Level	***Long-tail terminology***	冈上肌肌腱断裂试验是对冈上肌肌腱是否存在断裂进行检查。(*supraspinatus tendon* rupture test is to check whether the supraspinatus tendon is ruptured.)
	Synonym terminology	...本药品对过敏性鼻炎和上呼吸道感染引起的鼻充血有效，可用于感冒或鼻窦炎... (...This medicine is effective for nasal congestion caused by allergic rhinitis and *upper respiratory tract infection*, and can be used for *colds* or sinusitis...)
	Terminology combination	...糖尿病性视网膜病(diabetic retinopathy)是糖尿病性微血管病变中最重要的表现... (...*Diabetic retinopathy (DR)* is the most important manifestation of diabetic microangiopathy...)
Sentence-Level	***Paraphrasing***	***Passage***:...如果在嘴角烂了或结痂的地方进行冷敷，一方面冷敷物品不干净的话会造成感染；另一方面局部温度降低了之后，反而会延缓伤口的愈合。...(...If you apply a cold compress on a rotten or crusted corner of the mouth, on the one hand, if the cold compress is not clean, it will *cause infection*; on the other hand, when the local temperature is lowered, it will *delay the healing of the wound*...) ***Question***:为什么嘴角烂了或结痂不建议进行冷敷？(Why is it *not recommended* to apply cold compresses when the corners of the mouth are rotten or crusted?)

Table 1: Two levels of challenges in processing Chinese medical texts. The blue and underscore contents in brackets indicate why this example belongs to its corresponding "Characteristics" category. (Best viewed in color.)

deep learning leads to a variety of models, such as the QANet (Yu et al., 2018), SAN (Liu et al., 2018). Graph neural networks have been used in MRC recently by modeling the relations between entities in the passage (Ding et al., 2019) and multi-grained tokens representation (Zheng et al., 2020).

Pre-trained Language Models and Knowledge Fusion. Pre-trained language models (e.g., BERT (Devlin et al., 2019), ERNIE-THU (Zhang et al., 2019), K-BERT (Liu et al., 2020a)) have successfully improved the performance of the MRC task, which even exceed the human level in some datasets. This is because these models obtain better token representations and capture lexical and syntactic knowledge in different layers (Guan et al., 2019). For specific domain, there also have some pre-trained models (Beltagy et al., 2019; Zhang et al., 2020).

A potential drawback is that pre-trained language models of open domains only learn general representations, lacking domain-specific knowledge to deepen the understanding of entities and other nouns (Ostendorff et al., 2019) (which are often the answers in span-extraction MRC tasks). Without proper descriptions of such entities in the passage, MRC models often fail to understand and extract key information (Das et al., 2019). Hence, the explicit fusion of knowledge in MRC models is vital for learning context-aware token representations (Pan et al., 2019; Qiu et al., 2019b; Liu et al., 2020b). Instead of encoding entities appearing in both knowledge bases and passages into the MRC model only (Chen et al., 2018), our proposed model encodes all the triples from a medical KG and then employs heuristic rules to retrieve relevant entities. This practice allows the model to acquire deeper understanding of domain-specific terms.

3 The CMedMRC Dataset

In this section, we briefly describe the collection process and provide an analysis on various aspects of the CMedMRC dataset. For more dataset collection and statistical analysis of dataset details, we refer readers to the Appendix A and Appendix B.

3.1 Dataset Collection Process

The dataset collection process follows the SQuAD-style (Rajpurkar et al., 2016) rather than collecting question-answer pairs as in Google Natural Questions (Kwiatkowski et al., 2019). Our medical text corpus is collected from DXY Medical [2], an authoritative medical knowledge source in China. The general data collection process of CMedMRC consists of four major steps: passage collection, question-answer pair collection, support sentence selection and additional answer construction. Briefly speaking, during the passage collection process, we filter the corpus to generate high-quality medical passages. A group of human annotators are required to ask questions on medical knowledge and annotate the answers from these passages. The annotation results are in the form of question-answer pairs. Following SQuAD, we ask annotators to provide 2 additional answers for each question in the DEV and TEST sets.

Since people are concerned about the scientific explanation and sources of answers in the medical domain, we ask annotators to select the support sentence of their annotated answer similar to those of CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018). Finally, CMedMRC consists of three

[2] http://www.dxy.cn/

Skills	Example	Percentage
Token matching	***Passage***: ...急性羊水过多较少，见多发生在孕20～24周，羊水急剧增多，子宫短期内明显增大... (...it is less likely to secrete too much acute *amniotic fluid*. The disease is most common in the 20 to 24 weeks of *pregnancy*. The amniotic fluid increases sharply with the uterus enlarged significantly in the short term...) ***Question***: 怀孕期间羊水什么时候分泌过多? (When does the *amniotic fluid* secrete too much during *pregnancy*?) ***Answer***: 20～24周 (20～24 weeks)	31%
Co-reference resolution	***Passage***: ...尖锐湿疣有「割韭菜」的臭名声，它的治疗瓶颈在于病毒不进入血循环，因此机体无法产生免疫应答，所以容易反复复发。...(...*genital warts* has a bad reputation of cutting leeks. The bottleneck of *its* treatment is that the virus does not enter the blood circulation, so the body cannot produce an immune response and *it* is easy to relapse repeatedly....) ***Question***: 为什么尖锐湿疣易反复 发作? (Why *genital warts* is easy to relapse repeatedly?) ***Answer***: 病毒不进入血循环，因此机体无法产生免疫应答 (The virus does not enter the blood circulation, so the body cannot produce an immune response)	11%
Multi-sentence reasoning	***Passage***: ... 老年人应在医师指导下使用。5.肝、肾功能不全者慎用。6.孕妇及哺乳期妇女慎用。... (...*The elderly should take the medicine under the guidance of a physician. 5. Use with caution in patients with liver and kidney insufficiency. 6. Use with caution in pregnant and lactating women.*) ***Question***: 哪些人群慎用此药品? (Which groups of people should use this drug with caution?) ***Answer***:老年人应在医师指导下使用。5.肝、肾功能不全者慎用。6.孕妇及哺乳期妇女慎用 (The elderly should take the medicine under the guidance of a physician. 5. Use with caution in patients with liver and kidney insufficiency. 6. Use with caution in pregnant and lactating women.)	18%
Implicit causality	***Passage***: ... 不是所有的白细胞减少都必须治疗的， 关键看白细胞减少的程度、机体的一般状态以及医生的建议；...... 因为无症状的白细胞减少对生活的影响是很小的； ... (...*Not all leukopenia must be treated.* The key depends on the degree of leukopenia, the general state of the body and the doctor's advice;......*Because asymptomatic leukopenia has little impact on life*...) ***Question***: 为什么不是所有的白细胞减少都要进行治疗? (Why do *not all leukopenia have to be treated?*) ***Answer***:因为无症状的白细胞减少对生活的影响是很小的 (Because asymptomatic leukopenia has little impact on life.)	22%
Domain knowledge	***Passage***:...发病率居遗传性血小板功能缺陷疾病的首位。血栓细胞衰弱发病多见于幼年，发病率为 0.01/万... (...The incidence is the highest in hereditary platelet dysfunction diseases. The incidence of *thrombotic cell weakness* is more common in childhood with an incidence rate of 0.01 / 10,000...) ***Question***:血小板无力症的发病率约为多少? (What is the incidence rate of *blood platelet weakness?*) ***Answer***: 0.01/万 (0.01/10,000)	16%

Table 2: Reading comprehension skills of models required to answer questions in CMedMRC. The blue and underscore contents in brackets indicate why the sample belongs to its category. (Best viewed in color)

parts: 12,700 training samples, 3,630 development samples and 1,823 testing samples.

3.2 Quality Control

During the dataset collection process, we take the following measures to ensure the quality of the dataset. i) The knowledge source (DXY Medical) contains high-quality medical articles which are written by medical personnel and organized based on different topics in the medical domain. ii) Our annotators are all engaged in medical-related professions rather than annotators with short-term guidance only. iii) We further hire 12 medical experts to check all the collected samples rather than checking a randomly selected sample only. The experts remove out-of-domain questions and questions that are unhelpful to medical practice. In this stage, the experts are divided into two groups and cross-check their judgments.

3.3 Challenges of Understanding Texts

Due to the closed-domain property of our dataset, there are some domain-specific textual features in both passages and questions that the model needs to understand. Based on our observations of the CMedMRC, we summarize the following two major challenges. These challenges can be also regarded as key reasons why some recent state-of-the-art MRC models cannot address the medical MRC task on CMedMRC well.

Lexical-Level: i) Long-tail terminology means these medical terms occur very infrequently and are prone to Out-Of-Vocabulary (OOV) problems. ii) Synonym terminology means that some medical terms may express the same meaning, but there is a distinction between colloquial expressions and professional terms. The above two points require the model to have rich domain knowledge to solve. iii) Terminology combination means these terms are usually formed by a combination of multiple terms, while one term is the attributive of another. This does not only require the model to have domain knowledge but also poses challenges to phrase segmentation in specific domains.

Sentence-Level: Paraphrasing means some words in questions are semantically related to certain tokens in passages, but are expressed differently. Consider the last question in Table 1. When the model tries to answer the "not-recommended" question, it should focus on negative terms ("cause infection" and "delay the healing of the wound").

3.4 Reasoning Skills for MRC Models

We randomly select 100 samples from the development set to analyze what skills the model should have in order to answer the questions correctly. We divide the reasoning skills corresponding to these samples into five major categories, namely token matching, co-reference resolution, multi-sentence reasoning, implicit causality and domain knowledge. Examples are shown in Table 2. It is particularly noteworthy that the fifth type is the need of domain knowledge to answer medical questions. Consider the example:

Passage: ... The incidence of thrombotic cell weakness is most common in childhood with an incidence rate of 0.01 / 10,000...

Question: What is the incidence rate of blood platelet weakness?

Answer: 0.01/10,000.

We know that the *blood platelet* in the question refers to the *thrombotic cell* described in the passage through the medical knowledge base. It shows that the rich information of the knowledge base can help the model obtain a better understanding of domain terms to improve the MRC performance.

4 The CMedBERT Model

4.1 Task Formulation and Model Overview

For our task, the input includes a medical question Q together with the passage P. Let $\{p_1, p_2, \cdots p_m\}$ and $\{q_1, q_2, \cdots q_n\}$ represent the passage and question tokens, respectively. In the answer prediction task, the goal is to train an MRC model which extracts the answer span $\{p_i, p_{i+1}, \cdots p_j\}$ $(0 \le i \le j \le m)$ from P that correctly answers the question Q. Additionally, the model is required to predict the support sentence tokens $\{p_k, p_{k+1}, \cdots p_l\}$ $(0 \le k \le l \le m)$ from P to provide additional medical knowledge and to enhance interpretability of the extracted answers. We constrain that $\{p_k, p_{k+1}, \cdots p_l\}$ must form a complete sentence, instead of incomplete semantic units and the support sentence tokens contain the answer span. The high-level architecture of the CMedBERT model is shown in Figure 2. It mainly includes four modules: BERT encoding, knowledge embedding and retrieval, heterogeneous feature fusion and multi-task training.

4.2 BERT Encoding

This module is used to learn context-aware representations of question and the passage tokens. For each input pair (the question Q and the passage P), we treat $[\langle CLS\rangle, Q, \langle SEP\rangle, P, \langle SEP\rangle]$ as the input sequences for BERT. We denote $\{h_i\}_{i=1}^{m+n+3}$ as the hidden layer representations of tokens, where m and n are the length of passage tokens and question tokens, respectively.

4.3 Knowledge Embedding and Retrieval

In the knowledge bases, relational knowledge is stored in the form of (*subject, relation, object*) triples. In order to fuse knowledge into token representations, we first encode all entities in the knowledge base into a low-dimensional vector space. Here, we employ PTransE (Lin et al., 2015a) to learn entity representations, and denote the underlying entity embedding as e_i. Because existing medical NER tools do not have high coverage over our corpus, we consider five types of token strings as candidate entities: noun, time, location, direction and numeric. Two matching strategies are then employed to retrieve relevant entities from the knowledge base: (i) The two strings match exactly. (ii) The number of overlapped tokens is larger than a threshold. After relevant entities are retrieved, we can fuse the knowledge into contextual representations, introduced below.

4.4 Heterogeneous Feature Fusion

In this module, we fuse heterogeneous entity features retrieved from the knowledge base into the question and passage tokens representations.

Local Fusion Attention. We observe that each token is usually related to multiple entities of varying importance. Thus, we assign different weights to the entity embedding e_j corresponding to the token representation h_i using attention mechanism:

$$\alpha_{i,j} = \frac{exp(e_j^T W h_i)}{\sum_{k=1}^{K} exp(e_k^T W h_i)} \qquad (1)$$

where K is the number of entities and $\alpha_{i,j}$ represents the similarity between the j^{th} entity in the retrieved entity set and the i^{th} token. $W \in \mathbb{R}^{d_2 \times d_1}$ where d_1 is the dimension of BERT's output and d_2 is the dimension of entity embeddings. After fusing, the representation of the i^{th} token is: $\bar{e_i} = \sum_{k=1}^{K} \alpha_{i,k} e_k$. However, $\bar{e_i}$ is only related to retrieved entities, not other tokens in the question-passage pair.

Global Fusion Attention. In BERT, the output of the $[CLS]$ tag represents the entire sequence information learned by transformer encoders. We use

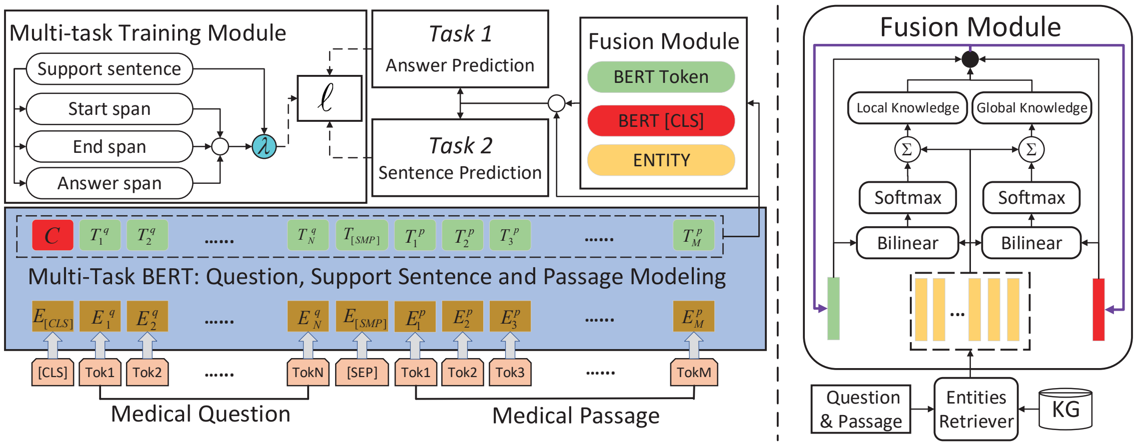


Figure 2: Model overview. The green box and the red box in the heterogeneous feature fusion layer represent the local and global token information, respectively. In the multi-task training module, the model learns the relationship between two tasks by dynamically learning the parameter λ. (Best viewed in color)

the token output $h_{[CLS]}$ to model the knowledge fusion representation of the entire entity collection that each token recalls:

$$\beta_{[CLS],j} = \frac{exp(e_j^T W h_{[CLS]})}{\sum_{k=1}^{K} exp(e_k^T W h_{[CLS]})} \tag{2}$$

$$\hat{e}_i = \sum_{k=1}^{K} \beta_{[CLS],k} e_k \tag{3}$$

where $\hat{e}_i$ is the global knowledge fusion result corresponding to the i^{th} token.

Gated Loop Layer. In order to fuse local and global results into token representations, we design a gated loop layer. The information of knowledge fusion is filtered through the gating mechanism in each loop of modeling. In the initialization stage, we simply have $h_i^0 = h_i$. In the l^{th} iteration, we have the following update process:

$$G_i^\ell = \sigma(tanh(W[h_i^\ell, \bar{e}_i, \hat{e}_i])) \tag{4}$$

$$h_i^{\ell+1} = G_i^\ell \odot h_i^\ell \tag{5}$$

This process runs for L loops and this fusion process output is h_i^L. The loop process mimics the human's behavior of reading the passage repeatedly to find the most accurate answers.

4.5 Multi-task Training

The output layer of CMedBERT is extended from BERT. We first concatenate two types of token representations and calculate the probability of the i^{th} token being selected in the support sentence as follows:

$$o_i = \sigma(W[h_i, h_i^L]),\ p_i^{support} = \sigma(W o_i) \tag{6}$$

We also calculate its probabilities as the starting and the ending positions of the answer span, respectively:

$$p_i^{start} = \frac{exp(w_1^T o_i)}{\sum_j exp(w_1^T o_j)},\ p_i^{end} = \frac{exp(w_2^T o_i)}{\sum_j exp(w_2^T o_j)}$$

The loss function of the answer prediction task is the negative log-likelihood of the starting and ending positions of ground-truth answer tokens:

$$\mathcal{L}_{\mathcal{A}} = -\frac{1}{N}\sum_{j=1}^{N}(logp_{y_j^{start}}^{start} + logp_{y_j^{end}}^{end}) \tag{7}$$

For the extraction of support sentences, the loss function is defined by cross-entropy:

$$\mathcal{L}_{\mathcal{S}} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M}(y_j^{support} logp_i^{support}) \tag{8}$$

where N is the number of samples and M is the length of input sequences. y_j^{start},y_j^{end} is the starting and ending positions of ground-truth of the j^{th} token. Furthermore, if the token is in the support sentence, the token label $y_j^{support}$ is set to 1, and 0 otherwise.

The representations of the support sentence are related to the positions of the answer. In order to better model the relationship between two tasks, we dynamically learn the coefficient between the loss values of two tasks. Let h_{su} and o_{sp} be self-attended, averaged pooled representations of the support sentence and the answer span. o_{st}, o_{ed} are the start and end position token representations of

Model	Answer				Support Sentence			
	Exact Match (EM)		F1		Exact Match (EM)		F1	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
DrQA	42.00%	37.45%	58.66%	57.15%	5.07%	5.88%	30.24%	32.52%
BERT_base	64.83%	68.31%	81.08%	83.74%	21.42%	17.70%	52.27%	48.34%
ERNIE	65.49%	68.57%	81.17%	83.86%	20.71%	16.32%	49.88%	45.81%
KT-NET♠	64.58%	69.03%	81.06%	84.18%	15.42%	13.48%	49.37%	46.45%
MC-BERT	66.58%	68.62%	81.23%	83.98%	20.08%	16.77%	47.33%	44.53%
KMQA	66.19%	68.45%	81.20%	83.79%	20.63%	16.54%	49.27%	45.97%
CMedBERT♣	69.00%	72.84%	82.68%	85.38%	25.17%	24.18%	52.36%	49.69%
CMedBERT♠	**70.33%**	**72.91%**	**83.43%**	**85.64%**	**25.58%**	**25.51%**	**52.67%**	**52.41%**

Table 3: The results of multi-task prediction (answer and support sentence) over CMedMRC.

the answer, respectively. We have:

$$\gamma_{st}, \gamma_{ed}, \gamma_{sp} = h_{su}[o_{st}, o_{ed}, o_{sp}]^T \quad (9)$$

$$H_A = \sigma(W[\gamma_{st} o_{st}, \gamma_{ed} o_{ed}, \gamma_{sp} o_{sp}]) \quad (10)$$

where $\gamma_{st}, \gamma_{ed}, \gamma_{sp}$ are the weight coefficients between the supporting sentence and the start/end/total token representations of the answer span. The loss value coefficient of two tasks λ and the total loss $\mathcal{L}$ are as follows:

$$\lambda = max\{0, \cos(H_A, h_{su})\} \quad (11)$$

$$\mathcal{L} = \mathcal{L}_A + \lambda \mathcal{L}_S \quad (12)$$

We minimize the total loss $\mathcal{L}$ to update our model parameters in the training process.

5 Experiments and Result Analysis

5.1 Experimental Setups

We evaluate CMedBERT on CMedMRC, and compare it against six strong baselines: DrQA (Chen et al., 2017), BERT_base (Devlin et al., 2019), ERNIE (Zhang et al., 2019), KT-NET (Yang et al., 2019), MC-BERT (Zhang et al., 2020) and KMQA (Li et al., 2020). KT-NET is the first model to leverage rich knowledge to enhance pre-trained language models for MRC. MC-BERT is the first Chinese biomedical pre-trained model fine-tuned on BERT_base. We only use the encoder layer in KMQA removing the answer layer due to the different answer type.

For evaluation, we use EM (Exact Match) and F1 metrics for answer and support sentence tasks. We calculate character-level overlaps between prediction and ground truth for the Chinese language, rather than token-level overlaps for English. To assess the difficulty of solving CMedMRC tasks, we select 100 testing samples to evaluate human performance. Human scores of EM and F1 are 85.00% and 96.69% for answer prediction, respectively.

Model	Exact Match (EM)		F1	
	Dev	Test	Dev	Test
DrQA	34.50%	32.10%	56.67%	56.64%
BERT_base	62.39%	68.29%	81.48%	83.70%
ERNIE	63.18%	66.92%	81.74%	83.41%
KT-NET♠	64.64%	66.26%	82.48%	83.61%
MC-BERT	63.39%	68.38%	81.86%	83.88%
KMQA	64.37%	67.48%	81.95%	83.74%
CMedBERT♣	68.00%	72.11%	82.50%	85.33%
CMedBERT♠	**69.83%**	**72.84%**	**83.02%**	**85.54%**
Human	-	85.00%	-	96.69%

Table 4: Result of single-task (answer prediction). ♣ and ♠ indicate that CMedBERT uses BERT_base and MC-BERT as the encoder, respectively.

In the implementation, we set the learning rate as 5e-5 and the batch size as 16, and the max sequence length as 512. Other BERT's hyper-parameters are the same as in Google's settings [3]. Each model is trained for 2 epochs by the Adam optimizer (Kingma and Ba, 2015). Results are presented in average with 5 random runs with different random seeds. Other implementation details are in Appendix C.

5.2 Model Results

Table 3 and Table 4 show the multi-task and single-task results on the CMedMRC development and testing sets. CMedBERT has a great improvement compared to four strong baseline models in both tasks. Specifically, our CMedBERT outperforms the state-of-the-art model by a large margin in multi-task results, with **+3.88%EM / +1.46%F1**

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

Model	Answer		Sentence	
	EM	F1	EM	F1
CMedBERT♠	**72.91%**	**85.64%**	**25.51%**	**52.41%**
w/o Local Att.	68.93%	83.89%	19.75%	49.45%
w/o Global Att.	71.71%	84.96%	17.59%	47.01%
w/o λ	71.91%	85.09%	21.09%	48.80%

Table 5: Ablation study of CMedBERT (testing set).

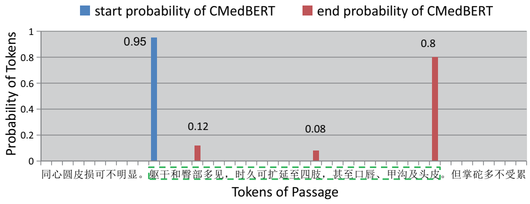
improvements, which shows the effectiveness of our model. Meanwhile, in the support sentence task, our model also has the best performance, improving (**+7.81%EM / +4.07%F1**) over the testing set. In single task evaluation, we remove the support sentence training module and the dynamic parameter for loss function module. Our model improves (**+4.46%EM / +1.66%F1**) over the best baseline model. In addition, we find that using support sentence prediction as an auxiliary task and the pre-training technique in medical domain can further improve the performance of CMedBERT.

5.3 Ablation study

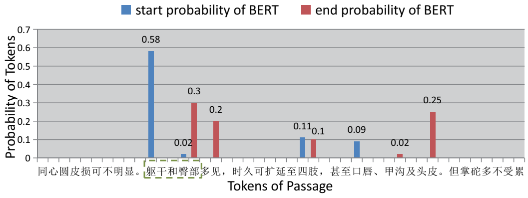
In Table 5, we choose three important model components for our ablation study and report the results over the testing set. When the dynamic parameter λ of the loss function is removed from the model, the performance of the model on two tasks is decreased by (**-1.00%EM and -0.55%F1**) and (**-4.42%EM and -3.61%F1**), respectively. Without local attention, the EM performance in the answer prediction task decreases by (**-3.98%EM and -1.75%F1**). Experiments have shown that the model performs worse without the local fusion attention than without the global fusion attention and the dynamic parameter λ. However, the performance of support sentence task is decreased significantly by (**-7.92%EM and -3.61%F1**) without global fusion attention. It shows that local fusion attention is more important for extracting answer spans, while global fusion attention plays a larger role in support sentence prediction.

5.4 Case Study

In Figure 3, we use our motivation example to conduct a case study. In BERT, we can see that the difference among the probability values of different words is small, especially when predicting the probability of ending positions. The ending position probabilities of token “部” token “皮” are **0.3** and **0.25**, leading the model to extract the wrong



(a) Tokens probabilities of **CMedBERT♠**



(b) Tokens probabilities of **BERT_base**

Figure 3: Case study. Predicted answer spans are in the green dotted box. Product of the maximum starting and ending probabilities of CMedBERT is **0.76**, with BERT to be **0.174**.

answer span. However, in the knowledge retrieval module of CMedBERT, multiple entities representation are fused into the context-aware latent space representation to enhance the medical text semantic understanding. Therefore, in our CMedBERT model, the starting position probability is **0.95** and the end position probability is **0.8**. In this case, the CMedBERT model can easily choose the correct range of the answer span.

5.5 Discussion of Support Sentence Task

Compared with the answer prediction task, existing models have poor prediction results on the EM metric in the support-sentence task. In prediction results, we randomly select 100 samples for analysis. We divide the error types into the following three main types (see Appendix): i) starting position cross ii) ending position cross iii) answer substring. The most common error type is the answer substring, accounting for **46%**. In this error type, the predicted result of our model is part of the true result, which shows the model cannot predict long answers completely (Yuan et al., 2020) and reduce the accuracy of the results greatly.

6 Conclusion

In this work, we address medical MRC with a new dataset **CMedMRC** constructed. An in-depth analysis of the dataset is conducted, including statistics, characteristics, required MRC skills, etc. Moreover, we propose the **CMedBERT** model, which can help the pre-trained model better understand domain terms by retrieving entities from medical knowledge bases. Experimental results confirm the effectiveness of our model. In the future, we will explore how knowledge can improve the performance of models in the medical domain.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments. This work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904, and Alibaba Group through Alibaba Research Intern Program.

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *EMNLP*, pages 4220–4230.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*, pages 3613–3618.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *EMNLP*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *ACL*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *ACL*, pages 1870–1879.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *ACL*, pages 2406–2417.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *EMNLP*, pages 2174–2184.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *EMNLP*, pages 5882–5888.

Nicole K Dalmer. 2017. Questioning reliability assessments of health information on social media. *Journal of the Medical Library Association: JMLA*, 105(1):61.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. 2019. Building dynamic knowledge graphs from text using machine reading comprehension. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. 2017. Quasar: Datasets for question answering by search and reading. *CoRR*, abs/1707.03904.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *ACL*, pages 2694–2703.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in NLP. In *ICML*, pages 2454–2463.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *ACL*, pages 37–46.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NeurIPS*, pages 1693–1701.

Yimin Jing, Deyi Xiong, and Yan Zhen. 2019. Bipar: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. In *EMNLP*, pages 2452–2462.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*, pages 785–794.

Kenneth Lee, Kreshnik Hoti, Jeffery D Hughes, and Lynne M Emmerton. 2014. Interventions to assist health consumers to find reliable online health information: a comprehensive review. *PloS one*, 9(4):e94186.

Dongfang Li, Baotian Hu, Qingcai Chen, Weihua Peng, and Anqi Wang. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *EMNLP*, pages 1427–1438.

Yankai Lin, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015a. Modeling relation paths for representation learning of knowledge bases. In *EMNLP*, pages 705–714.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015b. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *CoRR*, abs/1907.01118.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020a. K-BERT: enabling language representation with knowledge graph. In *AAAI*, pages 2901–2908.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *ACL*, pages 1694–1704.

Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S. Yu. 2020b. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. *CoRR*, abs/2008.02434.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julián Moreno Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching BERT with knowledge graph embeddings for document classification. In *KONVENS*.

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie, and Dong Yu. 2019. Improving question answering with external knowledge. In *EMNLP*, pages 27–37.

Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019a. A survey on neural machine reading comprehension. *CoRR*, abs/1906.03824.

Delai Qiu, Yuanzhe Zhang, Xinwei Feng, Xiangwen Liao, Wenbin Jiang, Yajuan Lyu, Kang Liu, and Jun Zhao. 2019b. Machine reading comprehension using structural knowledge graph-aware network. In *EMNLP*, pages 5895–5900.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, pages 193–203.

Simon Suster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. In *NAACL-HLT*, pages 1551–1563.

Jian Wang, Hsuan-Chih Chen, Ralph Radach, and Albrecht Inhoff. 1999. *Reading Chinese script: A cognitive analysis*. Psychology Press.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *EMNLP*, pages 94–106.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *TACL*, 6:287–302.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *ACL*, pages 2346–2357.

Bishan Yang and Tom M. Mitchell. 2017. Leveraging knowledge bases in lstms for improving machine reading. In *ACL*, pages 1436–1446.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pages 2369–2380.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*.

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang. 2020. Enhancing answer boundary detection for multilingual machine reading comprehension. In *ACL*, pages 925–934.

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

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *ACL*, pages 1441–1451.

Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020. Document modeling with graph attention networks for multi-grained machine reading comprehension. In *ACL*, pages 6708–6718.

A Detailed Dataset Construction Process of CMedMRC

A.1 Medical Passage Curation

We use the following rules to obtain 20,000 passages from DXY as the inputs to human annotators:

- We use regular expressions to filter out images, tables, hyperlinks, etc. The English-Chinese translations of medical terms are also provided if the passages contain medical terms in English.
- We find that if we follow (Rajpurkar et al., 2016) to limit the lengths of the passages within 500 tokens, our human annotators could not ask 4 high-quality medical questions easily. Hence, our passage length limit is 1000 tokens.

A.2 Medical Question-Answer Pair Collection

We employ a group of annotators with professional medical background to generate question-answer pairs from the medical passages. Here are some general guidelines:

- We encourage our annotators to ask questions to which the answers are uniformly distributed in different positions of medical passages.
- For each medical passage, we limit the number of questions to 4.
- Each question should be strictly related to the medical domain. When creating the questions, no any part of the texts can be directly copied and pasted from the given medical passages.
- We limit the number of answer tokens to no more than 40.

A.3 Support Sentence Selection

In our dataset, we add an index to each sentence in the passages. Annotators are required to select the support sentence index and mark the range of the answer spans on the user interface.

A.4 Additional Answer Construction

To evaluate the human performance of our dataset and make our model more robust, we collect two additional answers for each question in the development and testing sets. We employ another 12 annotators for answer construction. Since our medical passage is relatively long, we show the questions and the passage contents again on the interface, together with the previously labeled support sentence indices.

Answer Type	Pct.	Example
Numeric	6%	20%
Time/Date	11%	1-3小时 (1-3 Hours)
Person	8%	儿童 (Child)
Location	5%	安徽, 云南, 湖北 (Anhui, Yunnan, Hubei)
Noun Phrase	18%	输卵管炎 (Salpingitis)
Verb Phrase	6%	清洗，干燥和粉碎 (Wash, dry and crush)
Yes/No	1%	不会感染 (Will not infect)
Description	44%	维生素缺乏 (Vitamin deficiency)
Other	1%	严重 (Severe)

Table 6: Statistical results for answer types.

	Train	Dev	Test
# Questions	12,700	3,630	1,823
Avg. tokens of passages	883.64	743.10	745.52
Avg. tokens of questions	15.40	14.85	15.23
Avg. tokens of answers	19.69	18.48	16.57
Avg. tokens of support sen.	57.50	48.19	42.70

Table 7: Statistical results of text length in our CMedMRC dataset.

B Statistical Analysis of the CMedMRC Dataset

B.1 Question and Answer Types

Due to the special characteristics of the Chinese language, the question types cannot be simply classified by prefix words of questions (Rajpurkar et al., 2016). Here, we manually define 8 common question types in the user annotation interface. The statistics of each question type are shown in Figure 4. The first seven question types usually correspond to special medical answers. For example, the *What* type refers to a question on the name of a drug or a disease, which accounts for more than half of the dataset. A third of the questions belong to the types of *How* and *Why*. The statistics of answer types are also shown in Table 6. The proportions of *Noun Phrase* and *Description* types

Error Type	Example	Percentage
Start position cross	***Ground-truth***: 当有急性炎症或者化脓时，会有剧烈疼痛；或者合并牙神经发炎时也会出现剧烈疼痛。 (*When there is acute inflammation or suppuration, there will be severe pain;* or when combined with dental nerve inflammation, there will also be severe pain) ***Prediction***: 有轻微的隐痛或胀痛；当有急性炎症或者化脓时，会有剧烈疼痛； (There is slight dull pain or pain; *when there is acute inflammation or suppuration, there will be severe pain;*)	25%
End position cross	***Ground-truth***: 以下人群高危：乙肝、丙肝病毒慢性感染者；患有类风湿关节、狼疮、硬皮病等免疫性疾病；吸烟。......遗传因素对本病起到一定作用。 (The following people are at high risk: people with chronic hepatitis B and C virus infections; suffering from immune diseases such as rheumatoid joints, lupus, and scleroderma; smoking. ...... *Genetic factors play a role in this disease.*) ***Prediction***: 目前认为血管炎是一种自身免疫性疾病，...... 遗传因素对本病起到一定作用。一些药物也可以引起血管炎，还有一些感染（如丙肝、乙肝、梅毒）也可以引起血管炎。 (*At present, vasculitis is considered to be an autoimmune disease, ... Genetic factors play a role in this disease.* Some drugs can also cause vasculitis, and some infections (such as hepatitis C, hepatitis B, and syphilis) can also cause vasculitis.)	21%
Answer substring	***Ground-truth***:这些药物具有抗炎、改善毛细血管通透性、减轻水肿、止痛等作用，同时对日光性皮炎有很好的治疗作用。 (These drugs have *anti-inflammatory, improve capillary permeability, reduce edema, pain relief, etc.*, and have a good therapeutic effect on solar dermatitis.) ***Prediction***:具有抗炎、改善毛细血管通透性、减轻水肿、止痛等作用 (*Anti-inflammatory, improve capillary permeability, reduce edema, relieve pain, etc.*)	46%
Other	***Ground-truth***:抗组胺药第一代的经典代表药「马来酸氯苯那敏」就是一个，它俗称扑尔敏，在多年临床应用中没有发现对胎儿有明显的致畸或其他严重危害。 (One of the classic representative drugs of the first generation of antihistamines is ”Chlorpheniramine Maleate”. It is commonly known as Chlorpheniramine. It has not been found to have obvious teratogenic or other serious harm to the fetus in many years of clinical application.) ***Prediction***:但临床上也有一些药物是经过多年验证，只要注意把握用药时间和药量，即使让孕妇吃也不会有事的 (However, there are also some drugs that have been verified for many years in clinical practice. As long as you pay attention to the time and amount of medication, it will be fine even if pregnant women take it.)	8%

Table 8: Three typical error answer types in support sentence task. The blue and underscore contents in brackets indicate why the sample belongs to its corresponding category.

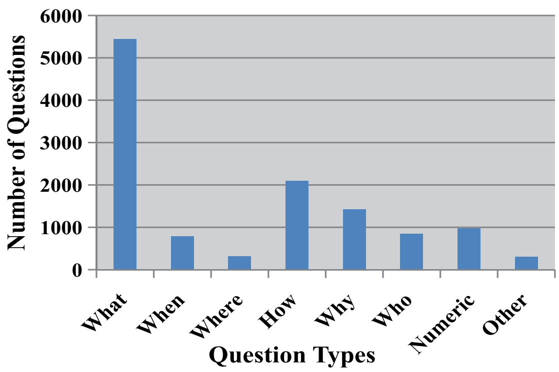


Figure 4: The number of questions that belong to each question types in CMedMRC.

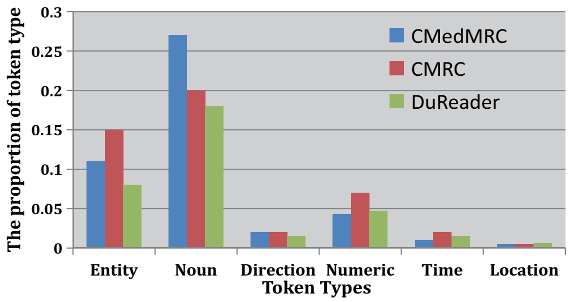


Figure 5: Proportions of entities and five frequently appearing POS tags in three Chinese MRC datasets.

are relatively large. The results are consistent with Figure 4, since most of *What* questions need to be answered with the above two answer types. Table 7 shows the text length of four input data.

B.2 Analysis of Domain Knowledge

We further analyze to what degree there exists domain knowledge in CMedMRC, in terms of medical entities and other terms. In this study, we employ the POS and NER toolkits[4] to tag medical entities and terms from 100 samples in the development set of CMedMRC. We also compare the statistics against those of two other Chinese MRC datasets, namely CMRC (Cui et al., 2019) and DuReader (He et al., 2018). The proportions of entities and five frequent POS tags in the three datasets are summarized in Figure 5. Comparing to the other two open-domain datasets, the proportion of entities in CMedMRC is very high (**11%**). In addition, the proportion of nouns (**27%**) is much higher than the other four POS tags in CMedMRC. The most likely cause is that existing models have difficulty recognizing all the medical terms, and treat them as common nouns. Among the three Chinese datasets, CMedMRC has the largest proportion (**38%**) of nouns and entities. Therefore, it is difficult for pre-trained language models to understand so many medical terms without additional medical background knowledge.

[4]We use jieba toolkit with additional medical term dictionaries. See `https://pypi.org/project/jieba/`.

C Experimental Settings

C.1 Medical Knowledge Base and Corpora

The underlying medical knowledge base is constructed by DXY, containing 44 relation types and over 4M relation triples. The KGs embedding trained by TransR (Lin et al., 2015b) on DXY-KG [5] containing 152,508 entities. In knowledge retrieval, the threshold of overlapped tokens is set to half of its own length. The medical pre-training corpora used in ERNIE-THU(Zhang et al., 2019) contains 5,937,695 text segments with 3,028,224,412 tokens (4.9 GB) after pre-processing.

C.2 Additional Training Details

In average, the training time for DrQA, BERT_base, MC-BERT, KT-NET, ERNIE, KMQA and CMedBERT takes 10, 16, 16, 27, 29, 28 and 25 minutes per epoch on a TiTAN RTX GPU. All the models are implemented by the PyTorch deep learning framework [6].

[5] https://portal.dxy.cn/
[6] https://pytorch.org/