CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark

Mosha Chen * **Zhen Bi** * **Xiaozhuan Liang** * **Lei Li** * **Ningyu Zhang** † **Xin Shang**

Kangping Yin **Chuanqi Tan** **Jian Xu** **Fei Huang** **Luo Si** **Yuan Ni** **Guotong Xie**

Zhifang Sui **Baobao Chang** **Hui Zong** **Zheng Yuan** **Linfeng Li** **Jun Yan**

Hongying Zan **Kunli Zhang** **Buzhou Tang**† **Qingcai Chen** †

CBLUE Team ‡

Abstract

Artificial Intelligence (AI), along with the recent progress in biomedical language understanding, is gradually offering great promise for medical practice. With the development of biomedical language understanding benchmarks, AI applications are widely used in the medical field. However, most benchmarks are limited to English, which makes it challenging to replicate many of the successes in English for other languages. To facilitate research in this direction, we collect real-world biomedical data and present the first Chinese Biomedical Language Understanding Evaluation (CBLUE) benchmark: a collection of natural language understanding tasks including named entity recognition, information extraction, clinical diagnosis normalization, single-sentence/sentence-pair classification, and an associated online platform for model evaluation, comparison, and analysis. To establish evaluation on these tasks, we report empirical results with the current 11 pre-trained Chinese models, and experimental results show that state-of-the-art neural models perform by far worse than the human ceiling. Our benchmark is released at `https://tianchi.aliyun.com/dataset/dataDetail?dataId=95414&lang=en-us`.

1 Introduction

Artificial intelligence is gradually changing the landscape of healthcare, and biomedical research [35]. With the fast advancement of biomedical datasets, biomedical natural language processing (BioNLP) has facilitated a broad range of applications such as biomedical text mining, which leverages textual data in Electronic Health Records (EHRs). For example, BioNLP methods can be employed to provide recommendations for specialized healthcare to those most at risk during pandemics (COVID-19) using the text and information in EHRs.

A key driving force behind such improvements and rapid iterations of models is the use of general evaluation datasets and benchmarks [9]. Pioneer benchmarks, such as BLURB [10], PubMedQA

*Equal contribution and shared co-first authorship.

†Corresponding author.

‡Author contributions are listed in the appendix.

Preprint. Under review.

[13], and others, have provided us with the opportunity to conduct research on biomedical language understanding and developing real-world applications. Unfortunately, most of these benchmarks are developed in English, which makes the development of the associated machine intelligence Anglo-centric. Meanwhile, other languages, such as Chinese, have unique linguistic characteristics and categories that need to be considered. Even though Chinese speakers account for a quarter of the world population, there have been no existing Chinese biomedical language understanding evaluation benchmarks.

To address this issue and facilitate natural language processing studies in Chinese, we take the first step in introducing a comprehensive **C**hinese **B**iomedical **L**anguage **U**nderstanding **E**valuation (**CBLUE**) benchmark with eight biomedical language understanding tasks. These tasks include named entity recognition, information extraction, clinical diagnosis normalization, short text classification, question answering (in transfer learning setting), intent classification, semantic similarity, and so on.

We evaluate several pre-trained Chinese language models on CBLUE and report their performance. The current models still perform by far worse than the standard of single-human performance, leaving room for future improvements. We also conduct a comprehensive analysis using case studies to indicate the challenges and linguistic differences in Chinese biomedical language understanding. We intend to develop a universal GLUE-like open platform for the Chinese BioNLP community, and this work is a small step in that direction. Overall, the main contributions of this study are as follows:

- We propose the first Chinese biomedical language understanding benchmark, an open-ended, community-driven project with eight diverse tasks. The proposed benchmark serves as a platform for the Chinese BioNLP community and encourages new dataset contributions.
- We report a systematic evaluation of 11 Chinese pre-trained language models to understand the challenges derived by these tasks. We release the source code of the baselines as a toolkit at `https://github.com/CBLUEbenchmark/CBLUE` for future research purposes.

2 Related Work

Several benchmarks have been developed to evaluate general language understanding over the past few years. GLUE [29] is one of the first frameworks developed as a formal challenge affording straightforward comparison between task-agnostic transfer learning techniques. SuperGLUE [28], styled after GLUE, introduce a new set of more difficult language understanding tasks, a software toolkit, and a public leaderboard. Other similarly motivated benchmarks include DecaNLP [22], which recast a set of target tasks into a general question-answering format and prohibit task-specific parameters, and SentEval [2], which evaluate explicitly fixed-size sentence embeddings. Non-English benchmarks include RussianSuperGLUE [25] and CLUE [34], which is a community-driven benchmark with nine Chinese natural language understanding tasks. These benchmarks in the general domain provide a north star goal for researchers and are part of the reason we can confidently say we have made great strides in our field.

For BioNLP, many datasets and benchmarks have been proposed [30, 18, 33] which promote the biomedical language understanding [1, 17, 16]. Tsatsaronis et al. [27] propose biomedical language understanding datasets as well as a competition on large-scale biomedical semantic indexing and question answering. Jin et al. [13] propose PubMedQA, a novel biomedical question answering dataset collected from PubMed abstracts. Pappas et al. [23] propose BioRead, which is a publicly available cloze-style biomedical machine reading comprehension (MRC) dataset. Gu et al. [10] create a leaderboard featuring the Biomedical Language Understanding & Reasoning Benchmark (BLURB). Unlike a general domain corpus, the annotation of a biomedical corpus needs expert intervention and is labor-intensive and time-consuming. Moreover, most of the benchmarks are based on English; ignoring other languages means that potentially valuable information may be lost, which can be helpful for generalization.

In this study, we focus on Chinese and aim to develop the first Chinese biomedical language understanding benchmark. Note that Chinese is linguistically different from English and other Indo-European languages, necessitating an evaluation BioNLP benchmark designed explicitly for Chinese.

Dataset	Task	Train	Dev	Test	Metrics
CMeEE	NER	15,000	5,000	3,000	Micro F1
CMeIE	Information Extraction	14,339	3,585	4,482	Micro F1
CHIP-CDN	Diagnosis Normalization	6,000	2,000	10,192	Micro F1
CHIP-STS	Sentence Similarity	16,000	4,000	10,000	Macro F1
CHIP-CTC	Sentence Classification	22,962	7,682	10,000	Macro F1
KUAKE-QIC	Intent Classification	6,931	1,955	1,994	Accuracy
KUAKE-QTR	Query-Document Relevance	24,174	2,913	5,465	Accuracy
KUAKE-QQR	Query-Query Relevance	15,000	1,600	1,596	Accuracy

Table 1: Task descriptions and statistics in CBLUE. CMeEE and CMeIE are sequence labeling tasks. Others are single sentence or sentence pair classification tasks.

3 CBLUE Overview

CBLUE consists of 8 biomedical language understanding tasks in Chinese. We will introduce the task definitions, detailed data collection procedures, and characteristics of CBLUE followingly.

3.1 Tasks

CMeEE Chinese Medical Named Entity Recognition, a dataset first released in CHIP2020[4], is used for CMeEE task. Given a pre-defined schema, the task is to identify and extract entities from the given sentence and classify them into nine categories: disease, clinical manifestations, drugs, medical equipment, medical procedures, body, medical examinations, microorganisms, and department.

CMeIE Chinese Medical Information Extraction, a dataset that is also released in CHIP2020 [11], is used for CMeIE task. The task is aimed at identifying both entities and relations in a sentence following the schema constraints. There are 53 relations defined in the dataset, including 10 synonymous sub-relationships and 43 other sub-relationships.

CHIP-CDN CHIP Clinical Diagnosis Normalization, a dataset that aims to standardize the terms from the final diagnoses of Chinese electronic medical records, is used for the CHIP-CDN task. Given the original phrase, the task is required to normalize it to standard terminology based on the International Classification of Diseases (ICD-10) standard for Beijing Clinical Edition v601.

CHIP-CTC CHIP Clinical Trial Classification, a dataset aimed at classifying clinical trials eligibility criteria, which are fundamental guidelines of clinical trials defined to identify whether a subject meets a clinical trial or not [38], is used for the CHIP-CTC task. All text data are collected from the website of the Chinese Clinical Trial Registry (ChiCTR) [5], and a total of 44 categories are defined. The task is like text classification; although it is not a new task, studies and corpus for the Chinese clinical trial criterion are *still limited*, and we hope to promote future researches for social benefits.

CHIP-STS CHIP Semantic Textual Similarity, a dataset for sentence similarity in the non-i.i.d. (non-independent and identically distributed) setting, is used for the CHIP-STS task. Specifically, the task aims to transfer learning between disease types on Chinese disease questions and answer data. Given question pairs related to 5 different diseases (The disease types in the training and testing set are different), the task intends to determine whether the semantics of the two sentences are similar.

KUAKE-QIC KUAKE Query Intent Classification, a dataset for intent classification, is used for the KUAKE-QIC task. Given the queries of search engines, the task requires to classify each of them into one of 11 medical intent categories defined in KUAKE-QIC, including diagnosis, etiology analysis, treatment plan, medical advice, test result analysis, disease description, consequence prediction, precautions, intended effects, treatment fees, and others.

[4] `http://cips-chip.org.cn/`

[5] `http://chictr.org.cn/`

KUAKE-QTR KUAKE Query Title Relevance, a dataset used to estimate the relevance of the title of a query document, is used for the KUAKE-QTR task. Given a query (e.g., "Symptoms of vitamin B deficiency"), the task aims to find the relevant title (e.g., "The main manifestations of vitamin B deficiency").

KUAKE-QQR KUAKE Query-Query Relevance, a dataset used to evaluate the relevance of the content expressed in two queries, is used for the KUAKE-QQR task. Similar to KUAKE-QTR, the task aims to estimate query-query relevance, which is an essential and challenging task in real-world search engines.

3.2 Data Collection

Since machine learning models are mostly data-driven, data plays a critical role, and it is pretty often in the form of a static dataset [8]. We collect data for different tasks from diverse sources, including clinical trials, EHRs, medical books, and search logs from real-world search engines. As biomedical data may contain private information such as the patient's name, age, and gender, **all collected datasets are anonymized and reviewed by the IRB committee of each data provider to preserve privacy.** We introduce the data collection details followingly.

Collection from Clinical Trials

Clinical trial eligibility criteria text is collected from ChiCTR, a non-profit organization that provides registration for clinical trial information for public research use. Eligibility criteria text is organized as a paragraph in the inclusion criteria and exclusion criteria in each trial registry file. Meaningless text was excluded, such as "The criteria is as follows", and the remained text was annotated to generate the CHIP-CTC dataset.

Collection from EHRs

We obtain the final diagnoses of the medical records from several Class A tertiary hospitals and sample a few diagnosis items from different medical departments to construct the CHIP-CDN dataset for research purposes. The diagnosis items are randomly sampled from the items not covered by the common medical synonyms dict. **No privacy information is involved in the final diagnoses.**

Collection from Medical Forum and Textbooks

Due to the COVID-19 pandemic, online consultation becomes more and more popular via the Internet. To promote data diversity, we select the online questions by patients to build the CHIP-STS dataset. Note that most of the questions are chief complaints. To ensure the authority and practicability of the corpus, we also select medical textbooks of Pediatrics [31], Clinical Pediatrics [26] and Clinical Practice[6]. We collect data from these sources to construct the CMeIE and CMeEE datasets.

Collection from Search Engine Logs

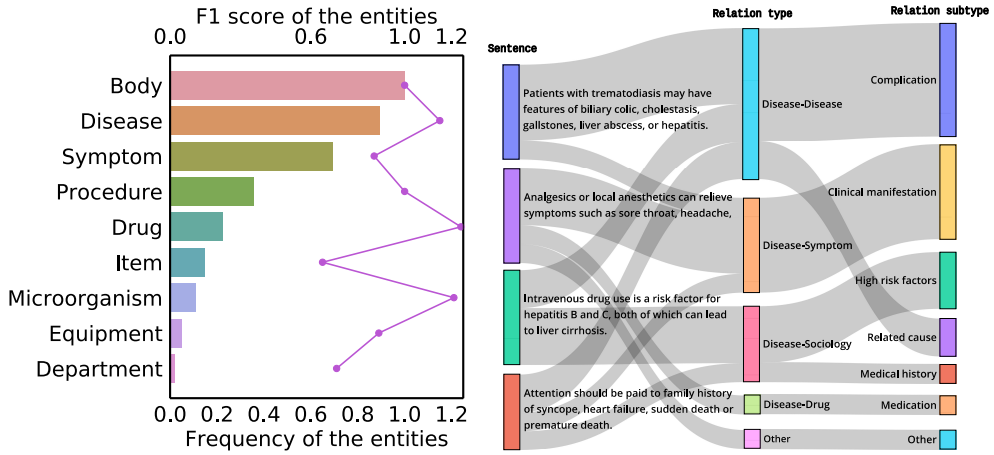
We also collect search logs from real-world search engines like the Alibaba KUAKE Search Engine[7]. First, we filter the search queries in the raw search logs by the medical tag to obtain candidate medical texts. Then, we sample the documents for each query with non-zero relevance scores (i.e., to determine if the document is relevant to the query). Specifically, we divide all the documents into three categories, namely high, middle, and tail documents, and then uniformly sample the data to guarantee diversity. We leverage the data from search logs to construct KUAKE-QTC, KUAKE-QTR, and KUAKE-QQR datasets.

3.3 Annotation

Each sample is annotated by three to five crowd workers, and the annotation with the majority of votes is taken to estimate human performance. During the annotation phase, we add control questions

[6] http://www.nhc.gov.cn/

[7] https://www.myquark.cn/



(a) CMeEE (b) CMeIE

Figure 1: Analysis of the named entity recognition and information extraction datasets. (a) illustrates the entity (coarse-grained) distribution in CMeEE and the impact of data distribution on the model's performance. We set entity type Body with the maximum number of entities to 1.0, and others to the ratio of number or F1 score to Body. (b) shows the relation hierarchy in CMeIE.

to prevent dishonest behaviors by the crowd workers. Consequently, we reject any annotations made by crowd workers who fail in the training phase and do not adopt the results of those who achieved low performance on the control tasks. We maintain strict and high criteria for approval and review at least 10 random samples from each worker to decide whether to approve or reject all their HITs. We also calculate the average inter-rater agreement between annotators using Fleiss' Kappa scores [7], finding that five out of six annotations show almost perfect agreement ($\kappa = 0.9$).

3.4 Characteristics

Utility-preserving Anonymization Biomedical data may be considered as a breach in the privacy of individuals because they usually contain sensitive information. Thus, we conduct utility-preserving anonymization following [15] to anonymize the data before releasing the benchmark.

Real-world Distribution To promote the generalization of models, all the data in our CBLUE benchmark follow real-world distribution without up/downsampling. As shown in Figure 1(a), our dataset follows long-tail distribution following Zipf's law so that all data will inevitably be long-tailed. However, long-tail distribution has no significant effect on performance. Further, some datasets, such as CMedIE, have label hierarchy with both coarse-grained and fine-grained relation labels, as shown in Figure 1(b).

Diverse Tasks Setting Our CBLUE benchmark includes eight diverse tasks, including named entity recognition, relation extraction, and single-sentence/sentence-pair classification. Besides the independent and i.i.d. scenarios, our CBLUE benchmark also contains a specific transfer learning scenario supported by the CHIP-STS dataset, in which the testing set has a different distribution from the training set.

3.5 Leaderboard

We provide a leaderboard for users to submit their own results on CBLUE. The evaluation system will give final scores for each task when users submit their prediction results. The platform offers 60 free GPU hours from Aliyun[8] to help researchers develop and train their models.

[8] https://tianchi.aliyun.com/notebook-ai/

3.6 Distribution and Maintenance

Our CBLUE benchmark was released online on April 1, 2021. Up to now, more than **300** researchers have applied the dataset, and over **80** teams have submitted their model predictions to our platform, including medical institutions (Peking Union Medical College Hospital, etc.), universities (Tsinghua University, Zhejiang University, etc.), and companies (Baidu, JD, etc.). We will continue to maintain the benchmark by attending to meet new requests and adding new tasks.

Model	CMeEE	CMeIE	CDN	CTC	STS	QIC	QTR	QQR	Avg.
BERT-base	62.1	54.0	55.4	69.2	83.0	84.3	60.0	**84.7**	69.1
BERT-wwm-ext-base	61.7	54.0	55.4	70.1	83.9	84.5	60.9	84.4	69.4
RoBERTa-large	62.1	54.4	56.5	**70.9**	84.7	84.2	60.9	82.9	69.6
RoBERTa-wwm-ext-base	62.4	53.7	56.4	69.4	83.7	85.5	60.3	82.7	69.3
RoBERTa-wwm-ext-large	61.8	**55.9**	55.7	69.0	85.2	85.3	62.8	84.4	70.0
ALBERT-tiny	50.5	35.9	50.2	61.0	79.7	75.8	55.5	79.8	61.1
ALBERT-xxlarge	61.8	47.6	37.5	66.9	84.8	**84.8**	62.2	83.1	66.1
ZEN	61.0	50.1	57.8	68.6	83.5	83.2	60.3	83.0	68.4
MacBERT-base	60.7	53.2	57.7	67.7	84.4	84.9	59.7	84.0	69.0
MacBERT-large	**62.4**	51.6	**59.3**	68.6	**85.6**	82.7	**62.9**	83.5	69.6
PCL-MedBERT	60.6	49.1	55.8	67.8	83.8	84.3	59.3	82.5	67.9
Human	67.0	66.0	65.0	78.0	93.0	88.0	71.0	89.0	77.1

Table 2: Performance of baseline models on CBLUE benchmark.

3.7 Reproducibility

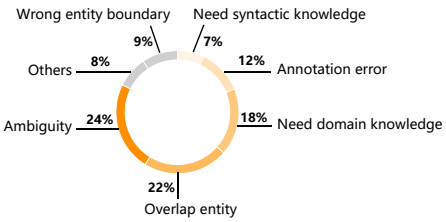
To make it easier to use the CBLUE benchmark, we also offer a toolkit implemented in PyTorch [24] for reproducibility. Our toolkit supports mainstream pre-training models and a wide range of target tasks. Different from existing pre-training model toolkits [37], the toolkit is aimed at fast validating performance on the CBLUE benchmark.

4 Experiments

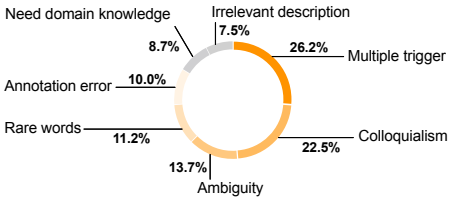
Baselines We conduct experiments with baselines based on different Chinese pre-trained language models. We add an additional output layer (e.g., MLP) for each CBLUE task and fine-tune the pre-trained models. Code for reproducibility is available in `https://github.com/CBLUEbenchmark/CBLUE`.

Models We evaluate CBLUE on the following public available Chinese pre-trained models:

- BERT-base [5]. We use the base model with 12 layers, 768 hidden layers, 12 heads, and 110 million parameters.
- BERT-wwm-ext-base [4]. A Chinese pre-trained BERT model with whole word masking.
- RoBERTa-large [21]. Compared with BERT, RoBERTa removes the next sentence prediction objective and dynamically changes the masking pattern applied to the training data.
- RoBERTa-wwm-ext-base/large. RoBERTa-wwm-ext is an efficient pre-trained model which integrates the advantages of RoBERTa and BERT-wwm.
- ALBERT-tiny/xxlarge [14]. ALBERT is a pre-trained model with two objectives: Masked Language Modeling (MLM) and Sentence Ordering Prediction (SOP), which shares weights across different layers in the transformer.
- ZEN [6]. A BERT-based Chinese text encoder enhanced by N-gram representations, where different combinations of characters are considered during training.
- Mac-BERT-base/large [3]. Mac-BERT is an improved BERT with novel MLM as a correction pre-training task, which mitigates the discrepancy of pre-training and fine-tuning.



(a) Error Analysis in CMeEE



(b) Error Analysis in KUAKE-QIC

Figure 2: We conduct error analysis on dataset CMeEE and QIC. For CMeEE, we divide error cases into 6 categories, including ambiguity, need domain knowledge, overlap entity, wrong entity boundary, annotation error, and others (long sequence, rare words, etc.). For KUAKE-QIC, we divide error cases into 7 categories, including multiple triggers, colloquialism, ambiguity, rare words, annotation error, irrelevant description, and need domain knowledge.

- PCL-MedBERT[9]. A pre-trained medical language model proposed by the Intelligent Medical Research Group at the Peng Cheng Laboratory, with excellent performance in medical question matching and named entity recognition.

		CMeEE	CMeIE	CDN	CTC	STS	QIC	QTR	QQR
Trained annotation	annotator 1	69.0	62.0	60.0	73.0	94.0	87.0	75.0	80.0
	annotator 2	62.0	65.0	69.0	75.0	93.0	91.0	62.0	88.0
	annotator 3	69.0	67.0	62.0	80.0	88.0	83.0	71.0	90.0
	avg	66.7	64.7	63.7	76.0	91.7	87.0	69.3	86.0
	majority	**67.0**	**66.0**	**65.0**	**78.0**	**93.0**	**88.0**	**71.0**	**89.0**
	best model	62.4	55.9	59.3	70.9	85.6	85.5	62.9	84.7

Table 3: Human performance of two-stage evaluation scores with the best-performed model. "avg" refers to the mean score from the three annotators. "majority" indicates the performance taken from the majority vote of amateur humans. Bold text denotes the best result among human and model prediction.

We implement all baselines with PyTorch [24]. Note that BERT-base, ALBERT-tiny/xxlarge, and RoBERTa-large are representatives of pre-trained language models. BERT-wwm-ext-base, RoBERTa-wwm-ext-base/large, ZEN, Mac-BERT-base/large utilize the specific characteristics (e.g., words and phrases) of the Chinese language. PCL-MedBERT further utilize domain-adaptive pre-training [12], which can consistently improve performance on tasks in the biomedical domain. We tune all the hyper-parameters based on the performance of each model on the development set. We implement each experiment five times and calculate the average performance. All the training details can be found in the appendix.

4.1 Benchmark Results

We report the results of our baseline models on the CBLUE benchmark in Table 2. We notice that the model obtain better performance with larger pre-trained language models. We also observe that models which use whole word masking do not always yield better performance than others in some tasks, such as CTC, QIC, QTR, and QQR, indicating that tasks in our benchmark are challenging, and more sophisticated technologies should be developed. Further, we find that ALBERT-tiny achieves comparable performance to base models in tasks of CDN, STS, QTR, and QQR, illustrating that smaller models may also be efficient in specific tasks. Finally, we notice that PCL-MedBERT, which tends to be state-of-the-art in Chinese biomedical text processing tasks, while does not perform as well as we expected. This further demonstrates the difficulty of our benchmark, and contemporary models may find it difficult to quickly achieve outstanding performance.

[9] https://code.ihub.org.cn/projects/1775

4.2 Human Performance

For all of the tasks in CBLUE, we ask human amateur annotators with no medical experience to label instances from the testing set and compute the annotators' majority vote against the gold label annotated by specialists. Similar to SuperGLUE [28], we first need to train the annotators before they work on the testing data. Annotators are asked to annotate some data from the development set; then, their annotations are validated against the gold standard. Annotators need to correct their annotation mistakes repeatedly so that they can master the specific tasks. Finally, they annotate instances from the testing data, and these annotations are used to compute the final human scores. The results are shown in Table 3 and the last row of Table 2. In most tasks, humans tend to behave better than machine learning models. We analyze the human performance detailedly in the next section.

Sentence	Word	Label	RO	MB
血液生化分析的结果显示维生素B缺乏率约为12%～19%。	血液生化分析	Ite	Pro	Pro
The results of blood biochemical analysis show that vitamin B lack rate is about 12% to 19%.	blood biochemical analysis	Ite	Pro	Pro
皮疹可因宿主产生特异性的抗毒素抗体而减少。	抗毒素抗体	Bod	O	Bod
The rash can be reduced by the host producing specific anti-toxin antibodies.	anti-toxin antibodies	Bod	O	Bod
根据遗传物质的结构和功能改变的不同，可将遗传病分为五类：1.染色体病指染色体数目异常，或者染色体结构异常，包括缺失、易位、倒位等	缺失, 易位, 倒位	Sym, Sym, Sym	O	Sym, Sym, Sym
According to the structure and function of genetic material, genetic diseases are divided into five categories: 1. Chromosomal diseases refer to abnormal chromosome number or chromosome structure abnormalities, including deletions, translocations, inversions...	deletions, translocations, inversions	Sym, Sym, Sym	O	Sym, Sym, Sym

Table 4: Case studies in CMeEE. We evaluate roberta-wwm-ext and PCL-MedBERT on 3 sampled sentences, with their gold labels and model predictions. Ite (medical examination items), Pro (medical procedure), Bod (body), and Sym (clinical symptoms) are labeled for medical named words. O means that the model fails to extract the entity from sentences. RO=roberta-wwm-ext, MB=PCL-MedBERT.

4.3 Case studies

We choose two datasets: CMeEE and KUAKE-QIC, a sequence labeling and classification task, respectively, to conduct case studies. As shown in Figure 2, we report the statistics of the proportion of various types of error cases[10]. For CMeEE, we notice that *overlap entity*, *ambiguity*, *need domain knowledge*, *annotation error* are major reasons that result in the prediction failure. Furthermore, there exist many instances with *overlap entity*, which may lead to confusion for the named entity recognition task. While in the analysis for KUAKE-QIC, almost half of bad cases are due to *multiple triggers* and *colloquialism*. *Colloquialism* is natural in search queries, which means that some descriptions of the Chinese medical text are too simplified, colloquial, or inaccurate.

We show some cases on CMeEE in Table 4. In the second row, we notice that given the instance of "皮疹可因宿主产生特异性的抗毒素抗体而减少 (*Rash can be reduced by the host producing specificanti-toxin antibodies.*)", ROBERTA and PCL-MedBERT obtain different predictions. The reason is that there exist medical terminologies such as "抗毒素抗体 (*anti-toxin antibodies*)". ROBERTA can not identify those tokens correctly, but PCL-MedBERT, pre-trained on the medical corpus, can successfully make it. Moreover, PCL-MedBERT can accurately extract entities "缺失,易位,倒位 (*eletions, translocations, inversions*)" from the long sentences, which is challenging for other models.

[10]See definitions of errors in the appendix.

We further show some cases on KUAKE-QIC in Table 5. In the first case, we notice that both BERT and BERT-ext fail to obtain the intent label of the query "请问淋巴细胞比率偏高、中性细胞比率偏低有事吗? (*Does it matter if the ratio of lymphocytes is high and the ratio of neutrophils is low?*)", while MedBERT can obtain the correct prediction. Since "淋巴细胞比率 (*ratio of lymphocytes*)" and "中性细胞比率 (*ratio of neutrophils*)" are biomedical terminologies, and the general pre-trained language model has to leverage domain knowledge to understand those phrases. Moreover, we observe that all model obtain incorrect predictions for the query "咨询：请问小孩一般什么时候出水痘 (*Consultation: When do children usually get chickenpox?*)" in the second case. Note that there exists lots of colloquial text in search queries (*colloquialism*), which have different distributions, thus, mislead the model predictions.

Query	**Model**			**Gold**
	BERT	BERT-ext	MedBERT	
请问淋巴细胞比率偏高、中性细胞比率偏低有事吗? Does it matter if the ratio of lymphocytes is high and the ratio of neutrophils is low?	病情诊断 Diagnosis	病情诊断 Diagnosis	指标解读 Test results analysis	指标解读 Test results analysis
咨询：请问小孩一般什么时候出水痘? Consultation: When do children usually get chickenpox?	其他 Other	其他 Other	其他 Other	疾病表述 Disease description
老人收缩压160，舒张压只有40多，是什么原因？怎么治疗? The systolic blood pressure of the elderly is 160, and the diastolic blood pressure is only more than 40. What is the reason? How to treat?	病情诊断 Diagnosis	病情诊断 Diagnosis	病情诊断 Diagnosis	治疗方案 Treatment

Table 5: Case studies in KUAKE-QIC. We evaluate the performance of baselines with 3 sampled instances. The correlation between Query and Title is divided into 3 levels (0-2), which means '*poorly related or unrelated*', '*related*' and '*strongly related*'. BERT = BERT-base, BERT-ext = BERT-wwm-ext-base, MedBERT = PCL-MedBERT.

As shown in Table 4 and Table 5, compared with other languages, Chinese language is very colloquial even in medical texts. Furthermore, Chinese is also a tonal language, and the meaning of a word changes according to its tone, which usually causes confusion and difficulties for machine reading. In summary, we conclude that **tasks in CBLUE are not easy to solve since the Chinese language has unique characteristics**, and more robust models that fully understand the semantics of Chinese, especially the informal or formal usages in the medical domain, should be taken into consideration.

4.4 Limitations

Although our CBLUE offers diverse settings, there are still some tasks not covered by the benchmark, such as medical dialogue generation [20, 19, 36] or medical diagnosis [32]. We encourage researchers in both academics and industry to contribute new datasets. Besides, our benchmark is static; thus, models may still achieve outstanding performance on tasks but fail on simple challenge examples and falter in real-world scenarios. We leave this as future works to construct a platform including dataset creation, model development, and assessment, leading to more robust and informative benchmarks.

4.5 Conclusion and Future Work

In this paper, we present a Chinese Biomedical Language Understanding Evaluation (CBLUE) benchmark, which consists of eight natural language understanding tasks, along with an online leaderboard for model evaluation. We evaluate 11 current language representation models on CBLUE and analyzed their results. The results illustrate the limited ability of state-of-the-art models to handle some of the more challenging tasks. In contrast to English benchmarks such as GLUE/SuperGLUE and BLURB, whose model performance already matches human performance, we observe that this is far from the truth for Chinese biomedical language understanding. We hope our benchmark can help promote developing stronger natural language understanding models in the future.

4.6 Broader Impact

The COVID-19 (coronavirus disease 2019) pandemic has had a significant impact on society, both because of the severe health effects of COVID-19 and the public health measures implemented to slow its spread. A lack of information fundamentally causes many difficulties experienced during the outbreak; attempts to address these needs caused an information overload for both researchers and the public. Biomedical natural language processing—the branch of artificial intelligence that interprets human language—can be applied to address many of the information needs making urgent by the COVID-19 pandemic. Unfortunately, most language benchmarks are in English, and no biomedical benchmark currently exists in Chinese. Our benchmark CBLUE, as the first Chinese biomedical language understanding benchmark, can serve as an open testbed for model evaluations to promote the advancement of this technology.

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics, 2019.

[2] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018.

[3] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*, 2020.

[4] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018.

[6] Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. Zen: pre-training chinese text encoder enhanced by n-gram representations. *arXiv preprint arXiv:1911.00720*, 2019.

[7] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[8] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[9] Pieter Gijsbers, Erin LeDell, Janek Thomas, Sébastien Poirier, Bernd Bischl, and Joaquin Vanschoren. An open source automl benchmark. *arXiv preprint arXiv:1907.00909*, 2019.

[10] 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. *CoRR*, abs/2007.15779, 2020.

[11] T. Guan, H. Zan, X. Zhou, H. Xu, and K Zhang. *CMeIE: Construction and Evaluation of Chinese Medical Information Extraction Dataset*. Natural Language Processing and Chinese Computing, 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I, 2020.

[12] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics, 2020.

[13] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics, 2019.

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

[15] Hyukki Lee, Soohyung Kim, Jong Wook Kim, and Yon Dohn Chung. Utility-preserving anonymization for health data publishing. *BMC Medical Informatics Decis. Mak.*, 17(1):104:1–104:12, 2017.

[16] 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.

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

[18] J. Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, A. P. Davis, C. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016, 2016.

[19] Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. Graph-evolving meta-learning for low-resource medical dialogue generation. *CoRR*, abs/2012.11988, 2020.

[20] Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *CoRR*, abs/2010.07497, 2020.

[21] 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.

[22] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730, 2018.

[23] Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. Bioread: A new dataset for biomedical reading comprehension. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[25] Tatiana Shavrina, Alena Fenogenova, Anton A. Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. Russiansuperglue: A russian language understanding evaluation benchmark. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4717–4726. Association for Computational Linguistics, 2020.

[26] Xiaoming Shen and Yonghao Gui. *Clinical Pediatrics 2nd edn*. People's Medical Publishing House, 2013.

[27] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28, 2015.

[28] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019.

[29] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[30] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: the covid-19 open research dataset. *CoRR*, abs/2004.10706, 2020.

[31] Weiping Wang, Kun Song, and Liwen Chang. *Pediatrics 9th edn*. People's Medical Publishing House, 2018.

[32] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai. Task-oriented dialogue system for automatic diagnosis. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 201–207. Association for Computational Linguistics, 2018.

[33] Y. Wu, Ruibang Luo, H. Leung, H. Ting, and T. Lam. Renet: A deep learning approach for extracting gene-disease associations from literature. In *RECOMB*, 2019.

[34] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A chinese language understanding evaluation benchmark. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4762–4772. International Committee on Computational Linguistics, 2020.

[35] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

[36] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. Meddialog: Large-scale medical dialogue datasets. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9241–9250. Association for Computational Linguistics, 2020.

[37] Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. UER: an open-source toolkit for pre-training models. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 241–246. Association for Computational Linguistics, 2019.

[38] Hui Zong, Jinxuan Yang, Zeyu Zhang, Zuofeng Li, and Xiaoyan Zhang. Semantic categorization of chinese eligibility criteria in clinical trials using machine learning methods. *BMC Medical Informatics Decis. Mak.*, 21(1):128, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.4.
   (c) Did you discuss any potential negative social impacts of your work? [Yes] See supplementary materials.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No]
   (b) Did you include complete proofs of all theoretical results? [No]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? See supplementary materials. [Yes] See supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See supplementary materials.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplementary materials.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See supplementary materials.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] because all the datasets either come from public resources or do not refer to ethics, which has been checked by the IRB committee of each data provider.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]