

基于 BERT 的中文医疗问答系统^①

王志明, 郑 凯

(华南师范大学 计算机学院, 广州 510631)
通信作者: 郑 凯, E-mail: david@scnu.edu.cn



摘 要: 现如今, 互联网中存在海量的医疗领域知识可以用于医疗病情诊断, 但传统的搜索引擎并无法根据病人的实际情况做出合理的判断, 无法满足使用需求. 因此, 本文主要开发基于知识图谱问答系统. 该系统面向医疗领域, 采用爬虫技术获取了大量医疗数据并将其存储在 Neo4j 图数据库构建医疗知识图谱中. 同时, 为了使系统能够进一步理解用户的医疗询问问句, 本文提出了基于 BERT 以及 BERT-BiLSTM-CRF 模型分别用于识别问句中的意图信息和实体信息的方法. 最后, 系统利用意图和实体信息在知识图谱中进行查询并为用户提供合适的回答, 完成了医疗问答系统的构建.

关键词: BERT; 知识图谱; 意图识别; 槽位填充

引用格式: 王志明, 郑凯. 基于 BERT 的中文医疗问答系统. 计算机系统应用, 2023, 32(6): 115-120. <http://www.c-s-a.org.cn/1003-3254/9140.html>

Chinese Medical Question Answering System Based on BERT

WANG Zhi-Ming, ZHENG Kai

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Nowadays, a large amount of medical domain knowledge on the Internet can be used for medical diagnosis, but traditional search engines cannot make reasonable judgments based on the actual situation of patients and fail to meet the needs of use. Therefore, this study mainly develops a question-answering system based on a knowledge graph. The system is applied to the medical field, which uses crawler technology to obtain a large amount of medical data and stores them in the constructed medical knowledge graph of the Neo4j graph database. At the same time, in order to enable the system to further understand the user's medical questions, this study proposes methods based on BERT and BERT-BiLSTM-CRF models for identifying intent information and entity information in questions, respectively. Finally, the system uses the intent and entity information to make a query in the knowledge graph and provides users with appropriate answers, thus completing the construction of a medical question-answering system.

Key words: BERT; knowledge graph (KG); intent recognition; slot filling

1 引言

目前, 我国的人工问诊制度还存在着许多问题. 主要体现在 3 个方面: 一是城乡医疗资源不平衡, 欠发达地区的医疗基础设施和专业医护人员相比发达地区还有一定的差距; 二是我国是一个人口大国, 目前的医疗资源还不足以满足所有人的医疗需求; 三是互联网的

医疗问答工作混乱, 信息来源不明. 且对于一些无需大费周章到医院问诊的小病, 例如感冒、头晕, 在线问答往往会夸大其词, 造成患者不必要的恐慌.

另外, 受限于中文自身的语言特性, 对医疗问句的理解有限成了中文医疗问答系统主要的性能瓶颈. 在这一方面, 近几年来也有学者进行过相关的研究: Jiang

^① 收稿时间: 2022-10-10; 修改时间: 2022-11-14; 采用时间: 2023-01-19; csa 在线出版时间: 2023-04-20
CNKI 网络首发时间: 2023-04-21

等人^[1]提出了一种字符串匹配的方式来对问句进行分类, Zhou 等人^[2]采用 TF-IDF 和贝叶斯算法对问句进行理解和分类, Gao 等人^[3]采用 TF-IDF 结合 Word2Vec 的方式理解中文问句. 但以上的研究中, 各模型对于中文语言的理解还不够深入, 且研究方向主要集中在问句分类上, 很少利用问句中的实体信息来增强对问句的理解.

基于上述问题, 本文从专业可靠的医疗网站上获取医疗信息, 构建了基于知识图谱的智能问答系统, 并利用富含语义信息的 BERT 模型, 结合意图识别和槽位填充等自然语言技术进一步加强了问答系统对用户问句的理解, 使得问答系统的回复更加精准有用. 本文的贡献总结如下.

(1) 利用爬虫技术, 从互联网上专业可靠的医疗平台上获取大量结构化数据, 并利用这些数据, 构建了一个医疗知识图谱, 并进一步构建了医疗智能问答系统.

(2) 提出了基于 BERT 的医疗意图分类模型, 用于准确识别出用户输入问句中的意图信息.

(3) 提出了基于 BERT+BiLSTM+CRF 的槽位填充模型, 用于准确识别出用户输入问句中的实体信息.

2 相关研究

自然语言理解 (natural language understanding, NLU) 是自然语言处理下的一个任务分支, 其目的在于研究如何使机器理解人类自然语言. 近些年来对于这一问题的研究层出不穷, 其主要的研究思路是将自然语言理解任务分解为意图识别和槽位填充两个子任务. 其中, 意图识别的目的是确定句子意图, 而槽位填充的目的是获取句子实体信息.

意图识别 (intent recognition) 通常被认为是文本分类问题, 因此其解决思路也是围绕着这一点展开的. Ravuri 等人^[4]提出了使用 RNN 和 LSTM 网络用于获取丰富的句子上下文关系从而进行意图识别. Hashemi 等人^[5]则提出了使用 CNN 网络获取句子中的卷积特征进行意图识别. 而 Firdaus 等人^[6]更是对 LSTM、CNN、GRU 等模型的集成模型如何用于意图识别进行了研究. 随着预训练模型的兴起, 使用 GloVe、Word2Vec、BERT 等预训练模型先进行词嵌入, 后应用于下游任务的方式也逐渐证明了其优越性^[7].

与意图识别任务类似, 槽位填充 (slot filling) 也通常被认为是序列标注任务中的命名实体识别任务

(named entity recognition, NER). Mesnil 等人^[8]和 Vu^[9]分别探索了如何利用 RNN 和 CNN 模型处理槽位填充任务, 也都取得了不错的效果. Louvan 等人^[10]则提出了一种更加高效地解决槽位填充的模型: BiLSTM-CRF, 同时该模型也被广泛用于中文相关数据集^[11].

3 基于 BERT 的中文医疗问答系统构建

本文构建的医疗问答系统主要分为 3 个模块, 分别是: 知识图谱模块, 意图识别模块, 槽位填充模块. 如图 1 所示, 在用户的医疗问句输入系统后, 系统将通过意图识别和槽位填充模块识别出问句中的意图信息和实体信息. 接着将这些信息代入到知识图谱中进行查询并最终将相应的答案输出. 下面我们将分别阐述这 3 个模块的具体实现.

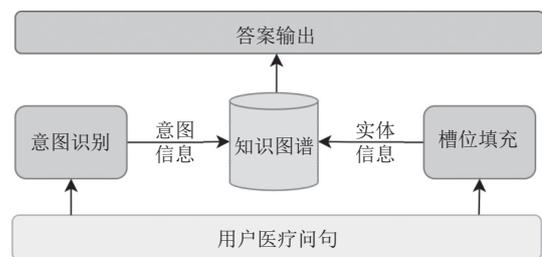


图 1 医疗问答系统简要流程图

3.1 构建知识图谱

知识图谱作为一种基于图的数据结构, 也是一种揭示实体之间关系的语义网络, 其通过节点和边的相互搭配组合可以以结构化的方式描述客观世界中的概念、实体及属性关系, 因此被广泛用于问答系统中^[12]. 与此同时, 构建知识图谱需要获取大量结构化数据, 因此本文利用爬虫技术, 从国内可靠的医疗数据网站——寻医问药网 (<http://www.xywy.com/>) 上爬取了大量的医疗数据用于构建本系统的知识图谱.

对于从网站上抓取至本地的原始数据, 我们利用 XPATH 和正则表达式等文本处理技术, 按照网页排版规律, 依次提取其中的实体、属性、关系信息, 并将文本中含有的“\n”“\t”等特殊字符一一去除, 最终以 JSON 格式的文件形式存储在本地, 实现了从非结构化数据到结构化数据的转换.

最后, 我们一共获取到了 5 类实体和 5 类关系. 其中, 每个疾病实体含有 14 种属性, 且疾病实体是其他实体之间的中心实体, 也是关系数量最多的实体. 在本

文系统构建的知识图谱中,一共有 30 402 个节点以及 206 873 对关系,其具体的实体、属性、关系信息如表 1 所示.

表 1 实体、属性、关系信息表

类型	包含内容
实体	疾病、药物、症状、菜谱、检查项目
属性	名称、描述、病因、预防措施、诊断科室、治疗方式、治愈时间、治愈率、费用、易感人群、发病率、推荐饮食、饮食禁忌、药物功能、制作方式、使用方法、单价
关系	并发疾病、相关症状、检查项目、推荐食谱、推荐药物

对于知识图谱的存储方式,本文采用了开源图形数据库 Neo4j 存储医疗知识图谱. Neo4j 提供了功能强大的 Cypher 语言,支持在知识图谱中进行复杂的关系查询. 我们利用 Neo4j 为我们获取到的每一个实体构建节点,并将其对应的属性信息存储到节点中,接着再根据实体间的关系信息,创建对应边连接实体. 通过这种方式,我们完成了知识图谱的构建和存储,图 2 展示了本系统中关于阿洪病部分知识在图数据库中的存储.

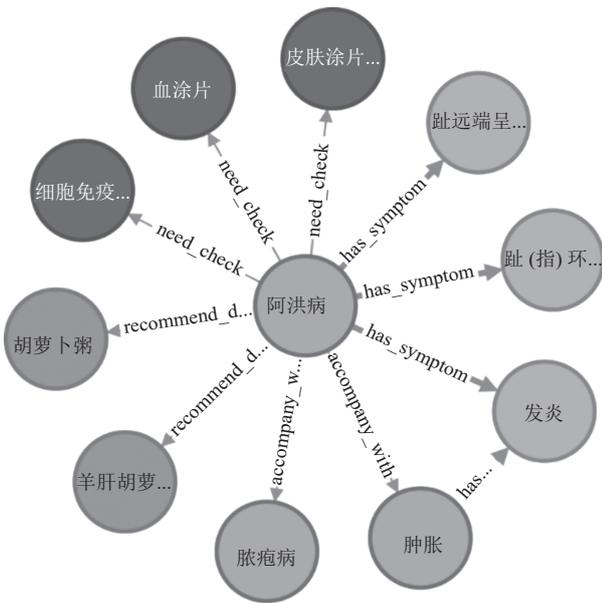


图 2 阿洪病在图数据库中的存储

3.2 意图识别

在系统获得用户的文本输入后,系统需要正确识别出用户的查询意图,才能继续查询相应的结果. 本文提出使用 BERT 模型来解决医疗问答系统中的意图识别任务,相比于以 Word2Vec 和 GloVe 为代表的静态词向量方法,作为动态词向量的 BERT 在预训练过程中充分计算了文本的上下文表示,可以解决静态词向

量无法解决的一词多义问题. 同时,通过 BERT 表示的文本,其向量稠密度也更高,语义信息的表示能力也更强.

BERT 处理问句意图分类问题的过程如图 3 所示. 对于一个问句文本序列 $x_1 x_2 \cdots x_n$, 在其前后加入特殊标记后得到 X , 其中 [CLS] 表示文本序列开始标记, [SEP] 表示文本序列之间的分隔标记.

$$X = [\text{CLS}] x_1 x_2 \cdots x_n [\text{SEP}] \quad (1)$$

接下来将 X 输入到 BERT 模型后得到句子的上下文语义表示 $h \in \mathbb{R}^{N \times d}$, 其中 N 表示句子长度, d 表示 BERT 的隐藏层维度.

$$h = \text{BERT}(X) \quad (2)$$

BERT 在文本分类任务中使用的是 [CLS] 标签对应的隐含层 h_0 , 即 h 的首个分量. 而为了预测输入文本对应的分类标签,需要将 h_0 接入一个全连接层并按照下式计算概率分布 $P \in \mathbb{R}^K$:

$$P = \text{Softmax}(h_0 W^o + b^o) \quad (3)$$

其中, K 表示分类标签数, $W^o \in \mathbb{R}^{d \times K}$ 表示全连接层的权重, $b^o \in \mathbb{R}^K$ 表示全连接层的偏置.

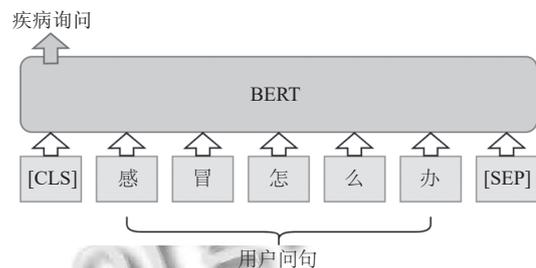


图 3 BERT 意图识别模型结构

3.3 槽位填充

在确定了用户输入文本的意图后,下一步要确定的就是其具体的实体信息,以便系统在知识库中查询具体的知识信息,而识别实体信息就是槽位填充任务的主要内容,因此本文提出使用 BERT-BiLSTM-CRF 模型来解决该问题. 如图 4 所示,该模型分为 3 层: BERT 层、BiLSTM 层、CRF 层. 输入文本在经过 BERT 模型获取了文本表示后,利用 BiLSTM 学习文本时序特征信息,最后再经过 CRF 层得到全局最优序列.

BERT 层与意图识别类似,其主要作用是将输入文本通过式 (1) 和式 (2) 后得到其上下文语义表示 h . 不同的是,槽位填充任务需要使用到完整的 h 作为 BiLSTM 层的输入.

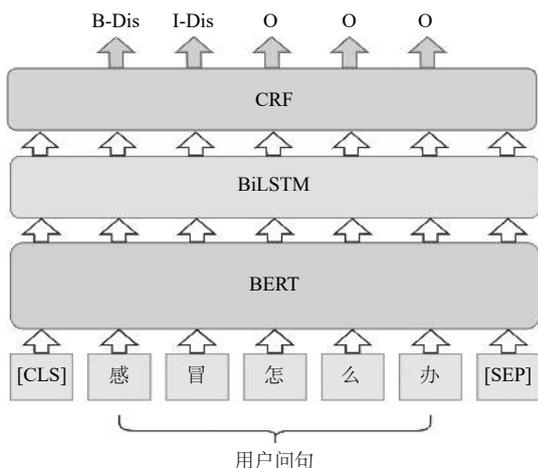


图4 BERT-BiLSTM-CRF 槽位填充模型结构

BiLSTM 表示双向的 LSTM 网络, 可以获得更全面的时序特征信息. 双向 LSTM 将前向向后向 LSTM 网络应用于每个训练序列, 并将两个 LSTM 网络连接到同一个输出层. 如图 5 所示, LSTM 计算单元包括 3 个门结构: 输入门、遗忘门和输出门. 具体的计算过程如下公式所示:

$$\begin{cases} f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \\ o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \\ u_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \\ c_t = f_t \odot c_{t-1} + i_t \odot u_t \\ h_t = o_t \odot \tanh(c_t) \end{cases} \quad (4)$$

其中, w_t, w_f, w_o 和 b_t, b_f, b_o 分别表示 LSTM 网络中的权重矩阵和偏置, i_t, f_t, o_t 分别表示输入门、遗忘门、输出门, x_t 和 h_t 分别表示 t 时刻的输入向量和隐藏层状态. u_t 和 c_t 则表示当前输入的单元状态和当前时刻的单元状态. 最终, 根据输出门和单元状态可以确定 LSTM 的最终输出 h_t . 对于双向 LSTM 网络, 我们用 \vec{h} 表示前向 LSTM 的输出结果, 用 \overleftarrow{h} 表示后向 LSTM 的输出结果, 将 \vec{h} 和 \overleftarrow{h} 拼接一起即可获得双向 LSTM 的最终输出 \vec{h} .

$$\begin{cases} \vec{h} = LSTM(x_t, \vec{h}_{t-1}) \\ \overleftarrow{h} = LSTM(x_t, \overleftarrow{h}_{t-1}) \\ \vec{h} = [\vec{h}, \overleftarrow{h}] \end{cases} \quad (5)$$

由于 LSTM 只能考虑句子的上下文信息, 不能考虑标记之间的依赖关系, 而 CRF 可以通过学习标记之间的依赖关系来确保标记有效, 因此我们在 BiLSTM 层后引入 CRF 层以弥补 BiLSTM 的缺陷. CRF 层使用以下公式用于对输入序列 $X = (x_1, x_2, \dots, x_n)$ 计算其预

测序列 $Y = (y_1, y_2, \dots, y_n)$ 的分数函数, 并使用极大似然估计法预测序列 Y 的概率. 其中, $P_{i,j}$ 表示第 i 个词的第 j 个标签的分数, $W_{i,j}$ 表示从 i 标签转移 j 标签的分数.

$$\begin{aligned} score(X, Y) &= \sum_{i=1}^n (P_{i,y_i} + W_{y_i, y_{i+1}}) \\ P(Y|X) &= \frac{\exp(score(x, y))}{\sum_{y'} \exp(score(x, y'))} \end{aligned} \quad (6)$$

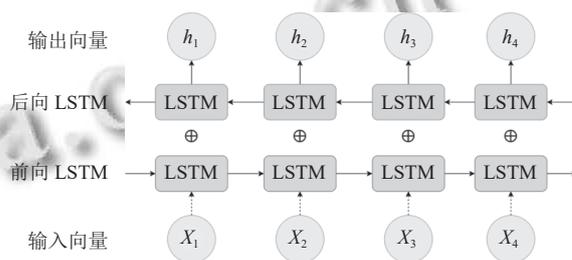


图5 BiLSTM 结构图

4 实验与结果分析

4.1 实验数据集选择

为了评估我们提出的模型, 我们分别使用 CBLUE^[13] 中的 QIC 数据集和 CCKS 2017 Task 2 的数据集来对我们的意图识别和槽位填充模型进行实验. 表 2 为两个数据集的数据情况.

表2 数据集数据情况

数据集	训练集	测试集	类别数
QIC	6931	1955	10
CCKS 2017	29866	9493	5

QIC 为医疗意图问句分类数据集, 包含了疾病表述、治疗方案、就医建议、注意事项等 11 种意图. 由于其数据量分布不均, 因此本文使用了前 10 种意图进行实验.

CCKS 2017 Task 2 为医学电子病历命名实体识别数据集, 包含了疾病、症状、检查项目、治疗方式和身体部位这 5 类实体标记. 该数据集采用了 BIO 标记方式, 其中 B 表示字符属于实体的开始位置, I 表示字符属于实体的其他位置, O 表示字符不属于任何实体.

4.2 评价指标

本文采用精确率 P , 召回率 R 和 $F1$ 值来评判意图识别和槽位填充两个模型的性能, 各评价指标的计算方法如下:

$$\begin{cases} P = \frac{\text{正确识别的标签个数}}{\text{识别出的所有标签个数}} \times 100\% \\ R = \frac{\text{正确识别的标签个数}}{\text{所有标签个数}} \times 100\% \\ F1 = \frac{2 \times P \times R}{P + R} \times 100\% \end{cases}$$

4.3 实验设置与结果分析

在我们的实验中,我们采用了 BERT Base 预训练模型,该模型包含了 12 层的编码器,768 个隐藏单元,12 个注意力,110M 个参数。另外,我们使用 PyTorch 搭建了我们提出的模型。对于意图分类模型,我们设置序列长度为 45,而对于槽位填充模型,我们设置序列长度为 128,隐藏层大小为 128。两个模型的学习率均为 $1E-5$,训练次数为 100。

为了评估我们提出的模型的性能,我们将我们的模型与几个基线模型进行了比较。对于意图分类模型,我们对比了几种经典的文本分类模型,包括 Fasttext、LSTM、Text-CNN。而对于槽位填充模型,我们将我们的模型与 LSTM、LSTM-CRF、BiLSTM、BiLSTM-CRF 模型对比。表 3 和表 4 显示了我们提出的模型和其他基线模型在 QIC 和 CCKS2017 数据集上的实验结果。

表 3 QIC 数据集实验结果 (%)

模型	P	R	F1
Fasttext	70.84	62.15	66.21
LSTM	72.72	68.63	70.61
Text-CNN	73.40	70.23	71.78
BERT	82.35	80.96	81.65

表 4 CCKS2017 数据集实验结果 (%)

模型	P	R	F1
LSTM	75.59	78.29	76.92
LSTM-CRF	83.59	85.28	84.42
BiLSTM	82.42	84.67	83.53
BiLSTM-CRF	84.69	87.31	85.98
BERT-BiLSTM-CRF	86.51	90.48	88.45

从表 3 可以看出,LSTM 和 Text-CNN 模型在意图识别任务上表现效果相近,而基于预训练模型 BERT 的意图识别效果要明显好于其他的意图识别模型。从表 4 可以观察到,双向 LSTM 网络比单向的 LSTM 模型在识别实体效果上有优势,且使用了 CRF 层后的模型准确率也会比使用前有一定的提升。而我们提出的综合了预训练模型 BERT、双向 LSTM 网络和 CRF

层的模型在各方面指标上都优于其他模型。

实验证明,我们提出的两个模型对比其他基线模型效果显著。这说明了我们的两个模型能够更好地表示单词的语义信息,能够很好地应对医疗领域问答系统的意图识别和槽位填充任务。

5 系统优化

由于数据集中的医疗信息并不能完全覆盖知识图谱中存储的医疗信息,为了提高医疗问答系统的问答能力,我们将知识图谱中未能覆盖到的实体与数据集中相同类别的实体进行了替换并生成了新的数据。我们利用这些数据将我们的意图识别和槽位填充模型进行二次训练,使得模型的识别能力有了一定的增强,能够与我们构建的知识图谱中的医疗信息相适应。

另外,我们搭建了一个简单的图形交互界面,以供用户使用,系统的使用界面如图 6 所示。



图 6 医疗问答系统图形化界面

6 结论与展望

本文的主要工作是基于互联网上可靠的医疗专业知识,构建了一个高质量的医疗知识图谱。基于此知识图谱,我们提出了基于 BERT 的意图识别模型和基于 BERT-BiLSTM-CRF 的槽位填充模型来改进传统的医疗问答系统,使系统对用户的问句文本理解更深入,返回的答案更加精准。另外,我们还针对公开数据集和自建知识图谱信息覆盖不充分的问题做了一定的优化,

同时也设计了图形化界面以供用户进行使用。最终,我们成功构建了一个可靠的中文医疗问答系统,该系统能够帮助解决人们对日常疾病相关知识的需求,具有很高的应用价值。值得注意的是,本文提出的系统构造方法具有一般性,可用于构建其他垂直领域的知识图谱问答系统,对其他领域基于知识图谱的问答系统研究具有一定的参考价值。在未来的工作中,我们将进一步扩展知识图谱的内容以及增强问答系统的问答能力,以提高系统回答的性能。

参考文献

- 1 Jiang ZX, Chi CY, Zhan YY. Research on medical question answering system based on knowledge graph. *IEEE Access*, 2021, 9: 21094–21101. [doi: [10.1109/ACCESS.2021.3055371](https://doi.org/10.1109/ACCESS.2021.3055371)]
- 2 Zhou CY, Guan RC, Zhao CT, *et al.* A Chinese medical question answering system based on knowledge graph. *Proceedings of the 15th IEEE International Conference on Big Data Science and Engineering (BigDataSE)*. Shenyang: IEEE, 2021. 28–33.
- 3 Gao RX, Li C. Knowledge question-answering system based on knowledge graph of traditional Chinese medicine. *Proceedings of the 9th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. Chongqing: IEEE, 2020. 27–31.
- 4 Ravuri SV, Stolcke A. Recurrent neural network and LSTM models for lexical utterance classification. *Proceedings of the 16th Annual Conference of the International Speech Communication Association*. Dresden: INTERSPEECH, 2015. 135–139.
- 5 Hashemi HB, Asiaee A, Kraft R. Query intent detection using convolutional neural networks. *Proceedings of the 2016 International Conference on Web Search and Data Mining, Workshop on Query Understanding*. 2016. 1–5.
- 6 Firdaus M, Bhatnagar S, Ekbal A, *et al.* Intent detection for spoken language understanding using a deep ensemble model. *Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence*. Nanjing: Springer, 2018. 629–642.
- 7 Korpusik M, Liu Z, Glass JR. A comparison of deep learning methods for language understanding. *Proceedings of the 20th Annual Conference of the International Speech Communication Association*. Graz: Interspeech, 2019. 849–853.
- 8 Mesnil G, Dauphin Y, Yao KS, *et al.* Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, 23(3): 530–539. [doi: [10.1109/TASLP.2014.2383614](https://doi.org/10.1109/TASLP.2014.2383614)]
- 9 Vu NT. Sequential convolutional neural networks for slot filling in spoken language understanding. *Proceedings of the 17th Annual Conference of the International Speech Communication Association*. San Francisco: Interspeech, 2016. 3250–3254.
- 10 Louvan S, Magnini B. Exploring named entity recognition as an auxiliary task for slot filling in conversational language understanding. *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. Brussels: ACL, 2018. 74–80.
- 11 Zhang WT, Jiang SH, Zhao S, *et al.* A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition. *Proceedings of the 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. Xiangtan: IEEE, 2019. 166–169.
- 12 郑泳智, 朱定局, 吴惠淼, 等. 知识图谱问答领域综述. *计算机系统应用*, 2022, 31(4): 1–13. [doi: [10.15888/j.cnki.csa.008418](https://doi.org/10.15888/j.cnki.csa.008418)]
- 13 Zhang NY, Chen MS, Bi Z, *et al.* CBLUE: A Chinese biomedical language understanding evaluation benchmark. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin: ACL, 2022. 7888–7915.

(校对责编: 牛欣悦)