Bioinformatics, 39(1), 2023, btac817
https://doi.org/10.1093/bioinformatics/btac817
Advance Access Publication Date: 20 December 2022
Original Paper

Data and text mining

A benchmark for automatic medical consultation system: frameworks, tasks and datasets

Wei Chen [1], Zhiwei Li[1], Hongyi Fang[1], Qianyuan Yao[1], Cheng Zhong[1], Jianye Hao[2], Qi Zhang[3], Xuanjing Huang[3], Jiajie Peng [4,5]* and Zhongyu Wei[1,5]*

[1]School of Data Science, Fudan University, Shanghai 200433, China, [2]College of Intelligence and Computing, Tianjin University, Tianjin 300072, China, [3]School of Computer Science, Fudan University, Shanghai 200433, China, [4]School of Computer Science, Northwestern Polytechnical University, Xi'an 710000, China and [5]Research Institute of Automatic and Complex Systems, Fudan University, Shanghai 200433, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 13, 2022; revised on November 8, 2022; editorial decision on December 13, 2022; accepted on December 20, 2022

Abstract

Motivation: In recent years, interest has arisen in using machine learning to improve the efficiency of automatic medical consultation and enhance patient experience. In this article, we propose two frameworks to support automatic medical consultation, namely doctor–patient dialogue understanding and task-oriented interaction. We create a new large medical dialogue dataset with multi-level fine-grained annotations and establish five independent tasks, including *named entity recognition*, *dialogue act classification*, *symptom label inference*, *medical report generation* and *diagnosis-oriented dialogue policy*.

Results: We report a set of benchmark results for each task, which shows the usability of the dataset and sets a baseline for future studies.

Availability and implementation: Both code and data are available from https://github.com/lemuria-wchen/imcs21.

Contact: jiajiepeng@nwpu.edu.cn or zywei@fudan.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Online medical consultation has shown great potential in improving the quality of healthcare services while reducing cost (Singh *et al.*, 2018), especially in the era of raging epidemics, such as *Coronavirus* (Singhal, 2020). This fact has accelerated the emergence of online medical communities like *SteadyMD* (https://www.steadymd.com) and *Haodafu* (https://www.haodf.com/). These platforms provide a medium for doctors and patients to communicate with each other remotely, which is called telemedicine (Wootton, 2001).

Typically, in telemedicine, the patient first provides a brief summary of their physical condition, i.e. self-report, then the doctor communicates with the patient to learn more about the patient's health condition. After sufficient inquiry, the doctor may make a diagnosis and provide further medical advice. The electronic record of this process is called Medical Consultation Record (MCR). Figure 1 demonstrates an example of MCR, which consists of patient's self-report, plain text of dialogue and corresponding disease category.

Recently, researchers have paid close attention to develop automatic approaches to facilitate online consultation service. Research topics include medical named entity recognition (Zhou *et al.*, 2021), drug recommendation (Zheng *et al.*, 2021), automatic text-based diagnosis (Chen *et al.*, 2020), health question answering (He *et al.*, 2020), medical report generation (Joshi *et al.*, 2020) and diagnostic policy (Wei *et al.*, 2018). Although progresses have been made to support automatic medical consultation from different perspectives, there is still a large gap between existing work and real-world application. We summarize this gap to two major limitations: (i) lack of design of frameworks and tasks for automatic medical consultation; and (ii) lack of benchmark datasets to support the development of research and application.

In this article, we make the first step to build a framework for automatic medical consultation and propose several tasks to cover the entire procedure. Two modes of frameworks are proposed to support both static and dynamic scenarios, namely, *dialogue understanding* and *task-oriented interaction*. The understanding framework aims to extract structured information from the dialogue context and generate useful labels to describe the dialogue state, which can include the patient's health status, patient's intention, etc. The interaction framework is designed to learn the dialogue policy (DP), i.e. to select the next action based on the current dialogue state, such as asking the patient whether he or she has a certain symptom. We create a corpus called IMCS-21 with multi-level fine-grained annotations to support

©The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023

Self-Report

My baby has diarrhea, five or six times a day.

Dialogue

……

Doctor*: What does the baby's stool look like? Is it watery stools?*

Patient*: Sometimes water, sometimes egg soup, and sometimes mushy.*

Doctor*: Does your baby have a fever or vomiting?*

Patient: *No, the child is in good spirits.*

Doctor: *Has the baby taken any medicines?*

Patient: *Medilac-vita and montmorillonite.*

……

Disease Category

Indigestion.

Fig. 1. An example of MCR, where the text is translated from Chinese

the research and application development of five tasks under the two modes. We develop widely used neural-based models for each task and report a set of benchmark results, which shows the usability of the corpus and sets a baseline for future studies. We conduct a comprehensive analysis of our corpus and tasks to show great future opportunities.

The *main contributions* of this article can be summarized as follows: (i) we propose a design of frameworks and tasks for automatic medical consultation and introduce IMCS-21, a large-scale annotated medical dialogue corpus, whose superiority makes it potentially a great benchmark for medical dialogue modeling; and (ii) we created neural-based models for each task and report a set of benchmark results. We will continue to track the progress of these tasks.

2 Related work

To build an automatic medical consultation system, learning from a large amount of actual doctor–patient conversations and directly imitate human behavior may be the best strategy. There are already a few medical dialogue corpus introduced by previous studies. These corpuses can be roughly divided into two categories.

One such category is original medical conversations corpus between patients and doctors with no annotations. MedDialog (Zeng *et al.*, 2020) is a large-scale medical dialogue dataset that contains a Chinese dataset with 3.4 million conversations covering 172 specialties of diseases and an English dataset with 0.26 million conversations covering 96 specialties of diseases. KaMed (Li *et al.*, 2021) is a knowledge aware medical dialogue dataset that contains over 60 000 medical dialogue sessions and is equipped with external medical knowledge from Chinese medical knowledge platform. The tasks built on these corpuses are usually response generation in dialogue systems, on which researchers can build automated medical chatbots. However, the responses generated by such end-to-end chatbots lack interpretability and controllability, which has strong limitations in healthcare applications.

Another category is the annotated doctor–patient medical dialogue corpus. The annotated content of these corpuses is related to the task they focus on and the researchers establish a series of medical dialogue modeling tasks including natural language understanding, natural language generation and DP. MSL (Shi *et al.*, 2020) is a dataset for slot filling task, which aims to transform a natural language medical query in which colloquial expressions exist into the formal representation with discrete logical forms to perform correct query. CMDD (Lin *et al.*, 2019), SAT (Du *et al.*, 2019) and MIE (Zhang *et al.*, 2020) are datasets for medical information extraction task, which is extract mentioned entities and their corresponding status. MZ (Wei *et al.*, 2018), DX (Xu *et al.*, 2019) and RD/SD (Zhong *et al.*, 2022) are datasets that contain structured symptom features to learn the DP for symptom-based automatic diagnosis. Chunyu (Lin *et al.*, 2021) is a dataset for end-to-end diagnosis-oriented response generation task. MedDG (Liu *et al.*, 2022) contains more than 17K conversations with annotated entities, and two medical dialogue tasks are established. One is the next sentence entity prediction and the other is the dialogue response generation.

One challenge of existing datasets is the medical label insufficiency. The majority of datasets only provide one specific medical labels, e.g. medical entities. These labels are too coarse to accurately describe the patient's state and intent. Another challenge is the small scale of existing annotated datasets, typically on the order of hundreds of dialogues. In the Supplementary Material, we present the comparative details between IMCS-21 and existing medical (dialogue) datasets.

3 Automatic medical consultation: frameworks and tasks

We present our design of frameworks and tasks for automatic medical consultation system in Figure 2.

3.1 Dialogue understanding framework

The understanding framework includes four tasks: Named Entity Recognition (NER), Dialogue Act Classification (DAC), Symptom Label Inference (SLI) and Medical Report Generation (MRG).

Named entity recognition: Medical NER task aims to recognize predefined medical named entities from medical texts (Zhou *et al.*, 2021). Medical related entities are widely present in actual doctor–patient conversations, and NER is a basic task for extracting medical semantics.

Dialogue act classification: DAC is the task of classifying an utterance with respect to the function it serves in a dialogue, i.e. the act the speaker is performing (Liu *et al.*, 2017). In medical dialogues, the identification of dialogue act is an important aspect in analyzing the doctor's and patient's intent and what they are trying to convey.

Symptom label inference: Symptoms are the main topics discussed in medical dialogues and an important basis for doctors to make a diagnosis (Lin *et al.*, 2019). The goal of SLI task is to identify mentioned symptoms from the dialogue, align them to standardized names and determine whether a patient suffers from these symptoms. SLI task generates a clearer structured symptom features about the patient.

Medical report generation: Medical report captures and summarizes the important parts of the medical conversation needed for clinical decision making and subsequent follow ups (Joshi *et al.*, 2020). As a way to record and convey medical information, MRG task addresses a practical need and plays an important role in medical practice.

3.2 Task-oriented interaction framework

The interaction framework controls the process of man–machine dialogue, which is to determine the future dialogue strategy based on historical information. For interaction framework, we introduce Diagnosis-oriented Dialogue Policy (DDP) task, which follows the setting of task-oriented dialogue system (Wei *et al.*, 2018).

Diagnosis-oriented dialogue policy: The DDP task aims to learn the optimal policy for symptom-based automatic disease diagnosis. The policy is expected to efficiently find potential symptoms of patients and make a correct diagnosis, through several turns of interaction. It is worth noting that the training data required for DDP is exactly the structured symptom features the SLI task needs to predict.

4 Medical dialogue corpus: IMCS-21

In this section, we present our collection and analysis of the annotated dataset. The raw data come from Muzhi (http://muzhi.baidu.com), a Chinese online health community that provides professional medical consulting service for patients. We collect extensive MCRs for 10 pediatric diseases. After removing some incomplete samples and samples with too short dialogues, we annotate the filtered samples to form our medical dialogue corpus, which we call IMCS-21.

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023

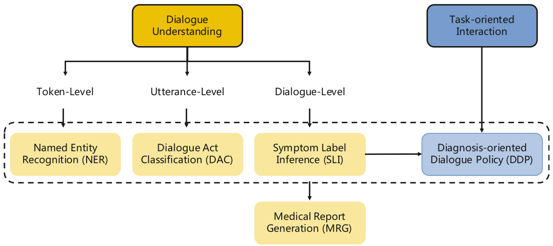

Dialogue Understanding
Task-oriented Interaction
Token-Level
Utterance-Level
Dialogue-Level
Named Entity Recognition (NER)
Dialogue Act Classification (DAC)
Symptom Label Inference (SLI)
Diagnosis-oriented Dialogue Policy (DDP)
Medical Report Generation (MRG)

Fig. 2. Framework and task design for automatic medical consultation

4.1 Annotation scheme

The annotation scheme is designed by medical experts with consideration of our task design as well as actual scenarios of online consultation.

Specifically, we collect multi-level annotations for each MCR, including token level, utterance level and dialogue level. At token level, the annotator is asked to find medical named entities; at utterance level, the intention of each utterance is annotated; at dialogue level, the symptom features are collected, and the medical report is manually written.

Medical named entity: We define five main categories of entities for annotation after discussing with domain experts, i.e. *symptom (SX)*, *drug name (DN)*, *drug category (DC)*, *examination (EX)* and *operation (OP)*. These categories of entities, we believe are important for understanding the doctor–patient dialogue. Among them, *drug name* represents a specific drug name while *drug category* represents a class of drugs with a certain efficacy. For example, 'aspirin' belongs to *DN* while 'anti-inflammatory drug' belongs to *DC*. Besides, *OP* represents related medical operations, such as 'infusion', 'atomization', etc.

Inside–outside–beginning (BIO) (Ramshaw and Marcus, 1999) tagging scheme is employed, where 'B' and 'I' determine the boundary of an entity. For each Chinese character, 'B' stands for the beginning of the entity, 'I' means inside and 'O' means other. This results in 11 possible tags for tokens. We assign an initial label to each token using a rule-based algorithm (Aho and Corasick, 1975) to prompt the annotation process.

Dialogue act: The categories of dialogue acts are determined according to the specific content of the utterance. It can be broadly divided into two big categories: *request (R)* and *inform (I)*, one means '*ask for information*', and another means '*tell the information*'. We further categorize the content of information conveyed as: *basic information (BI)*, *symptom (SX)*, *etiology (ETIOL)*, *existing exam and treatment (EET)*, *medical advice (MA)*, *drug recommendation (DR)*, *precautions (PRCTN)* and *diagnose (DIAG)*. There are both request and inform versions for all categories except *DIAG*. Utterances that does not fall into the above categories will be labeled as *other (OTR)*.

This results in a total of 16 possible categories of dialogue acts in our scheme. Each utterance in a dialogue is tagged with one of these categories. In this article, we use abbreviations to denote specific dialogue acts. For example, *R-SX* is abbreviated for *request-symptom*, which represents the intent of asking someone for relevant symptoms. To be more intuitive, we demonstrate several examples for each entity category and dialogue act category in the Supplementary Material.

Symptom label: In order to clarify the relationship between the symptoms appearing in the dialogue and the patient, each symptom entity is additionally tagged with a label: *Positive (POS)*, *Negative (NEG)* or *Not Sure (NS)*, to indicate whether the patient has the symptom. The symptom label determines the relationship between the symptom and the patient. The annotator can infer the symptom label by observing the utterance where the symptom entity is located and its context. Besides, all identified symptoms are normalized by linking them to the most relevant one on SNOMED-CT2 (https://www.snomed.org/snomed-ct), which can unify different expressions of the same symptom into one standard name. Symptoms mentioned in self-report are also identified and normalized.

Table 1. Statistics of IMCS-21

Statistics	Avg.
# of utterances per dialogue	40
# of characters per utterance	523
# of characters per self-report	57
# of entities per dialogue (annotated)	26
# of characters per medical report (annotated)	88

Medical report: Annotators are also required to write a report in specified format to summarize the medical consultation case. It contains six parts: (i) *chief complaint*: patient's main symptoms or signs; (ii) *present disease*: description of main symptoms; (iii) *auxiliary examination*: the patient's existing examinations, examination results, records, etc.; (iv) *past medical history*: previous health conditions and illnesses; (v) *diagnosis*: diagnosis of disease; and (vi) *suggestions*: doctor's suggestions of inspection recommendations, drug treatment and precautions. Annotators are required to fill in these parts and leave it blank if the part is not mentioned in the dialogue.

4.2 Inter-annotator agreement

For the annotation of medical dialogues, we develop a web-based tool, which can be utilized for general-purpose multi-turn dialogue labeling. We recruit undergraduates and postgraduates in medical school to annotate our corpus. All annotators are people who are willing to participate and over the age of 18.

Two annotations per dialogue are gathered, and inconsistent parts are further finalized by a third annotator. We use Cohen's kappa coefficient (Banerjee *et al.*, 1999) to estimate the inter-annotator agreement. For the annotations of medical named entities and dialogue acts, the kappa coefficients are 83.11% and 76.41%, respectively; for the annotations of symptom labels, the kappa coefficients is 92.71%; for medical reports, both reports are remained for reference. These results show that the consistency between the annotators is satisfactory.

4.3 Corpus analysis

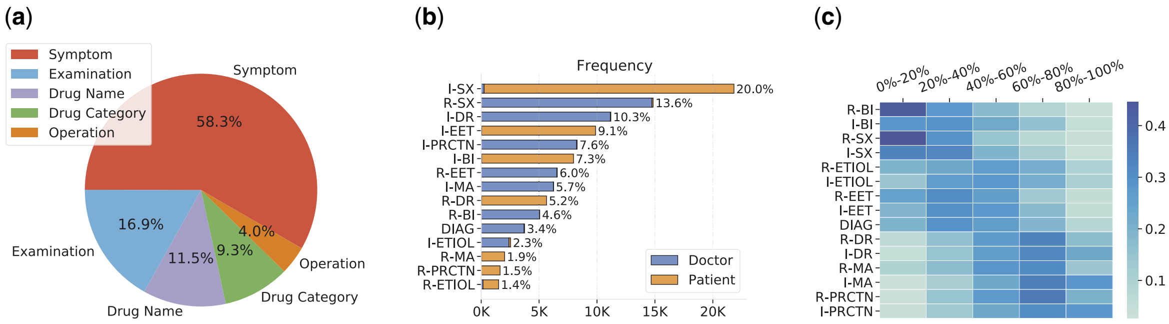
Corpus statistics: IMCS-21 contains a total of 4116 annotated samples with 164 731 utterances, which covers 10 pediatric diseases: *bronchitis*, *fever*, *diarrhea*, *upper respiratory infection*, *dyspepsia*, *cold*, *cough*, *jaundice*, *constipation* and *bronchopneumonia*. Each dialogue contains an average of 40 utterances, 523 Chinese characters (580 characters if including self-report) and 26 entities (see in Table 1).

Dialogue content analysis: The distribution of number of entity categories and dialogue act categories are shown in Figure 3a and b. Briefly, *symptom* entities appear the most in conversations, about 58.3%. Similarly, the two categories with the highest proportion of dialogue acts are *I-SX* and *R-SX*. This indicates that doctor–patient conversations mainly discuss the patient's symptoms. Examinations, drugs, advice and precautions are also common topics, this suggests that patients try to find medical solutions in consultations.

It is worth noting that for any specific category of dialogue act, it is either almost from doctors or patients (Fig. 3b). However, there are some exceptions. For example, the category *I-SX* means telling the other about the symptoms, which intuitively will only come from the patient who tells the doctor his symptom. But sometimes the doctor may remind the patient what symptoms they actually have, based on their previous vague description. For example, the utterance '*your body temperature is relatively high, it is febrile*' from the doctor will be labeled as *I-SX*.

Dialogue structure analysis: Figure 3c presents the positional distribution characteristics of dialogue act categories. We divide utterances in a dialogue into five parts according to their locations. For example, 0–20% means the sentences appeared in the first fifth of

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023



# of entity categories

dialogue acts

position of dialogue acts

Fig. 3. Pie chart for number of entity categories (a), and bar chart (b) and heat-map (c) for the number and locations of dialogue acts, respectively. Note that, we exclude category other in the statistics of dialogue acts.

the conversation. From the figure, we conclude that with the in-depth of medical consultation, the focus gradually shifts from symptoms to drugs, treatments and precautions. Clearly, there are certain regularities in the structure of medical dialogues for the purpose of diagnosis.

Symptoms analysis: We present the statistics of symptoms and symptom labels in Table 2. Each self-report and dialogue contains 1.7 and 6.6 (unique) symptom entities on average. In the dialogue, the number of non-positive symptom entities account for nearly 40%, which means that a large proportion of the symptoms in the conversation may not be related to the patient.

Medical reports analysis: In the annotation of medical reports, without being provided with true disease labels, the annotators are required to populate the *diagnosis* part with the patient's disease they infer from the conversation. Therefore, the accuracy of the content of this part can roughly assess how well the annotator understands the dialogue. By regex matching, we find that in 84.7% of the reports, the content of the *diagnosis* part contains the text of the actual disease or the key concepts, which ensures the quality of medical reports acceptably.

It is worth noting that some diseases are hard to distinguish from others, or are themselves a symptom of other diseases. In this case, annotators are easily confused. For example, when the real disease is *Cold*, only 65.6% of the reports contain the key concepts of cold in the *diagnosis* part. When the disease is *Jaundice*, the proportion is as high as 98.1%.

Besides, the *Present disease* and *Suggestions* part has about 30 and 20 words on average respectively, which occupy the main content of medical reports, while a considerable percentage (about 60%) of *past medical history* is empty, because this part is less involved in the dialogue.

5 IMCS-21 as a new benchmark

As introduced in Section 3, we break down the medical consultation modeling into two modes of frameworks, comprising a total of five tasks. To show the potential usefulness of IMCS-21, we establish a standard split for IMCS-21 at the dialogue level, and report a benchmark result for each of the task: NER, DAC, SLI, MRG and DDP. The split is consistent across all tasks, consisting of a training set with 2472 dialogues, a develop set with 833 dialogues and a test set with 811 dialogues.

Before presenting the experimental results, we first introduce some notations. Let $X = \{x_1, x_2, \ldots, x_T\}$ be a piece of doctor–patient dialogue, where $x_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$ is the i-th utterance and $x_{i,j} \in \mathcal{V}$ is the j-th token in x_i. The self-report and the disease category of the patient are denoted as x_0 and $y \in \mathcal{D}$, respectively, then a MCR can be represented as: $\{x_0, X, y\}$.

The formalization of each task will be introduced in each subsection, and readers can refer to the notations in Table 3 to better understand our task and evaluation settings. For each task, we only report the baseline models, evaluation metrics and experimental results, the details of experimental settings are provided in the Supplementary Material.

Table 2. Statistics of symptoms and symptom labels

	Self-report	Dialogue (POS/NEG/NS)	Total
# of Symptoms	1.7	4.0/1.6/1.0	8.3

Table 3. List of notations for task formalization

Sign	Description	Dimension
$\mathcal{D}$	The set of all diseases	10
$\mathcal{B}$	The set of all BIO tags for named entities	11
$\mathcal{A}$	The set of all dialogue act categories	16
$\mathcal{S}$	The set of all normalized symptom names	331
$\mathcal{L}$	The set of symptom labels	3
$\mathcal{V}$	Vocabulary of source and target tokens	3138

Note: The *dimension* column denotes the size of the set represented by these notations in our task settings. There are three elements in $\mathcal{L}$, namely *POS, NEG and NS.*

5.1 Named entity recognition

Task formalization: Robust medical NER is the first step in understanding doctor–patient conversations. The NER task is designed to automatically predict the boundaries and categories of pre-defined medical named entities contained in the dialogue. Formally, the NER task aims to predict the BIO label $b_i^j \in \mathcal{B}$ for each token $x_{i,j}$ given the utterance x_i.

Experimental settings: We use several popular Chinese named entity models as baselines, including: (i) Lattice LSTM (Zhang and Yang, 2018a), an extension of Char-LSTM that incorporates lexical information into native LSTM; (ii) BERT (Devlin *et al.*, 2019), a bidirectional Transformer encoder with large-scale language pre-training; (iii) ERNIE (Zhang *et al.*, 2019); an improved BERT that adopts entity-level masking and phrase-level masking during pre-training; (iv) FLAT (Li *et al.*, 2020), a flat-lattice Transformer that converts the lattice structure into a flat structure consisting of spans; (v) LEBERT (Liu *et al.*, 2021), a lexicon enhanced BERT for Chinese sequence labeling, which integrates external lexicon knowledge into BERT layers by a lexicon adapter layer; (vi) MC-BERT (Zhang *et al.*, 2022), a BERT model pre-trained on 20M Chinese biomedical sentences using whole entity masking and whole span masking; (vii) ERNIE-Health (Zhang *et al.*, 2019), a language model (LM) based on ERNIE pre-trained on 126.9G biomedical Chinese dataset, including medical dialogues, scientific articles on medicine and

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023

Table 4. Experimental results for medical NER task

Models	$P\uparrow$	$R\uparrow$	$F1\uparrow$
Lattice LSTM (Zhang and Yang, 2018a)	89.37	90.84	90.10
BERT-CRF (Devlin *et al.*, 2019)	88.46	92.35	90.37
ERNIE (Zhang *et al.*, 2019)	88.87	92.27	90.53
FLAT (Li *et al.*, 2020)	88.76	92.07	90.38
LEBERT (Liu *et al.*, 2021)	86.53	**92.91**	89.60
MC-BERT (Zhang *et al.*, 2022)	88.92	92.18	90.52
ERNIE-Health (Zhang *et al.*, 2019)	**89.71**	2.82	**91.24**

Note: The up arrows and down arrows indicate that the higher the better and the lower the better for the number in the column, respectively. All the numbers are percentage values, with the highest value highlighted. It is the same for other tables of experimental results.

All the boldface values in Table are significantly at the 5% significance level.

healthcare, electronic medical records and electronic textbooks on medicine and clinical pathology.

For evaluation, we report token-level metrics, including precision (P), recall (R) and F1 score ($F1$).

Experimental results: In Table 4, we present the experimental results of the NER task. All baselines achieve $F1$ scores around 90%. Among them, ERNIE-Health has the highest precision and $F1$ score, while LEBERT performs best on recall. The domain-specific LMs demonstrate a weak advantage over the generic domain LMs. The decent performance of the metrics shows that the NER task for medical dialogue is highly feasible in our settings.

5.2 Dialogue act classification

Task formalization: Dialogue act (DA) directly reflects the fine-grained intention of the speaker. Formally, the goal of DAC task is to identify the DA category of each utterance, i.e. to predict the DA label $a_i \in \mathcal{A}$ for each utterance x_i.

Experimental settings: DAC is a typical text classification task. Baseline models include non-pre-trained models: TextCNN (Chen, 2015), TextRNN (Liu *et al.*, 2016), TextRCNN (Lai *et al.*, 2015) and DPCNN (Johnson and Zhang, 2017), generic domain pre-trained models: BERT (Devlin *et al.*, 2019) and ERNIE (Zhang *et al.*, 2019) and biomedical pre-trained models: MC-BERT (Zhang *et al.*, 2022) and ERNIE-Health (Zhang *et al.*, 2019).

We report four metrics for evaluations, including precision (P), recall (R), $F1$ score ($F1$) and accuracy (Acc).

Experimental results: From the results in Table 5, it can be seen that the pre-trained model has obvious advantages over the traditional neural models in DAC task, and the benefits brought by the domain specific models are slightly weak. The ERNIE-Health model achieves the best results in the classification task, with the classification accuracy of 82.37% achieved. The performance of MC-BERT is not as expected both in NER and DAC task, which may be related to the size and distribution of pre-training data.

5.3 Symptom label inference

Symptom features are the key information to describe the patient's health condition and also the structured training data required for DDP task. The goal of the SLI task is to identify the patient's symptom features from the self-report and the dialogue, and it consists of two cognate sub-tasks: SLI-EXP and SLI-IMP.

Task formalization: The SLI-EXP task aims to find out the patient's self-provided symptoms in self-report x_0, which is called *explicit symptoms*, denoted by $\{s_1, \ldots, s_k\}$, where $s_i \in \mathcal{S}$. In the contrast, the SLI-IMP task aims to find the symptoms and the corresponding labels in the dialogue X, which is called *implicit symptoms*, denoted by $\{(s_{k+1}, l_{k+1}), \ldots, (s_n, l_n)\}$, where $s_j \in \mathcal{S}$ and $l_j \in \mathcal{L}$. Compared with the SLI-EXP task, the SLI-IMP task not only needs to identify symptoms, but also predicts the labels of symptoms. We do not need to predict symptom labels in the SLI-EXP task because symptoms in self-report are always positive.

Table 5. Experimental results for DAC task

Models	$P\uparrow$	$R\uparrow$	$F1\uparrow$	Acc $\uparrow$
TextCNN (Chen, 2015)	74.02	70.92	72.22	78.99
TextRNN (Liu *et al.*, 2016)	73.07	69.88	70.96	78.53
TextRCNN (Lai *et al.*, 2015)	73.82	72.53	72.89	79.40
DPCNN (Johnson and Zhang, 2017)	74.30	69.45	71.28	78.75
BERT (Devlin *et al.*, 2019)	75.35	77.16	76.14	81.62
ERNIE (Zhang *et al.*, 2019)	**76.18**	77.33	76.67	82.19
MC-BERT (Zhang *et al.*, 2022)	75.03	77.09	75.94	81.54
ERNIE-Health (Zhang *et al.*, 2019)	75.81	**77.85**	**76.71**	**82.37**

All the boldface values in Table are significantly at the 5% significance level.

Table 6. Experimental results of symptom recognition in SLI-EXP and SLI-IMP task

Task	Models	Example level			Label level		
		SA $\uparrow$	HL $\downarrow$	HS $\uparrow$	$P\uparrow$	$R\uparrow$	$F1\uparrow$
SLI-EXP	BERT-MLC	73.24	10.10	84.58	86.33	**93.14**	89.60
	MC-BERT-MLC	**75.34**	**9.31**	**85.10**	**88.47**	92.72	**90.54**
SLI-IMP	BERT-MLC	34.16	39.52	82.22	84.98	94.81	89.63
	MC-BERT-MLC	35.14	37.84	82.78	85.41	**95.26**	90.07
	BERT-MTL	37.24	35.32	84.49	96.05	87.04	**91.62**
	MC-BERT-MTL	**37.48**	**34.98**	**85.34**	**95.68**	87.56	91.44

Note: The value of hamming loss (HL) is multiplied by 1e4.

All the boldface values in Table are significantly at the 5% significance level.

Experimental settings: We treat the SLI task as a multi-label classification (MLC) problem, where the label space is $\mathcal{S}$ for SLI-EXP task and $\mathcal{S} \times \mathcal{L}$ for SLI-IMP task. We use BERT (Devlin *et al.*, 2019) and MC-BERT (Zhang *et al.*, 2022) as the encoder and obtain the latent vector of the self-report or the entire conversation, which is then mapped into the label space using an MLP layer. The training objectives are the binary Cross-Entropy loss between sigmoid activations of MLP outputs and actual labels. The model is denoted as BERT-MLC and MC-BERT-MLC.

For SLI-IMP task, we additionally propose a multi-task learning (MTL)-based model (Zhang and Yang, 2018b) called BERT-MTL (or MC-BERT-MTL) that can utilize the BIO labels in NER during training. BERT-MTL has three additional MLP layers on top of BERT (Devlin *et al.*, 2019). The role of these three MLP layers is to predict the BIO label of each token, the normalized name of each symptom entity and the label of each symptom entity. For the first MLP, the input is the hidden vector of each token obtained by BERT encoder, and for the latter two MLPs, the input is the average hidden vector of each symptom entity. The output label space of the three MLP layers are $\mathcal{B}$, $\mathcal{S}$ and $\mathcal{L}$, respectively. The three objectives are trained simultaneously to push the hidden vector of symptom entities to contain more contextual information. We provide the structure diagram of BERT-MTL model in the Supplementary Material.

We evaluate symptom recognition and symptom inference separately. For the evaluation of symptom recognition, we only focus on whether the mentioned symptom entities are found and the

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023

Table 7. Experimental results of symptom inference in SLI-IMP task

Task	Models	POS	NEG	NS	$F1$
SLI-IMP	BERT-MLC	**81.25**	46.53	59.14	62.31
	MC-BERT-MLC	80.80	41.30	58.15	60.08
	BERT-MTL	79.64	**53.87**	**60.20**	**64.57**
	MC-BERT-MTL	80.42	53.15	59.74	64.27

All the boldface values in Table are significantly at the 5% significance level.

symptom labels are ignored. We report two categories of metrics for MLC, including subset accuracy (SA), hamming loss (HL) and hamming score (HS) in example-based metrics, and precision (P), recall (R) and $F1$-score ($F1$) in label-based metrics (Zhang and Zhou, 2014). For symptom inference, we report the macro $F1$ score ($F1$) only for those entities that are correctly predicted, $F1$ scores for each symptom label (*POS, NEG* and *NS*) are also reported.

Experimental results: The performance of the SLI task is listed in Tables 6 and 7. For symptom recognition, the performance of SA shows that the strict prediction of symptoms is very challenging, especially for implicit symptoms from the entire dialogue (only about 37%). It is probably due to the exponential growth of the prediction space, as up to dozens of symptoms can be mentioned in a single dialogue. In non-strict cases, both SLI-EXP and SLI-IMP tasks can achieve good performance, with label-level $F1$ scores reach about 90%. Moreover, the BERT-MTL model that utilizes BIO labels obtains slight better performance in the symptom recognition in SLI-IMP task, which is also intuitive. Compared with BERT, MC-BERT has some weak advantages in symptom recognition.

For symptom inference in SLI-IMP task, MT-based models have an obvious advantage in the identification for the *NEG* and *NS* categories of symptoms, while MC-BERT seems to have no positive effect. It can be seen that inferring the two categories of symptoms is obviously harder than *POS*, since it often requires more contextual information. Especially for the symptoms that appear in the doctor's utterances, it is likely that the patient's response needs to be observed and analyzed to determine the labels. The results suggest that more efforts are needed to improve the symptom inference.

5.4 Medical report generation

Task formalization: The medical report is the summarized patient profile according to the MCR. Formally, the MGR task aims to generate a piece of text $R = \{r_1, \dots, r_m\}$ based on the self-report and the dialogue, where $r_i \in \mathcal{V}$.

Experimental settings: We treat the MGR task as a text-to-text generation problem. The baseline models include: (i) Seq2seq (Nallapati *et al.*, 2016), a LSTM-based encoder–decoder model with attention mechanism; (ii) Pointer-Generator (PG) (See *et al.*, 2017), an improved Seq2Seq model that allows tokens from the source to be directly copied during decoding; (iii) Transformer (Vaswani *et al.*, 2017), the basic model most commonly used in pre-training that based solely on attention mechanisms; (iv) T5 (Xue *et al.*, 2021), a unified Text-to-Text Transformer pre-trained on large text corpus; (v) ProphetNet (Qi *et al.*, 2021), a large-scale pre-trained generative Transformer based on future prediction strategies; and (vi) Bio-ProphetNet, a biomedical generative LM based on ProphetNet that we pre-train on the MedDialog dataset.

We measure model performance on standard metrics of ROUGE scores (Lin, 2004) that widely used for evaluating automatic summarization task, including ROUGE-1/2/L (R-1/2/L). Besides, we also report Concept $F1$ score (C-F1) (Joshi *et al.*, 2020) to measure the model's effectiveness in capturing the medical concepts that are of importance, and Regex-based Diagnostic Accuracy (RD-Acc), to measure the model's ability to judge the disease. To compute C-F1, we use the medical entity extractor (BERT-CRF) trained in our NER task to match entities in the predicted summary to the gold summary, where medical entities in the predicted summary that are not present in the original medical report would be false positives and vice versa for false negatives. For RD-Acc, we use the same regex-based approach mentioned in Section 4.3.

Table 8. Experimental results for MRG task

Models	R-1 ↑	R-2 ↑	R-L ↑	C-F1 ↑	RD-Acc ↑
Seq2seq (Nallapati *et al.*, 2016)	54.15	38.86	50.89	35.46	39.33
PG (See *et al.*, 2017)	57.27	43.41	53.64	43.51	53.51
Transformer (Vaswani *et al.*, 2017)	53.99	39.38	49.78	37.19	45.75
T5 (Xue *et al.*, 2021)	60.97	44.18	57.63	47.35	49.32
ProphetNet (Qi *et al.*, 2021)	60.48	45.73	56.41	49.48	**61.90**
Bio-ProphetNet	**61.83**	**47.12**	**57.48**	**50.12**	61.15

All the boldface values in Table are significantly at the 5% significance level.

Experimental results: The results in Table 8 illustrate that pre-trained generative models can improve the ROUGE scores of medical reports, but the improved R-2 score compared to PG is quite limited. The improvement of the C-F1 score implies a stronger ability of the pre-trained model to capture medical concepts in the dialogue. Despite a high score on Rouge, T5 performs mediocrely on D-Acc. Overall, ProphetNet (Qi *et al.*, 2021) has the best performance, which may benefit from the pre-training on large-scale Chinese corpus and the future prediction strategies during decoding. Although pre-trained models improve the fluency of the generated texts, there are still great challenges in scenarios that are highly dependent on knowledge and reasoning.

5.5 Diagnosis-oriented dialogue policy

Task formalization: Different from the above tasks, the DDP task is dynamic task that requires interaction with the *patient simulator* $\mathcal{P}$. Given the patient's explicit symptoms, the goal of the DDP task is to collect the patient's implicit symptoms and predict the disease, within a given maximum number of interactions with $\mathcal{P}$.

In this article, the patient simulator $\mathcal{P}$ follows the design of Wei *et al.* (2018). It can be treated as a function, given the patient's id, and any symptom $s \in \mathcal{S}$ as input, $\mathcal{P}$ can output the patient's symptom label. An *Unknown (UNK)* label will be returned if the symptom s do not appear in the dialogue.

More specifically, the agent asks $\mathcal{P}$ for one symptom at each step, and after receiving a feedback, asks the next symptom, and repeating above for several turns, until the agent obtains enough information to make diagnosis.

The patient simulator can be designed to be more practical, i.e. the input and output are both natural language texts. In this case, we need a language transmitter to generate the text that conveys the semantics of the action selected by the agent, and a language interpreter based on the proposed understanding framework to parse the patient's response. This situation is one of our future research directions, but it is beyond the scope of this article since our DDP task focuses on policy learning with structured data.

Experimental settings: Baseline models include DQN (Wei *et al.*, 2018), KR-DQN (Xu *et al.*, 2019), REFUEL (Kao *et al.*, 2018), GAMP (Xia *et al.*, 2020) and HRL (Zhong *et al.*, 2022). Except for GAMP, all other methods are based on reinforcement learning (RL). In RL settings, at each turn of interaction, the agent chooses an action from the joint action space of all symptoms and diseases, and correct symptom queries and disease diagnoses are positively rewarded, then the policy can be learned by maximizing the empirical expected cumulative reward. In the contrast, GAMP is a GAN-based method that uses the GAN network to avoid generating randomized trials of symptom, and adds mutual information to encourage the model to select the most discriminative symptoms. We

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023

Table 9. Experimental results for DDP task

Models	SX-Rec ↑	DX-Acc ↑	Avg. # turns
UB-SVM (Noble, 2006)	—	0.706	—
DQN (Wei *et al.*, 2018)	0.047	0.408	9.75
KR-DQN (Xu *et al.*, 2019)	0.279	0.485	6.75
REFUEL (Kao *et al.*, 2018)	0.262	0.505	5.50
GAMP (Xia *et al.*, 2020)	0.067	0.500	1.78
HRL (Zhong *et al.*, 2022)	**0.295**	**0.556**	6.99

All the boldface values in Table are significantly at the 5% significance level.

set the maximum number of interactions between all agents and the patient simulator to 10.

Ideally, if the agent collects all the implicit symptoms, the disease classifier can utilize all the symptom features of the patient. In this case, the performance of disease classification is intuitively the best. We train support vector machine (SVM) (Noble, 2006) classifier with the complete symptoms (both explicit and implicit symptoms), and call the model upper-bound-SVM (UB-SVM). It is an invalid static agent, but its performance can provide a certain reference for dynamic agents.

To evaluate the agent, we report three most concerned metrics, namely symptom recall (SX-Rec), diagnostic accuracy (DX-Acc) and average number of turns (# Turns). Symptom recall measures the agent's ability to find implicit symptoms of the patient, diagnostic accuracy measures the agent's ability in disease classification and average number of turns indicates the efficiency of the diagnostic process.

Experimental results: From the results in Table 9, HRL obtains the best symptom recall and diagnostic accuracy compared to other baselines, with an acceptable average number of turns. HRL groups diseases and works in a combination of master and multiple workers, which is more in line with the actual medical division of labor. However, the symptom recall and diagnostic accuracy of existing models are still far from acceptable levels. It is worth noting that the SVM-UB model can achieve a diagnostic accuracy of 70%, suggesting that the performance of dynamic agents can be expected to be improved if the agents are able to find more implicit symptoms.

6 Conclusions

In this article, we propose a design of frameworks and tasks for automatic medical consultation system to support both static and dynamic medical scenarios. We introduce a new medical dialogue dataset called IMCS-21 with multi-level fine-grained annotations and establish five tasks under the proposed framework. We develop widely used neural-based models for each task and demonstrate experimental results to give an insight about the performance of different tasks. The experimental results show that the validity and potential of the corpus make it expected to be an important benchmark for automated medical consultation systems.

Funding

This work was partially supported by the Natural Science Foundation of China [71991471 and 6217020551]; Science and Technology Commission of Shanghai Municipality [20dz1200600 and 21QA1400600]; and Zhejiang Lab [2019KD0AD01].

Conflict of Interest: none declared.

References

Aho,A.V. and Corasick,M.J. (1975) Efficient string matching: an aid to bibliographic search. *Commun. ACM*, **18**, 333–340.

Banerjee,M. *et al.* (1999) Beyond kappa: a review of interrater agreement measures. *Can. J. Stat.*, **27**, 3–23.

Chen,J. *et al.* (2020) Towards interpretable clinical diagnosis with Bayesian network ensembles stacked on entity-aware CNNs. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020, Online. pp. 3143–3153.

Chen,Y. (2015) Convolutional neural network for sentence classification. Master's Thesis, University of Waterloo.

Devlin,J. *et al.* (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Du,N. *et al.* (2019) Extracting symptoms and their status from clinical conversations. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 915–925.

He,Y. *et al.* (2020) Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 4604–4614.

Johnson,R. and Zhang,T. (2017) Deep pyramid convolutional neural networks for text categorization. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2017, Vancouver, Canada. pp. 562–570.

Joshi,A. *et al.* (2020) Dr. summarize: global summarization of medical dialogue by exploiting local structures. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics, pp. 3755–3763.

Kao,H.-C. *et al.* (2018) Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, EMNLP 2018, Brussels, Belgium, Vol. **32**.

Lai,S. *et al.* (2015) Recurrent convolutional neural networks for text classification. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2015, Austin, Texas, USA*.

Li,D. *et al.* (2021) Semi-supervised variational reasoning for medical dialogue generation. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021*, Online. pp. 544–554.

Li,X. *et al.* (2020) FLAT: Chinese NER using flat-lattice transformer. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020, Online. Association for Computational Linguistics, pp. 6836–6842.

Lin,C.-Y. (2004) Looking for a few good metrics: automatic summarization evaluation-how many samples are enough? In: *NTCIR 2004*, Tokyo, Japan.

Lin,S. *et al.* (2021) Graph-evolving meta-learning for low-resource medical dialogue generation. In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, AAAI 2021, Online. AAAI Press, pp. 13362–13370.

Lin,X. *et al.* (2019) Enhancing dialogue symptom diagnosis with global attention and symptom graph. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), EMNLP 2019, Hong Kong, China*. pp. 5033–5042.

Liu,P. *et al.* (2016) Recurrent neural network for text classification with multi-task learning. In: *IJCAI'16*, New York, US. AAAI Press, pp. 2873–2879.

Liu,W. *et al.* (2021) Lexicon enhanced Chinese sequence labeling using BERT adapter. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL 2021, Bangkok, Thailand. Association for Computational Linguistics, pp. 5847–5858.

Liu,W. *et al.* (2022) MedDG: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In: *CCF International Conference on Natural Language Processing and Chinese Computing*, NLPCC 2022, Guilin, China. Springer, pp. 447–459.

Liu,Y. *et al.* (2017) Using context information for dialog act classification in DNN framework. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark*. pp. 2170–2178.

Nallapati,R. *et al.* (2016) Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: *Proceedings of the 20th*

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023

SIGNLL Conference on Computational Natural Language Learning. Association for Computational Linguistics, Berlin, Germany, pp. 280–290.

Noble,W.S. (2006) What is a support vector machine? *Nat. Biotechnol.*, **24**, 1565–1567.

Qi,W. *et al.* (2021) ProphetNet-X: large-scale pre-training models for English, Chinese, multi-lingual, dialog, and code generation. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, ACL 2021, Bangkok, Thailand. pp. 232–239.

Ramshaw,L.A. and Marcus,M.P. (1999) Text chunking using transformation-based learning. In: *Natural Language Processing Using Very Large Corpora*. Springer, Dordrecht, pp. 157–176.

See,A. *et al.* (2017) Get to the point: summarization with pointer-generator networks. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2017, Vancouver, Canada. pp. 1073–1083.

Shi,X. *et al.* (2020) Understanding medical conversations with scattered keyword attention and weak supervision from responses. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI 2020, New York, USA, Vol. **34**. pp. 8838–8845.

Singh,A.P. *et al.* (2018) Online medical consultation: a review. *Int. J. Community Med. Public Health*, **5**, 1230–1232.

Singhal,T. (2020) A review of coronavirus disease-2019 (COVID-19). *Indian J. Pediatr.*, **87**, 281–286.

Vaswani,A. *et al.* (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems*, NIPS 2017, California, USA, Vol. **30**.

Wei,Z. *et al.* (2018) Task-oriented dialogue system for automatic diagnosis. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL 2018, Melbourne, Australia. pp. 201–207.

Wootton,R. (2001) Telemedicine. *BMJ*, **323**, 557–560.

Xia,Y. *et al.* (2020) Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI 2020, New York, USA, Vol. **34**. pp. 1062–1069.

Xu,L. *et al.* (2019) End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI 2019, Honolulu, Hawaii, USA, Vol. **33**. pp. 7346–7353.

Xue,L. *et al.* (2021) mT5: a massively multilingual pre-trained text-to-text transformer. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL 2021, Online. Association for Computational Linguistics, pp. 483–498.

Zeng,G. *et al.* (2020) MedDialog: large-scale medical dialogue dataset. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP 2020, Online.

Zhang,M.-L. and Zhou,Z.-H. (2014) A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, **26**, 1819–1837.

Zhang,N. *et al.* (2022) CBLUE: a Chinese biomedical language understanding evaluation benchmark. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, pp. 7888–7915.

Zhang,Y. and Yang,J. (2018a) Chinese NER using lattice LSTM. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pp. 1554–1564.

Zhang,Y. and Yang,Q. (2018b) An overview of multi-task learning. *Natl. Sci. Rev.*, **5**, 30–43.

Zhang,Y. *et al.* (2020) MIE: a medical information extractor towards medical dialogues. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020, Online. pp. 6460–6469.

Zhang,Z. *et al.* (2019) ERNIE: enhanced language representation with informative entities. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 1441–1451.

Zheng,Z. *et al.* (2021) Drug package recommendation via interaction-aware graph induction. In: *Proceedings of the Web Conference 2021*, WWW 2021, Ljubljana, Slovenia. pp. 1284–1295.

Zhong,C. *et al.* (2022) Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, **38**, 3995–4001.

Zhou,B. *et al.* (2021) An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL 2021, Bangkok, Thailand. pp. 6214–6224.

Downloaded from https://academic.oup.com/bioinformatics/article/39/1/btac817/6947983 by guest on 08 March 2023