Supplementary Materials of "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 ‡

1 CBLUE Background

Standard datasets and shared tasks have played essential roles in promoting the development of AI technology. Taking the Chinese BioNLP community as an example, the CHIP (China Health Information Processing) conference releases biomedical-related shared tasks every year, which has extensively advanced Chinese biomedical NLP technology. However, some datasets are no longer available after the end of shared tasks, which has raised issues in the data acquisition and future research of the datasets.

In recent years, we can obtain state-of-the-art performance for many downstream tasks with the help of pre-trained language models. A significant trend is the emergence of multi-task leaderboards, such as GLUE (General Language Understanding Evaluation) and CLUE (Chinese Language Understanding Evaluation). These leaderboards provide a fair benchmark that attracts the attention of many researchers and further promotes the development of language model technology. For example, Microsoft has released BLURB (Biomedical Language Understanding & Reasoning Evaluation) at the end of 2020 in the medical field. Recently, the Tianchi platform has launched the CBLUE (Chinese Biomedical Language Understanding Evaluation) public benchmark under the guidance of the CHIP Society. We believe that the release of the CBLUE will further attract researchers' attention to the medical AI field and promote the development of the community.

CBLUE 1.0[4] comprises the previous shared tasks of the CHIP conference and the dataset from Alibaba QUAKE Search Engine, including named entity recognition, information extraction, clinical diagnosis normalization, single-sentence/sentence-pair classification.

---

*Equal contribution and shared co-first authorship.

†Corresponding author.

‡Author contributions are listed in the appendix.

[4] We release the benchmark following the CC BY-NC 4.0 license.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

2 Negative Impact

Although we ask domain experts and doctors to annotate all the corpus, there still exist some instances with wrong annotated labels. If a model was chosen based on numbers on the benchmark, this could cause real-world harm. Moreover, our benchmark lowers the bar of entry to work with biomedical data. While generally a good thing, it may dilute the pool of data-driven work in the biomedical field even more than it already it, making it hard for experts to spot the relevant work.

3 Detailed Task Introduction

3.1 Chinese Medical Named Entity Recognition Dataset (CMeEE)

Task Background As an essential subtask of information extraction, entity recognition has achieved promising results in recent years. Biomedical texts such as textbooks, encyclopedias, clinical trials, medical literature, electronic health records, and medical examination reports contain rich medical knowledge. Named entity recognition is the process of extracting medical terminologies, such as diseases and symptoms, from the above mentioned unstructured or semi-structured texts, and it can help significantly improve the efficiency of scientific research. CMeEE dataset is proposed for this purpose, and the original dataset was released at the CHIP2020 conference.

Task Description This task is defined as given the pre-defined schema and an input sentence to identify medical entities and to classify them into 9 categories, including disease (dis), clinical symptoms(sym), drugs (dru), medical equipment (equ), medical procedures (pro), body (bod), medical examination items (ite), microorganisms (mic), department (dep). For the detailed annotation instructions, please refer to the CBLUE official website, and examples are shown in Table 1.

Entity type	Entity subtype	Label	Example
疾病 disease	疾病或综合症 disease or syndrome 中毒或受伤 poisoned or injured 器官或细胞受损 damage to organs or cells	dis	尿潴留者易继发泌尿系感染 Patients with urinary retention are prone to secondary infections of the urinary system.
临床表现 clinical manifestations	症状 symptom 体征 physical sign	sym	逐渐出现呼吸困难、阵发性喘憋，发作时呼吸快而浅，并伴有呼气性喘鸣，明显鼻扇及三凹征 Then dyspnea and paroxysmal asthma may occur, along with shortness of breath, expiratory stridor, obvious flaring nares, and three-concave sign.
医疗程序 medical procedure	检查程序 check procedure 治疗 treatment 或预防程序 or preventive procedure	pro	用免疫学方法检测黑种病原体的特异抗原很有诊断价值，因其简单快速，常常用于早期诊断，诊断意义常较抗体检测更为可靠 It is of great diagnostic value to detect the specific antigen of a certain pathogen with immunoassay, a simple and quick assay that is intended for early diagnosis and proves more reliable than the antibody assay.

Table 1: Examples in CMeEE

Annotation Process The annotation guide was conducted by two medical experts from Class A tertiary hospitals and optimized during the trail annotation process. A total of 32 annotators had participated in the annotation process, including 2 medical experts who were also in charge of the annotation guideline, 4 experts from biomedical informatics field, 6 medical M.D., and 22 master students from computer science majors. The annotation lasted for about three months (from October 2018 to December 2018), as well as an additional month's time for curation. The total expense is about 50,000 RMB.

The annotation process was devided into two stages.

- Stage1: This stage was called the trail annotation phase. The medical experts gave training to the annotators to make sure they have a comprehensive understanding of the task. Two rounds of trail annotation were conducted by the annotators, with the purpose of getting familiar with the annotation task as well as discovering the unclear points of the guideline, and annotation problems were discussed and the medical experts improved the annotation guidelines according to the feedback iteratively.
- Stage2: For the first phase, each record was assigned to two annotators to label indepedently, and the medical experts and biomedical informatics experts would give in time help. The annotation results were compared automatically by the annotation tools (developed for CMeEE and CMeIE tasks) and any disagreement was rocorded and handed over to the next phase. At the second phase, medical experts and the annotators had a discussion for the disagreements records as well as other annotation problems, and the annotators made corrections. After the two stages, the IAA score (Kappa score) is 0.8537, which satisfied the research goal.

PII and IRB The corpus is collected from authorized medical textbooks or Clinical Practice, and no personally identifiable information or offensive content is involved in the text.

No PII is included in the above mentioned resources. The dataset does not refer to ethics, which has been checked by the IRB committee of the provider.

The original dataset format is a self-defined plain text format, to simplify the data pre-processing step, the CBLUE team has converted the data format to the unified JSON format with the permission of the data provider.

Evaluation Metrics This task uses strict Micro-F1 metrics.

Dataset Statistic This task has 15,000 training set data, 5,000 validation set data, 3,000 test set data. The corpus contains 938 files and 47,194 sentences. The average number of words contained per file is 2,355. The dataset contains 504 common pediatric diseases, 7,085 body parts, 12,907 clinical symptoms, and 4,354 medical procedures in total.

Dataset Provider The dataset is provided by:

- Key Laboratory of Computational Linguistics, Peking University, Ministry of Education, China
- Laboratory of Natural Language Processing, Zhengzhou University, China
- The Research Center for Artificial Intelligence, Peng Cheng Laboratory, China
- Harbin Institute of Technology, Shenzhen, China

3.2 Chinese Medical Information Extraction Dataset (CMeIE)

Task Background Entity and relation extraction is an essential information extraction task for natural language processing and knowledge graph (KG), which is used to detect pairs of entities and their relations from unstructured text. The technology of this task can apply to the medical field. For example, with entity and relation extraction, unstructured and semi-structured medical texts can construct medical knowledge graphs, which can serve lots of downstream tasks.

Task Description Given the schema and sentence, in which defines the relation (Predicate) and its related Subject and Object, such as (“subject_type”: “疾病”, “predicate”: “药物治疗”, “object_type”: “药物”). The task requires the model to automatically analyzing the sentence and then extract all the $Triples = [(S1, P1, O1), (S2, P2, O2)...]$ in the sentence. Table 2 shows the examples in the data set, and 53 SCHEMAs include 10 kinds of genus relations, 43 other sub-relations. The details are in the 53_schema.json file. For the detailed annotation instructions, please refer to the CBLUE official website, and examples are shown in Table 2.

Relation type	Relation subtype	Example
疾病_其他 disease_other	预防 prophylaxis	{’predicate’: ’预防-prevention’, ’subject’: ’麻风病-Leprosy’, ’subject_type’: ’疾病-disease’, ’object’: ’利福-rifampicin’, ’object_type’: ’其他-others’}
	阶段 phase	{’predicate’: ’阶段-phase’, ’subject’: ’肿瘤-tumor’, ’subject_type’: ’疾病-disease’, ’object’: ’I期-phase_‘’, ’object_type’: ’其他-others’}
	就诊科室 treatment department	{’predicate’: ’就诊科室-treatment_department’, ’subject’: ’腹主动脉瘤-abdominal_aortic_aneurysm’, ’subject_type’: ’疾病-disease’, ’object’: ’初级医疗保健医处-primary_medical_care_clinic’, ’object_type’: ’其他-others’}
疾病_其他治疗 disease_other treatment	辅助治疗 adjuvant therapy	{’predicate’: ’辅助治疗-adjuvant_therapy’, ’subject’: ’皮肤鳞状细胞癌-utaneous_squamous_cell_carcinoma’, ’subject_type’: ’疾病-disease’, ’object’: ’非手术破坏-non_surgical_destructio’, ’object_type’: ’其他治疗-other_treatment’}
	化疗 chemotherapy	{’predicate’: ’化疗-chemotherapy’, ’subject’: ’肿瘤-tumour’, ’subject_type’: ’皮肤鳞状细胞癌-cutaneous_squamous_cell_carcinoma’, ’object’: ’局部化疗-local_chemotherapy’, ’object_type’: ’其他治疗-other_treatment’}
	放射治疗 radiotherapy	{’predicate’: ’放射治疗-radiation_therapy’, ’subject’: ’非肿瘤性疼痛-non_cancer_pain’, ’subject_type’: ’疾病-disease’, ’object’: ’外照射-external_irradiation’, ’object_type’: ’其他治疗-other_treatment’}
疾病_手术治疗 disease_surgical treatment	手术治疗 surgical treatment	{’predicate’: ’手术治疗-surgical_treatment’, ’subject’: ’皮肤鳞状细胞癌-cutaneous _squamous_cell_carcinoma’, ’subject_type’: ’疾病-disease’, ’object’: ’传统手术切除-surgical_resection(traditional_therapy)’, ’object_type’: ’手术治疗-surgical_treatment’}

Table 2: Examples in CMeIE

Annotation Process The annotation guide was conducted by two medical experts from Class A tertiary hospitals and optimized during the trail annotation process. A total of 22 annotators had participated in the annotation process, including 2 medical experts who were also in charge of the annotation guideline, 2 experts from biomedical informatics field, 4 medical M.D., and 14 master students from computer science majors. The annotation lasted for about four months (from October 2019 to January 2020), which contained the annotation time as well as the curation time. The total expense is about 40,000 RMB.

Similar to the CMeEE dataset, the annotation process for CMeIE also contains the trail annotation stage and the formal annotation stage following the same pipeline. Besides, an additional step called the Chinese segmentation validation step was added for this dataset. The data provider has developed a segmentation tool for the medical texts which could generate the word segment as well as the POS tagging, and some specified POS tagging types (like 'disease', 'drug') could automatically check if there were potential missing named entities for this task, which could help assist the annotators to check the missing labels. The final IAA for this dataset is 0.83, which could satisfy the research purpose.

PII and IRB The corpus is collected from authorized medical textbooks or Clinical Practice, and no personally identifiable information or offensive content is involved in the text.

No PII is included in the above mentioned resources. The dataset does not refer to ethics, which has been checked by the IRB committee of the provider.

Evaluation Metrics The SPO results given by the participants need to be accurately matched. The strict Micro-F1 is used for evaluation.

Dataset Statistic This task has 14,339 training set data, 3,585 validation set data, 4,482 test set data. The dataset is from the pediatric corpus and common disease corpus. The pediatric corpus originates from 518 pediatric diseases, and the common disease corpus is derived from 109 common diseases. The dataset contains nearly 75,000 triples, 28,000 disease sentences, and 53 schemas.

Dataset Provider The dataset is provided by:

- Key Laboratory of Computational Linguistics, Peking University, Ministry of Education, China
- Laboratory of Natural Language Processing, Zhengzhou University, China
- The Research Center for Artificial Intelligence, Peng Cheng Laboratory, China
- Harbin Institute of Technology, Shenzhen, China

3.3 CHIP - Clinical Diagnosis Normalization Dataset (CHIP-CDN)

Task Background Clinical term normalization is a crucial task for both research and industry use. Clinically, there might be up to hundreds of different synonyms for the same diagnosis, symptoms, or procedures; for example, "heart tack" and "MI" both stand for the standard terminology "myocardial infarction". The goal of this task is to find the standard phrases (i.e., ICD codes) for the given clinical term. With the help of the standard code, it can help ease the burden of researchers for the statistical analysis of clinical trials; also, it can be helpful for the insurance companies on the DRGs or DIP-related applications. This task is proposed for this purpose, and the originally shared task was released at the CHIP2020 conference.

Task Description The task aims to standardize the terms from the final diagnoses of Chinese electronic medical records. No privacy information is involved in the final diagnoses. Given the original terms, it is required to predict its corresponding standard phrase from the standard vocabulary of "International Classification of Diseases (ICD-10) for Beijing Clinical Edition v601". For the detailed annotation instructions, please refer to the CBLUE official website. Examples are shown in Table 3.

Annotation Process The Chinese Diagnostic Normalization Data Set (CHIP-CDN) was annotated by the medical team of Yidu Cloud. They are all composed of people with medical background and clinician qualification certificates. This work took about 2 months, and the estimated cost was around 100,000 RMB in total.

The Chinese Diagnostic Normalization Data Set (CHIP-CDN) is completed by one round of labeling, one round of full audit, and one round of random quality inspection. Labeling and review are completed by ordinary labeling personnel with clinical qualifications, and random quality inspections are completed by high-level terminology experts.

Original terms	Normalization terms
右肺结节转移可能大 Possible nodule metastasis in the right lung	肺占位性病变## Space-occupying Lesion of the Lung 肺继发恶性肿瘤## Secondary Malignant Neoplasm of the Lung 转移性肿瘤 Metastatic Tumor
右肺结节住院 Hospitalization after detection of nodules in the right lung	肺占位性病变 Space-occupying Lesion of the Lung
左上肺胸膜下结节待查 Subpleural nodule in the left upper lung to be examined	胸膜占位 Space-occupying Lesion within the Pleural Space

Table 3: Examples in CHIP-CDN

PII and IRB The corpus is collected from EMR(electronical medical records) and only the final diagnoses field is used for the normalization research. The dataset dose not refer to ethics.

As shown in the example table, the final diagnoses has no PII included.

The original dataset format is a self-defined xlsx format, to unify the data pre-processing step, the CBLUE team has converted the data format to the JSON format with the permission of the data provider.

Evaluation Metrics The F1 score is calculated with (original diagnosis terms, standard phrases) pairs. Say, if the test set has m golden pairs, and the predicted result has n pairs, where k pairs are predicted correctly, then:

$$P = k/n, R = k/m, F1 = 2 * P * R/(P+R). \tag{1}$$

Dataset Statistic 8,000 training instances and 10,000 testing instances are provided. We split the original training set into 6,000 and 2,000 for the training and validation set, respectively.

Dataset Provider The dataset is provided by Yidu Cloud Technology Inc.

3.4 Clinical Trial Criterion Dataset (CHIP-CTC)

Task Background Clinical trials refer to scientific research conducted by human volunteers to determine the efficacy, safety, and side effects of a drug or a treatment method. It plays a crucial role in promoting the development of medicine and improving human health. Depending on the purpose of the experiment, the subjects may be patients or healthy volunteers. The goal of this task is to predict whether a subject meets a clinical trial or not. Recruitment of subjects for clinical trials is generally done through manual comparison of medical records and clinical trial screening criteria, which is time-consuming, laborious, and inefficient. In recent years, methods based on natural language processing have got successful in many biomedical applications. This task is proposed with the purpose of automatically classifying clinical trial eligibility criteria for the Chinese language, and the original task is released at the CHIP2019 conference. All the data comes from real clinical trials collected from the website of the Chinese Clinical Trial Registry (ChiCTR) [5], which is a non-profit organization providing registration for public research use. Each

Task Description A total of 44 pre-defined semantic categories are defined for this task, and the goal is to predict a given text to the correct category. For the detailed annotation instructions, please refer to the CBLUE official website. Examples of labeled data are shown in Table 4.

---

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

ID	Clinical trial sentence	Category
S1	年龄>80岁 Age: > 80	Age
S2	近期颅内或椎管内手术史 Recent intracranial/intraspinal surgery	Therapy or Surgery
S3	血糖<2.7mmol/L Blood glucose < 2.7 mmol/L	Laboratory Examinations

Table 4: Examples in CHIP-CTC

Annotation Process The CHIP-CTC corpus was annotated by three annotators. The first annotator is Zuofeng Li, a principal scientist in Philips Research China, with more than a decade of research experience in the biomedical domain. Other annotators were Zeyu Zhang (Ph.D. candidate) and Jinxuan Yang (Ph.D. candidate) in the biomedical informatics field from Tongji University. The annotation started in July 2019 and took about 1 month, further the corpus was used in CHIP 2019 shared task. The annotation was related to the annotator's research project, and no payment was required.

One experienced biomedical researcher (Z.L) and two raters (Z.Z and J.Y, Ph.D. candidate for biomedical informatics) of biomedical domains labeled the CHIP-CTC corpus with the 44 categories. First, they studied these categories' definitions, investigated a large amount of expression patterns of criteria sentences, and chose criteria examples of each category. Next, the two raters independently annotated the same 1000 sentences, then they checked annotations and discussed contradictions with Z.L until consensus was achieved. This step repeated 20 iterations and 20000 criteria sentences were annotated which were later used to calculate the inter-annotator agreement score (0.9920 by Cohen's kappa score). Finally, the remaining 18341 sentences were assigned to the two raters for annotation.

PII and IRB The corpus is collected from the Chinese Clinical Trial Registry (ChiCTR) website, which is a non-profit organization providing registration for public research use. For each registered clinical trail case on this website, it is already approved by the ethic committee of the organization. In addition, the annotation and corpus have also been reviewed and approved by Internal Committee on Biomedical Experiments (ICBE) in Philips. It is encouraged to use the corpus for academic research.

For each registered clinical trail report, no PII is included.

The original dataset format is a self-defined csv format, to unify the data pre-processing step, the CBLUE team has converted the data format to the JSON format with the permission of the data provider.

Evaluation Metrics The evaluation of this task uses Macro-F1. Suppose we have n categories, $C_1, ..., C_i, ..., C_n$. The accuracy rate P_i is the number of records correctly predicted to class C_i / the number of records predicted to be class C_i. Recall rate R_i = the number of records correctly predicted as the class C_i / the number of records of the real C_i class.

$$Average - F1 = (1/n) \sum_{i=1}^{n} \frac{2 * Pi * Ri}{Pi + Ri} \tag{2}$$

Dataset Statistic This task has 22,962 training set, 7,682 validation set, and 10000 test set.

Dataset Provider The dataset is provided by the School of Life Sciences and Technology, Tongji University and Philips Research China.

3.5 Semantic Textual Similarity Dataset (CHIP-STS)

Task Background CHIP-STS task aims to learn similar knowledge between disease types based on the Chinese online medical questions. Specifically, given question pairs from 5 different diseases, it is required to determine whether the semantics of the two sentences are similar or not. The originally shared task was released at the CHIP2019 conference.

Task Description The category represents the name of the disease type, including diabetes, hypertension, hepatitis, aids, and breast cancer. The label indicates whether the semantics of the questions are the same. If they are the same, they are marked as 1, and if they are not the same, they are marked as 0. Examples of labeling are shown in Table 5.

Question1	**Question2**	**Label**
糖尿病吃什么？ What should patients with diabetes eat?	糖尿病的食谱? What is the recommended dietary for patients with diabetes?	label:1
乙肝小三阳的危害? What is the harm of hepatitis B (HBsAg/HBeAb/HBcAb-positive)?	乙肝大三阳的危害? What is the harm of hepatitis B (HBsAg/HBeAg/HBcAb-positive)?	label:0

Table 5: Examples in CHIP-STS

Annotation Process The CHIP-STS corpus was annotated by five undergraduate annotators from medical colleges under the guidance of one surgeon and one physician. The task is relatively simple since it is a two-class classification one. The annotation process as well as the time of inspection lasted for two weeks. A total of 30,000 sentences pairs were annotated and the total annotation expense is 25,000 RMB.

The corpus was composed of five types of diseases, so each annotator was assigned two types of diseases to label, to guarantee that each type of disease was annotated by two raters. During the trail annotation process, each annotator was given 100 records to label, to test if they could understand the tasks thoroughly. Following that, the annotators start to label the remaining instances, and medical experts would give necessary help, like explaining the disease mechanism to assist the raters. Finally, each record was labeled by two different labelers and the disagreed pairs were selected for discussion and case study, the annotators would recheck the previous annotated results according to the experts' feedback. The IAA score was 0.93.

PII and IRB The corpus is collected from online questions from medical forum, and it doesn't refer to the ethics, which has been checked by the IRB committee of the provider.

During the annotation step, sentences with PHI information is discarded by the annotators manually. The CBLUE team has also validated the dataset record by record to guarentee there is no PII included.

The original dataset format is a self-defined csv format, to unify the data pre-processing step, the CBLUE team has converted the data format to the JSON format with the permission of the data provider.

Evaluation Metrics The evaluation of this task is Macro-F1.

Dataset Statistic This task has 16,000 training set, 4,000 validation set, and 10,000 tests set data.

Dataset Provider The dataset is provided by Ping An Technology.

3.6 KUAKE-Query Intent Classification Dataset (KUAKE-QIC)

Task Background In medical search scenarios, the understanding of query intent can significantly improve the relevance of search results. In particular, medical knowledge is highly specialized, and classifying query intentions can also help integrate medical knowledge to enhance the performance of search results. This task is proposed for this purpose.

Task Description There are 11 categories of medical intent labels, including diagnosis, etiology analysis, treatment plan, medical advice, test result analysis, disease description, consequence prediction, precautions, intended effects, treatment fees, and others. For the detailed annotation instructions, please refer to the CBLUE official website. Examples are shown in Table 6.

Intent	Sentences
病情诊断 disease diagnosis	最近早上起来浑身无力是怎么回事? Why do I always feel weak after I get up in the morning? 我家宝宝快五个月了，为什么偶尔会吐清水带? Why does my 5-month-old baby occasionally vomit clear liquid?
注意事项 precautions	哮喘应该注意些什么 What should patients with asthma pay attention to? 孕妇能不能吃榴莲 Can a pregnant woman eat durians? 柿子不能和什么一起吃 Which food cannot be eaten together with persimmons? 糖尿病人饮食注意什么啊? What should patients with diabetes pay attention to about their diet?
就医建议 medical advice	糖尿病该做什么检查? What examination should patients with diabetes receive? 肚子疼去什么科室? Which department should patients with stomachache visit?

Table 6: Examples in KUAKE-QIC

Annotation Process The KUAKE-QIC corpus was annotated by six annotators graduated from medical college, they were employed by Alibaba as full-time employee for the KUAKE department. They got passed the test for the specified annotation tasks before the annotation started. This task costed about 2 weeks and the annotation fee was 11,000 RMB with 22,000 labelled records, that's to say, 0.3 RMB / per record.

The annotation process was divided into three steps:

The first step was the trail annotation step, 2,000 records were selected for this stage. The annotators were grouped into 2 groups, each with 3 persons. The data provider had a strict metric for quality control, say, the IAA between the three persons within the same group must exceed 0.9.

The second stage is the formal annotation phase, and during this stage, 6 annotators were divided into three groups, each with 2 persons. A total of 20,000 records were annotated, IAA for this step was 0.9230.

The last step was the quality inspection step, sampling strategy was adapted and 300 records were sampled for validation, some common annotation problems were raised by the medical experts and the data would be fixed in a batch mode. In addition, some disagreed cases were made final decisions by the medical experts.

PII and IRB The corpus is collected from user queries from the KUAKE search engine, and it doesn't refer to the ethics, which has been checked by the IRB committee of the provider.

During the annotation step, sentences with PHI information or offensive information (like sexual queries) is discarded by the annotators manually. The dataset also got passed the data disclosure process of Alibaba.

The CBLUE team has also validated the dataset record by record to guarentee there is no PII included.

Evaluation Metrics Accuracy is used for the evaluation of this task.

Dataset Statistic This task has 6,931 training set data, 1,955 validation set data, and 1,994 test set data.

Dataset Provider The dataset is provided by Alibaba QUAKE Search Engine.

3.7 KUAKE- Query Title Relevance Dataset (KUAKE-QTR)

Task Background KUAKE Query Title Relevance is a dataset for query document (title) relevance estimation. For example, give the query "Symptoms of vitamin B deficiency", the relevant title should be "The main manifestations of vitamin B deficiency".

Task Description The correlation between Query and Title is divided into 4 levels (0-3), 0 is the worst, and 3 stands for the best match. For the detailed annotation instructions, please refer to the CBLUE official website. Examples are shown in Table 7.

Query	Title	Level
缺维生素b的症状 Symptoms of Vitamin B deficiency	维生素b缺乏症的主要表现 What are the major symptoms of Vitamin B deficiency?	3
大腿软组织损伤怎么办 How can I treat a soft tissue injury in the thigh?	腿部软组织损伤怎么办 What's the treatment for a soft tissue injury in the leg?	2
小腿抽筋是什么原因引起的 What causes lower leg cramps?	小腿抽筋后一直疼怎么办 How can I treat pains caused by lower leg cramps?	1
挑食是什么原因造成的 What is the cause of picky eating?	挑食是什么原因造成的 What is the cause of picky eating?	0

Table 7: Examples in KUAKE-QTR

Annotation Process The KUAKE-QTR corpus was annotated by a total of nine annotators, among which seven were from third-party crowd-sourcing undergraduates from medical colleges and two were from Alibaba full-time medical experts. The crowd-sourcing annotators were required to get trained and passed the annotation test before they could execute the task. The annotations lasted for 2 weeks and a total of 28,000 RMB was used.

Similar to the KUAKE-QIC task, the KUAKE-QTR annotation process was divided to three steps with minor changes:

The training and examination stage: The seven annotators got trained by the two FTE (full-time employee) experts to understand the tasks, then each one was given 200 records to label, which have ground-truth answer annotated by FTE experts. The precision must be above 85% to pass the test.

The second step was the formal annotation step, and each annotators were given 3,000 records to label, among which 100 were with golden labels by medical experts. The annotation tools would automatically evaluate the annotation quality by comparing the label between the annotators' ones and the golden ones. Help would be given to the annotators if necessary. Only the precision exceeding the threshold of 0.85 would be handed to the next round.

The last step was the quality inspection step, sampling strategy was adapted and 100 records were sampled for validation by the FTE medical experts, bad cases would be returned to the crowd-sourcing annotators to be fixed.

PII and IRB The corpus is collected from user queries from the KUAKE search engine, and it doesn't refer to the ethics, which has been checked by the IRB committee of the provider.

During the annotation step, sentences with PHI information or offensive information (like sexual queries) is discarded by the annotators manually. The dataset also got passed the data disclosure process of Alibaba.

The CBLUE team has also validated the dataset record by record to guarentee there is no PII included. One record with NULL label was discarded with the permission of the provider.

Evaluation Metrics Same as the KUAKE-QIC task, accuracy is used for the evaluation of this task.

Dataset Statistic This task has 24,174 training set data, 2,913 validation set data, and 54,65 test set data.

Dataset Provider This dataset is provided by Alibaba QUAKE Search Engine.

3.8 KUAKE - Query Query Relevance Dataset (KUAKE-QQR)

Task Background KUAKE Query-Query Relevance is a dataset that evaluates the relevance between two given queries to resolve the long-tail challenges for search engines. Similar to KUAKE-QTR, query-query relevance is an essential and challenging task in real-world search engines.

Task Description The correlation between Query and Title is divided into 3 levels (0-2), 0 is the worst, and 2 stands for the best correlation. For the detailed annotation instructions, please refer to the CBLUE official website. Examples are shown in Table 8.

Query	Query	Level
小孩子打呼噜是什么原因引起的 What causes children's snoring	小孩子打呼噜什么原因 What makes children snore?	2
双眼皮遗传规律 Heredity laws of double-fold eyelids	内双眼皮遗传 Heredity of hidden double-fold eyelids	1
白血病血常规有啥异常 What index of the CBC test will be abnormal for patients with leukemia?	白血病血检有哪些异常 What index of the blood test will be abnormal for patients with leukemia?	0

Table 8: Examples in KUAKE-QQR

Annotation Process The same as KUAKE-QTR except for the expense, which is 22,000 RMB in total.

PII and IRB The same as KUAKE-QTR.

Evaluation Metrics Same with the KUAKE-QIC and KUAKE-QTR tasks, accuracy is used for the evaluation metrics.

Dataset Statistic This task has 15,000 training set data, 1,600 validation set data, and 1,596 test set data.

Dataset Provider This dataset is provided by Alibaba QUAKE Search Engine.

4 Experiments Details

This section details the training procedures and hyper-parameters for each of the data sets. We utilize Pytorch to conduct experiments, and all running hyper-parameters are shown in the following Tables. There are two stages in CMeIE, namely, entity recognition (CMeEE-ER) and relation classification (CMeEE-RE). So we detail the hyper-parameters in CMeEE-ER and CMeEE-RE, respectively.

Requirements

- python3
- pytorch 1.7
- transformers 4.5.1

Method	Value
warmup_proportion	0.1
weight_decay	0.01
adam_epsilon	1e-8
max_grad_norm	1.0

Table 9: Common hyper-parameters for all CBLUE tasks

Model	epoch	batch_size	max_length	learning_rate
bert-base	5	32	128	4e-5
bert-wwm-ext	5	32	128	4e-5
roberta-wwm-ext	5	32	128	4e-5
roberta-wwm-ext-large	5	12	65	2e-5
roberta-large	5	12	65	2e-5
albert-tiny	10	32	128	5e-5
albert-xxlarge	5	12	65	1e-5
zen	5	20	128	4e-5
macbert-base	5	32	128	4e-5
macbert-large	5	12	80	2e-5
PCL-MedBERT	5	32	128	4e-5

Table 10: Hyper-parameters for the training of pre-trained models with a token classification head on top for named entity recognition of the CMeEE task.

- jieba
- gensim

Hyper-parameters for Specific Task is shown in Table 9-20

5 Error Analysis for Other Tasks

Ambiguity indicates that the instance has a similar context but different meaning, which mislead the prediction.

Need domain knowledge indicates that there exist biomedical terminologies in the instance which require domain knowledge to understand.

Need syntactic knowledge indicates that there exists complex syntactic structure in the instance, and the model fails to understand the correct meaning.

Model	epoch	batch_size	max_length	learning_rate
bert-base	7	32	128	5e-5
bert-wwm-ext	7	32	128	5e-5
roberta-wwm-ext	7	32	128	4e-5
roberta-wwm-ext-large	7	16	80	4e-5
roberta-large	7	16	80	2e-5
albert-tiny	10	32	128	4e-5
albert-xxlarge	7	16	80	1e-5
zen	7	20	128	4e-5
macbert-base	7	32	128	4e-5
macbert-large	7	20	80	2e-5
PCL-MedBERT	7	32	128	4e-5

Table 11: Hyper-parameters for the training of pre-trained models with a token-level classifier for subject and object recognition of the CMeIE task.

Model	**epoch**	**batch_size**	**max_length**	**learning_rate**
bert-base	8	32	128	5e-5
bert-wwm-ext	8	32	128	5e-5
roberta-wwm-ext	8	32	128	4e-5
roberta-wwm-ext-large	8	16	80	4e-5
roberta-large	8	16	80	2e-5
albert-tiny	10	32	128	4e-5
albert-xxlarge	8	16	80	1e-5
zen	8	20	128	4e-5
macbert-base	8	32	128	4e-5
macbert-large	8	20	80	2e-5
PCL-MedBERT	8	32	128	4e-5

Table 12: Hyper-parameters for the training of pre-trained models with a classifier for the entity pairs relation prediction of the CMeIE task.

Model	**epoch**	**batch_size**	**max_length**	**learning_rate**
bert-base	5	32	128	5e-5
bert-wwm-ext	5	32	128	5e-5
roberta-wwm-ext	5	32	128	4e-5
roberta-wwm-ext-large	5	20	50	3e-5
roberta-large	5	20	50	4e-5
albert-tiny	10	32	128	4e-5
albert-xxlarge	5	20	50	1e-5
zen	5	20	128	4e-5
macbert-base	5	32	128	4e-5
macbert-large	5	20	50	2e-5
PCL-MedBERT	5	32	128	4e-5

Table 13: Hyper-parameters for the training of pre-trained models with a sequence classification head on top for screening criteria classification of the CHIP-CTC task.

Param	**Value**
recall_k	200
num_negative_sample	10

Table 14: Hyper-parameters for the CHIP-CDN task. We model the CHIP-CDN task with two stages: recall stage and ranking stage. *num_negative_sample* sets the number of negative samples sampled for the training ranking model during the ranking stage. *recall_k* sets the number of candidates recalled in the recall stage.

Model	epoch	batch_size	max_length	learning_rate
bert-base	3	32	128	4e-5
bert-wwm-ext	3	32	128	5e-5
roberta-wwm-ext	3	32	128	4e-5
roberta-wwm-ext-large	3	32	40	4e-5
roberta-large	3	32	40	4e-5
albert-tiny	3	32	128	4e-5
albert-xxlarge	3	32	40	1e-5
zen	3	20	128	4e-5
macbert-base	3	32	128	4e-5
macbert-large	3	32	40	2e-5
PCL-MedBERT	3	32	128	4e-5

Table 15: Hyper-parameters for the training of pre-trained models with a sequence classifier for the ranking model of the CHIP-CDN task. We encode the pairs of the original term and standard phrase from candidates recalled during the recall stage and then pass the pooled output to the classifier, which predicts the relevance between the original term and standard phrase.

Model	epoch	batch_size	max_length	learning_rate
bert-base	20	32	128	4e-5
bert-wwm-ext	20	32	128	5e-5
roberta-wwm-ext	20	32	128	4e-5
roberta-wwm-ext-large	20	12	40	4e-5
roberta-large	20	12	40	4e-5
albert-tiny	20	32	128	4e-5
albert-xxlarge	20	12	40	1e-5
zen	20	20	128	4e-5
macbert-base	20	32	128	4e-5
macbert-large	20	12	40	2e-5
PCL-MedBERT	20	32	128	4e-5

Table 16: Hyper-parameters for the training of pre-trained models with a sequence classifier for the prediction of the number of standard phrases corresponding to the original term in the CHIP-CDN task.

Model	epoch	batch_size	max_length	learning_rate
bert-base	3	16	40	3e-5
bert-wwm-ext	3	16	40	3e-5
roberta-wwm-ext	3	16	40	4e-5
roberta-wwm-ext-large	3	16	40	4e-5
roberta-large	3	16	40	2e-5
albert-tiny	3	16	40	5e-5
albert-xxlarge	3	16	40	1e-5
zen	3	16	40	2e-5
macbert-base	3	16	40	3e-5
macbert-large	3	16	40	3e-5
PCL-MedBERT	3	16	40	2e-5

Table 17: Hyper-parameters for the training of pre-trained models with a sequence classifier for sentence similarity predication of the CHIP-STS task.

Model	**epoch**	**batch_size**	**max_length**	**learning_rate**
bert-base	3	16	50	2e-5
bert-wwm-ext	3	16	50	2e-5
roberta-wwm-ext	3	16	50	2e-5
roberta-wwm-ext-large	3	16	50	2e-5
roberta-large	3	16	50	3e-5
albert-tiny	3	16	50	5e-5
albert-xxlarge	3	16	50	1e-5
zen	3	16	50	2e-5
macbert-base	3	16	50	3e-5
macbert-large	3	16	50	2e-5
PCL-MedBERT	3	16	50	2e-5

Table 18: Hyper-parameters for the training of pre-trained models with a sequence classifier for query intention prediction of the KUAKE-QIC task.

Model	**epoch**	**batch_size**	**max_length**	**learning_rate**
bert-base	3	16	40	4e-5
bert-wwm-ext	3	16	40	2e-5
roberta-wwm-ext	3	16	40	3e-5
roberta-wwm-ext-large	3	16	40	2e-5
roberta-large	3	16	40	2e-5
albert-tiny	3	16	40	5e-5
albert-xxlarge	3	16	40	1e-5
zen	3	16	40	3e-5
macbert-base	3	16	40	2e-5
macbert-large	3	16	40	2e-5
PCL-MedBERT	3	16	40	3e-5

Table 19: Hyper-parameters of training the sequence classifier for the KUAKE-QTR task.

Model	**epoch**	**batch_size**	**max_length**	**learning_rate**
bert-base	3	16	30	3e-5
bert-wwm-ext	3	16	30	3e-5
roberta-wwm-ext	3	16	30	3e-5
roberta-wwm-ext-large	3	16	30	3e-5
roberta-large	3	16	30	2e-5
albert-tiny	3	16	30	5e-5
albert-xxlarge	3	16	30	3e-5
zen	3	16	30	2e-5
macbert-base	3	16	30	2e-5
macbert-large	3	16	30	2e-5
PCL-MedBERT	3	16	30	2e-5

Table 20: Hyper-parameters of training the sequence classifier for the KUAKE-QQR task.

Sentence	**Golden**	**RO**	**ME**
另一项研究显示，减荷鞋对内侧膝骨关节炎也没有效。 Another study showed that load-reducing shoes were not effective for medial knee osteoarthritis.	内侧膝骨关节炎| 辅助治疗| 减荷鞋 medial knee osteoarthritis, adjuvant therapy, load-reducing shoes	膝骨关节炎|辅助治疗|减荷鞋 medial knee osteoarthritis, adjuvant therapy, load-reducing shoes	膝骨关节炎| 辅助治疗|减荷鞋 medial knee osteoarthritis, adjuvant therapy, load-reducing shoes
精神疾病：焦虑和抑郁与失眠症高度相关。 Mental illness: anxiety and depression are related to insomnia.	焦虑|相关（导致）|失眠症 anxiety, related cause, insomnia	无|无|无 None|None|None	焦虑|相关（导致）|失眠症 anxiety, related cause, insomnia
在狂犬病感染晚期，患者常出现昏迷。 In the late stage of rabies infection, patients often appear comatose.	狂犬病|相关（转化）|昏迷 rabies, transform, comatose	无|无|无 None|None|None	无|无|无 None|None|None

Table 21: Error cases in CMeIE. We evaluate roberta-wwm-ext and PCL-MedBERT on 3 sampled sentences, with their gold labels and model predictions. Each label consists of subject | predicate | Object. None means that the model fails to predict. RO = roberta-wwm-ext, MB = PCL-MedBERT.

Sentence	**Label**	**RO**	**MB**
右第一足趾创伤性足趾切断 Right first toe traumatic toe cutting	单趾切断 Single toe cut	足趾损伤 Toe injury	单趾切断 Single toe cut
C3-4脊髓损伤 C3-4 spinal cord injury	颈部脊髓损伤 Neck spinal cord injury	脊髓损伤 Spinal cord injury	脊髓损伤 Spinal cord injury
肿瘤骨转移胃炎 Tumor bone metastatic gastritis	骨继发恶性肿瘤##转移性肿瘤##胃炎 Junior malignant tumor##Metastatic tumor##Gastritis	反流性胃炎##转移性肿瘤##胃炎 Reflux gastritis##Metastatic tumor##Gastritis	骨盆部肿瘤##转移性肿瘤##胃炎 Pelvic tumor##Metastatic tumor##Gastritis

Table 22: Error cases in CHIP-CDN. We evaluate roberta-wwm-ext and PCL-MedBERT on 3 sampled sentences, with their gold labels and model predictions. There may be multiple predicted values, separated by a "##". RO = roberta-wwm-ext, MB = PCL-MedBERT.

Overlap entity indicates there exist multiple overlapping entities in the instance.

Long sequence indicates that the input instance is very long.

Annotation error indicates that the annotated label is wrong.

Wrong entity boundary indicates that the instance has the wrong entity boundary.

Rare words indicates that there exist low-frequency words in the instance.

Multiple triggers indicates that there exist multiple indicative words which mislead the prediction.

Colloquialism (very common in the search queries) indicates that the instance is quite different from written language (e.g., with many abbreviations), thus, challenging the prediction model.

Irrelevant description indicates that the instance has lots of irrelevant information, which mislead the prediction.

Sentence	**Label**	**RO**	**MB**
既往多次行剖腹手术或腹腔广泛粘连者 Previous multi-time crashed surgery or abdominal adhesive	含有多类别的语句 Multiple	治疗或手术 Therapy or Surgery	治疗或手术 Therapy or Surgery
术前认知发育筛查（DST）发现发育迟缓 Preoperative cognitive development screening test(DST) finds development slow	诊断 Diagnostic	疾病 Disease	诊断 Diagnostic
已知发生中枢神经系统转移的患者 Patients who have been transferred in central nervous system	肿瘤进展 Neoplasm Status	疾病 Disease	疾病 Disease

Table 23: Error cases in CHIP-CTC. We evaluate roberta-wwm-ext and PCL-MedBERT on 3 sampled sentences, with their gold labels and model predictions. RO = roberta-wwm-ext, MB = PCL-MedBERT.

Query-A	**Query-B**	**Model**			**Gold**
		BE	BE+	MB	
汗液能传播乙肝病毒吗? Can sweat spread the hepatitis B virus?	乙肝的传播途径? How is hepatitis B transmitted?	0	0	0	1
哪种类型糖尿病? What type of diabetes?	我是什么类型的糖尿病? What type of diabetes am I?	1	1	1	0
如何防治艾滋病? How to prevent AIDS?	艾滋病防治条例。 AIDS Prevention and Control Regulations.	1	0	0	1

Table 24: Error cases in CHIP-STS. We evaluate performance of baselines with 3 sampled instances. The similarity between queries is divided into 2 levels (0-1), which means ’*unrelated*’ and ’*related*’. BE = BERT-base, BE+ = BERT-wwm-ext-base, MB = PCL-MedBERT.

Query-A	**Query-B**	**Model**			**Gold**
		BE	BE+	MB	
吃药能吃螃蟹吗? Can I eat crabs with medicine?	你好，吃完螃蟹后，可不可以吃药呢 Hello, does it matter to take medicine after eating crabs?	3	3	3	0
一颗蛋白卡路里。 Calories per egg white.	一个鸡蛋白的热量。 One egg white calories.	1	1	0	3
氨基酸用法用量。 Amino acid usage and dosage.	氨基酸的功效及用法用量。 Efficacy and dosage of amino acids.	2	2	2	1

Table 25: Error cases in KUAKE-QTR. We evaluate performance of baselines with 3 sampled instances. The correlation between Query and Title is divided into 4 levels (0-3), which means ’*unrelated*’, ’*poorly related*’, ’*related*’ and ’*strongly related*’. BE = BERT-base, BE+ = BERT-wwm-ext-base, MB = PCL-MedBERT.

Query-A	Query-B	Model			Gold
		BE	ZEN	MB	
益生菌是饭前喝还是饭后喝。 Should probiotics be drunk before or after meals.	益生菌是饭前喝还是饭后喝比较好。 Is it better to drink probiotics before or after meals	1	2	1	2
糖尿病能吃肉吗? Can diabetics eat meat?	高血糖能吃肉吗? Can hyperglycemic patients eat meat?	1	1	1	0
神经衰弱吃什么药去根? What drug does neurasthenic patient take effective?	神经衰弱吃什么药有效? What drug does neurasthenic patient take effective?	0	0	2	2

‘

Table 26: Error cases in KUAKE-QQR. We evaluate 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*'. BE = BERT-base, ZEN = ZEN, MB = PCL-MedBERT.

Contributions

Zhen Bi, Xiaozhuan Liang, Lei Li, Ningyu Zhang from Zhejiang University, AZFT Joint Lab for Knowledge Engine, Hangzhou Innovation Center wrote the paper.

Mosha Chen, Chuanqi Tan, Fei Huang, Luo Si from Alibaba Group and **Zheng Yuan** from the Center for Statistical Science, Tsinghua University contributed the CBLUE benchmark leaderboard and transformed the eight datasets from self-defined data format to unified JSON format.

Kunli Zhang from School of Information Engineering, Zhengzhou University, Peng Cheng Laboratory, China and **Baobao Chang** from Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Peng Cheng Laboratory, China contributed the dataset of CMeEE.

Hongying Zan from School of Information Engineering, Zhengzhou University, Peng Cheng Laboratory, China and **Zhifang Sui** from Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Peng Cheng Laboratory, China contributed the dataset of CMeIE.

Linfeng Li, Jun Yan from Yidu Cloud Technology Inc., Beijing, China contributed the dataset of CHIP-CDN.

Hui Zong from School of Life Sciences and Technology, Tongji University and Philips Research China contributed the dataset of CHIP-CTC.

Yuan Ni from Pingan Health Technology, Shanghai, China and **Guotong Xie** from Pingan Health Technology, China, Ping An Health Cloud Company Limited, China, Ping An International Smart City Technology Co., Ltd, China contributed the dataset of CHIP-STS.

Kangping Yin, Jian Xu from Alibaba Group and **Xin Shang** from School of Mathematical Science, Zhejiang University contributed the datasets of KUAKE-QIC, KUAKE-QTR, and KUAKE-QQR.

Buzhou Tang, Qingcai Chen from Harbin Institute of Technology (Shenzhen), Peng Cheng Laboratory, China advised the project, suggested tasks, and led the research.