情报学报　2021 年 11 月　第 40 卷　第 11 期
Journal of the China Society for Scientific and Technical Information, Nov. 2021, 40(11): 1234-1244

DOI: 10.3772/j.issn.1000-0135.2021.11.009

基于多任务学习和多态语义特征的中文疾病名称归一化研究

韩　普[1,2]，张展鹏[1]，张　伟[1]

（1. 南京邮电大学管理学院，南京　210003；2. 江苏省数据工程与知识服务重点实验室，南京　210023）

摘　要　为解决在线文本中存在大量疾病指称的问题，提出了基于多任务学习和多态语义特征的中文疾病名称归一化模型（multi-task attention-dictionary BERT GRU-CNN，MTAD-BERT-GCNN）。首先利用word2vec和Glove生成融合局部和全局的外部语义特征向量；其次将CNN（convolutional neural networks）和BERT（bidirectional encoder representations from transformers）作为基准模型进行对比实验；接着在CNN上引入GRU（gated recurrent unit）、LSTM（long short-term memory）、BiGRU（bi-directional gated recurrent unit）和BiLSTM（bi-directional long short-term memory）以提取文本间语义关系；然后，基于多任务学习视角，将上述模型与BERT相结合以捕获静态和动态语义信息；最后，引入医学词典生成注意力权重词典作为辅助任务以调节静态向量，从而进一步提升模型效果。在自建的中文疾病名称归一化数据集ChDND（Chinese disease normalization data）上进行实验。研究结果发现，MTAD-BERT-GCNN模型在Accuracy@10指标上可以达到89.60%的准确率，较基础的词级CNN和字级CNN分别提高了12.96%和5.12%。本研究在中文疾病名称归一化任务中引入了多任务学习思路，从语义向量和模型框架层面进行了优化，在中文医学知识图谱构建、信息抽取和自然语言理解中具有较好的应用价值。

关键词　疾病名称归一化；有监督学习；多任务学习；卷积神经网络；BERT

Chinese Disease Name Normalization Based on Multi-task Learning and Polymorphic Semantic Features

Han Pu[1,2], Zhang Zhanpeng[1] and Zhang Wei[1]

(1. *School of Management, Nanjing University of Posts & Telecommunications, Nanjing*　210003;
2. *Jiangsu Provincial Key Laboratory of Data Engineering and Knowledge Service, Nanjing*　210023)

Abstract: In order to solve the problem of a large number of disease designations in online texts, a Chinese disease name normalization model based on multi-task learning and polymorphic semantic features (multi-task attention-dictionary BERT GRU-CNN, MTAD-BERT-GCNN) is proposed. First, word2vec and Glove were used to generate external semantic feature vectors that integrate local and global semantics. Second, CNN and BERT were used as benchmark models for comparative experimental analysis. Third, GRU, LSTM, BiGRU, and BiLSTM were introduced on CNN to extract semantic relationships between texts. Next, from the perspective of multi-task learning, the above model was combined with BERT to capture static and dynamic semantic information. Finally, the medical dictionary was introduced to calculate the attention matrix as an auxiliary task to adjust the static vector, thereby further improving the model effect. Our experiments were carried out using the self-built Chinese disease name normalization dataset, ChDND. The experimental results found

收稿日期：2020-11-23；**修回日期**：2021-05-25

基金项目：国家社会科学基金项目“大数据环境下健康领域实体语义挖掘研究”（17CTQ022）。

作者简介：韩普，男，1983年生，博士，副教授，硕士生导师，主要研究领域为医疗健康语义分析，E-mail：hanpu@njupt.edu.cn；张展鹏，男，1996年生，硕士研究生，主要研究领域为实体归一化；张伟，男，2000年生，本科生，主要研究领域为自然语言处理。

that the MTAD-BERT-GCNN model achieved 89.60% accuracy on the Accuracy@10, which is higher than the basic word-level CNN, and the word-level CNN increased by 12.96% and 5.12%, respectively. This research introduces the concept of multi-task learning in the normalization task of Chinese disease names and optimizes it from the level of the semantic vector and model framework, which has good application value in the construction of Chinese medical knowledge graphs, information extraction, and natural language understanding.

Key words: disease name normalization; supervised learning; multi-task learning; convolutional neural network; bidirectional encoder representation from transformers (BERT)

1　引　言

近年来，随着互联网的飞速发展和公众信息素养的提升，微博、微信和在线健康社区等社会化媒体逐渐成为人们获取、传播和分享医疗健康知识的重要渠道，这些平台所产生的海量在线医疗健康数据已经成为医疗实体识别[1-2]、流行病预测[3-4]、情感分析[5-6]和药物不良反应[7-8]等多个研究的重要数据源。与电子病历中的专业化表述相比，在线医疗健康文本缺乏医疗术语规范，存在大量的疾病指称和口语化表达，这对在线医疗健康信息抽取和知识挖掘带来了极大的挑战。在这种背景下，将用户的非标准化表述映射到标准医学术语的疾病名称归一化任务[9-10]，受到了医疗健康信息抽取、知识库和知识图谱构建以及领域知识挖掘的重点关注[11-12]，目前已经成为自然语言处理和信息抽取中的一个重要研究领域。

疾病名称归一化任务的主要挑战表现是在线医疗健康文本中疾病指称与标准术语往往并没有字面上的关联，基于规则的方法难以从字符层面实现归一化；另外，在线医疗健康文本中的疾病指称与标准术语存在一对多或多对多等复杂关系，传统方法难以挖掘深层语义信息。与英文相比，中文文本表达方式和语法结构更为复杂，词汇间无分隔符号，一词多义和同形异义的现象较为普遍，导致语义分析的难度更大[13]。另外，中文疾病名称构词更为复杂，存在大量缩写和翻译词汇，也缺少类似于UMLS（unified medical language system）和SNOMED CT（the systematized nomenclature of human and veterinary medicine clinical terms）的疾病名称知识库资源[14-15]，使得中文疾病名称归一化面临着更大的挑战。与通常的术语相比，中文疾病名称专业性更强，尤其是在线医疗健康社区中不同用户的表述多种多样，并且有许多名称是从外文翻译而来，这些因素导致中文疾病名称归一化难度也远大于普通的术语标准化。

本研究基于多任务学习视角，将CNN（convolutional neural networks）、GRU（gated recurrent unit）、LSTM（long short-term memory）、BiGRU（bidirectional gated recurrent unit）、BiLSTM（bi-directional long short-term memory）与BERT（bidirectional encoder representations from transformers）相结合，以捕获静态和动态语义信息；同时引入注意力权重词典作为辅助任务生成注意力矩阵以调节静态向量，并将疾病名称归一化转化为分类任务；最后在中文数据集上进行实验，以验证多任务学习对中文疾病名称归一化的效果。

2　相关研究概述

根据所采用的研究方法，疾病名称归一化可以分为无监督学习和有监督学习。在有监督学习方法中，多任务学习和BERT是学界近期的关注重点。

2.1　无监督学习

无监督学习方法主要是指采用字典查找或字符串匹配的方法进行归一化。Ristad等[16]利用编辑距离计算字符串间的相似度，将归一化任务转化为相似度排序问题。2010年，美国医学图书馆年发布了MetaMap工具[17]，它首先通过词典遍历和浅层句法分析来识别名词短语，然后将生物医学文本与UMLS的CUIs建立映射关系。Tsuruoka等[18]利用逻辑回归计算字符串相似度以实现归一化，其效果优于传统的规则匹配方法。Yang[19]从UMLS和SNOMED CT中提取了疾病相关特征，并改进了基于规则的归一化方法。基于MetaMap工具，Khare等[20]建立了疾病和药物的映射关系，并将药物描述中的疾病作为候选名称，结果表明该方法在疾病名称归一化上可达到较好的效果。基于UMLS中的疾病变体规则，Kate[21]提出了自动学习临床术语变体的模型，从而对未包含在知识库中的术语进行归一化。Jonnagaddala等[22]提出了基于词典查找的方法进行疾病名称归一化，并引入同义词增强词典以进

一步提升实验效果。通过上述分析可知，一方面，传统的无监督学习方法依赖权威的医学词典或知识库，难以应对未收录疾病和疾病指称的情况；另一方面，该方法主要利用语言形态信息进行处理，难以结合深层语义信息进行疾病名称归一化。

2.2 有监督学习

有监督学习方法主要是指利用机器学习或深度学习模型进行任务分类的方法，该方法往往将疾病描述文本与疾病名称匹配视为文本分类任务，通过模型学习疾病描述特征表示以预测疾病分类，从而实现疾病名称归一化。基于成对学习思想，Leaman利用机器学习模型，构建了英文疾病名称归一化系统DNorm（disease name normalization）[10]。该系统利用计算相似度矩阵预测疾病描述文本与候选疾病名称的关系，其*F*值在NCBI（National Center for Biotechnology Information）疾病数据集实验中较MetaMap提升了25%。Shi等[23]利用字符级感知神经网络学习书面诊断描述和ICD（international cassification of diseases）编码的隐藏表示，并引入注意力机制，实现了书面诊断与ICD编码的归一化映射。Liu等[24]利用word2vec和TreeLSTM生成了分布式特征表示并提取候选疾病名称，通过计算疾病描述和候选疾病名称间相似度进行分类，在英文数据集上取得了较好的实验结果和较高的鲁棒性。通过学习文本内在语义关系，Limsopatham等[25]发现CNN在疾病名称归一化上的效果优于RNN（recurrent neural network），其实验准确率较DNorm高出13.79%。基于形态和语义信息，Li等[26]通过CNN计算疾病指称和候选疾病名称的语义相似度实现了生物医学概念归一化，实验结果明显优于基于规则的方法，验证了引入语义特征可提高疾病名称归一化效果。Tutubalina等[27]提出了基于注意机制的双向LSTM及GRU，并引入UMLS的TF-IDF（term frequency-inverse document frequency）特征和语义相似性特征，进一步验证了语义特征对疾病名称归一化的影响。Huang等[28]基于RNN和CNN实现了MIMIC-III（medical information mark for intensive care）数据集到ICD编码的映射，研究结果验证了RNN和CNN较传统的逻辑回归和随机森林等模型的疾病名称归一化效果均有明显提升。

与无监督学习相比，有监督学习不但弥补了无监督学习中无法处理未收录疾病名称的不足，而且通过大规模训练数据学习疾病特征，可充分利用文本语义信息进行疾病名称归一化。

2.3 多任务学习

多任务学习可联合训练多个子任务，通过共享参数提高模型的学习效率和泛化能力，近期在自然语言处理领域受到了学界的重点关注。Collobert等[29]在词性标注、命名实体识别和语义角色标注等任务中，提出了基于多任务学习的CNN模型，验证了多任务学习在自然语言处理上的优异表现。Liu等[30]基于LSTM设计了三种信息共享机制，使用特定任务的共享层对文本进行建模，研究发现子任务可以提升主分类任务效果；另外，Liu等[31]还在文本分类中提出对抗性的多任务学习框架，避免了共享和私有两种特征的相互干扰，实验结果表明所学习的共享知识可被迁移到新任务中。Yang等[32]以ELMo（embeddings from language models）作为向量嵌入提出了基于注意力的多任务BiLSTM-CRF模型，在电子病历数据集上进一步提升了医疗实体识别和归一化效果。Niu等[33]基于多任务学习思路提出了字符级CNN模型进行疾病名称归一化，较好地解决了未登录词的问题，并引入注意力机制优化模型效果，实验结果在AskApatient数据集上达到了84.65%的准确率。由上文可知，在自然语言处理任务的不同应用场景中，多任务学习得到了广泛的应用。本文将多任务学习思想引入中文疾病名称归一化研究中，利用多任务学习能够共享多个子任务间参数以共同提升主任务的优势，进一步推动中文疾病名称归一化研究进展。

2.4 BERT

BERT[34]是一种基于转换器的双向编码表征模型，在多个自然语言处理任务中表现优异[35-36]。Li等[37]对大规模标注的电子健康档案进行了BioBERT微调，进一步训练了EhrBERT、BioBERT和BERT，研究结果发现，这些模型在疾病名称归一化上的效果均优于DNorm。Xu等[38]基于BERT设计了列表分类器并利用正则化UMLS语义类型对候选概念进行排序，在疾病名称归一化上达到了较高的准确率。Ji等[39]基于微调预训练的BERT、BioBERT和ClinicalBERT进行疾病名称归一化，在ShARe/CLEF、NCBI和TAC2017ADR三种不同类型数据集上的实验均表明微调模型明显优于基线方法。此外，Kalyan等[40]提出了一种基于BERT和Highway的医学概念

标准化系统，研究发现在CADEC和PsyTAR数据集上的效果优于传统方法。本文基于多任务视角，结合当前主流的BERT模型，综合利用文本形态信息和深层语义信息进行中文疾病名称归一化实验，并引入多态语义特征以改进模型效果。

3　模型设计

本文设计的MTAD-BERT-GCNN模型结构如图1所示。首先，根据好大夫、求医问药、好问康、THUOCL（THU Open Chinese Lexicon）、CCKS2017（China Conference on Knowledge Graph and Semantic Computing-2017）和ICD-10（International Classification of Diseases-10）分别构建实验数据集、特征训练语料和注意力权重词典；其次，利用word2vec和Glove在特征训练语料上生成字词向量；接着分别将疾病描述文本转化为向量输入到子任务；然后利用GCNN（graph convolutional neural network）和BERT同时对输入向量进行特征训练和提取，并引入注意力权重词典以调节向量表示质量；最后，根据Softmax函数实现疾病名称归一化。其中，BERT输入的是动态语义向量，GCNN输入的是静态语义向量。因此，MTAD-BERT-GCNN模型可以通过多任务学习捕获特征向量的静态和动态语义信息，并利用共享权重参数优化多个子任务深度挖掘语义信息，从而提升实验效果。

3.1　数据准备

1）构建实验数据集

由于国内缺少公开的疾病名称归一化数据集，本文参照英文疾病名称归一化评测任务，构建了中文疾病名称归一化数据集（Chinese Disease Normalization Data，ChDND）。具体过程包含两部分。一是数据获取及处理。从好大夫在线网站爬取了

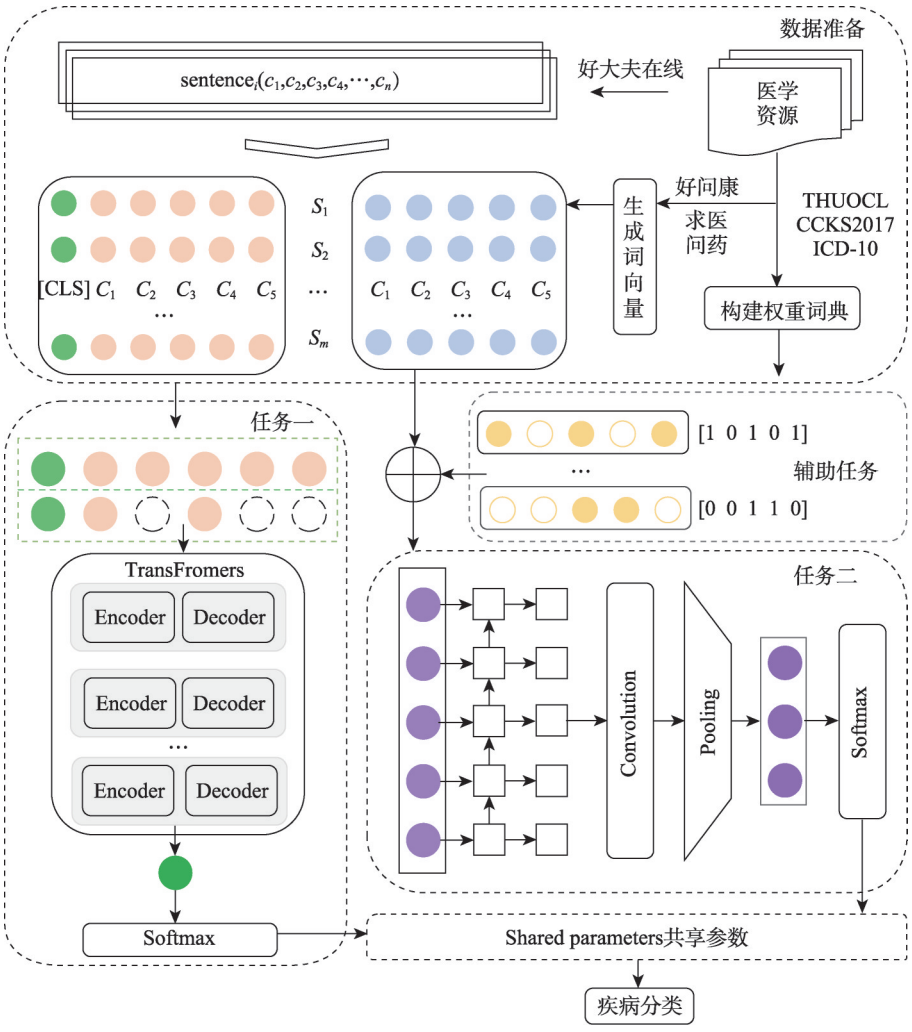

图1　MTAD-BERT-GCNN模型结构图

46140 条疾病描述和 537 个疾病名称，参照已有研究[27,41]，去除出现频次少于 10 的疾病名称及其对应描述，并分别生成词级和字级疾病描述；二是建立映射关系。基于网站的类别信息，将疾病描述与对应疾病名称建立多对一的映射关系。最后，本文构建的数据集 ChDND 包含了 407 个疾病名称和 42891 个疾病描述，平均每个疾病名称对应 105 个疾病描述。数据集 ChDND 的示例如表 1 所示。

表 1 中文疾病数据集实例

疾病名称	词级文本	字级文本
败血症	高烧呕吐无力大腿红肿医院诊断迁移性败血症年月开始突发高烧全身无力胃寒大腿红肿年月日出现这种情况	高烧呕吐无力大腿红肿医院诊断迁移性败血症年月开始突发高烧全身无力胃寒大腿红肿年月日出现这种情况
皮肌炎	线粒体阳性是不是肝不好我查了是说有免疫性肝病的可能结果是肝硬化还说肌炎可能是不确定的百度免疫性肝炎还挺严重然后百度说有蜘蛛痣我脖子真的有颗痣一直在长大还痒痒	线粒体阳性是不是肝不好我查了是说有免疫性肝病的可能结果是肝硬化还说肌炎可能是不确定的百度免疫性肝炎还挺严重然后百度说有蜘蛛痣我脖子真的有颗痣一直在长大还痒痒

2）生成特征向量

基于求医问药和好问康在线医疗社区问答语料，利用 word2vec 和 Glove 两种词向量训练模型，生成具有局部和全局语义特征的多特征融合向量，并作为本实验静态语义向量的输入。BERT 预训练向量是谷歌提供的中文预训练模型 BERT-Base-Chinese。

3）构建注意力权重词典

引入医学词典构建注意力权重词典以提高领域关键词的权重，降低非专业化表述的影响，进而提升关键特征的提取效果。本实验所采用的医学词汇，一方面，来源于 ICD-10 和 THUOCL 中的专业医学词汇；另一方面，抽取了 CCKS2017 电子病历数据集中的所有医疗实体。其中，ICD-10 是国际疾病分类，包含 1587 个疾病类别，本实验提取了 5634 个疾病特征词汇；THUOCL 是清华大学 NLP 组构建的中文词库，词表来自主流网站的社会标签、搜索热词和输入法词库，本实验提取了 18749 个专业医学词汇；CCKS2017 是 2017 年全国知识图谱与语义计算大会中文电子病历命名实体识别竞赛数据，包含 2505 条电子病历，本实验提取了 13802 个高频实体词汇。此外，ICD-10 是标准的医学术语，THUOCL 中的医学词汇符合医学术语规范；相比而言，经 CCKS2017 提取的词汇主要来自电子病历中医生表述，其规范性略低于医学词典。

3.2 关键技术

1）LSTM

长短时记忆网络[42]（LSTM）是 RNN 的变体，它可解决文本序列中的长期依赖问题，该模型由忘记门、输入门和输出门组成。其中，忘记门决定细胞状态丢弃的信息；输入门添加细胞状态中的新信息；输出门则判断细胞的状态特征，联合输入层中的细胞状态计算得到最终输出。

2）GRU

GRU[43]是 LSTM 的变体，它将三门结构替换为更新门和重置门两门结构，优化了网络结构，在联动表达式将前一节点和当前节点相结合以更新单元记忆。

3）BiLSTM、BiGRU

LSTM 和 GRU 均采用正向传播算法，仅能获取文本序列正向的上文语义信息，而忽略了后向序列的语义影响。BiLSTM 和 BiGRU 可以通过正反传播获取上下文全局语义特征。

4）CNN

卷积神经网络[44]（CNN）是一种前馈神经网络，它通过多个卷积核提取文本信息。该模型包含输入层、卷积层、池化层、连接层和输出层。其中，输入层将向量转换成张量矩阵；卷积层提取输入向量的局部特征和位置编码信息，利用卷积核进行首次特征提取；池化层对文本向量进行二次特征提取，通过降维保留关键信息；全连接层用于拼接和拟合池化后的特征向量以降低模型损失值；输出层根据任务目标选择不同函数并输出相应结果。

5）BERT

BERT 是一种基于转换器的双向编码表征模型，具有强大的特征提取功能。Transformer[45]是 BERT 的主要框架，它基于自注意力机制能够更全面地捕捉语句间的双向关系；BERT 基于掩藏语言模型（mask language model，MLM）突破了单项语言模型的限制，利用 MASK 随机替换输入特征以提高模型对特征的辨识度。在具体分类任务中，BERT 在

每条数据前插入[cls]标记，并将 Transformer 输出结果汇总到该标记，从而实现整个输入序列的信息汇总，从句向量角度实现分类任务。

3.3　多任务

1）任务一

任务一基于动态语义向量进行 BERT 微调和疾病描述映射。首先，文本 S_i 经过数据准备阶段转化为向量矩阵 $\boldsymbol{C}_i=([\text{cls}],c_1,c_2,c_3,\cdots,c_i,\cdots,c_n)$ 并输入到该任务，c_i 可与 BERT 预训练嵌入层建立唯一映射关系；其次，BERT 将输入向量转化为字向量特征 $\boldsymbol{W}_i$、位置特征 Pos_i 和分割嵌入 Seg_i 三种嵌入特征，并将三特征求和作为新的输入向量矩阵，其中 Seg_i 在单句文本分类时记为 0；接着，BERT 经多层 Transformer 生成微调后的动态语义向量，输入到下游任务计算分类向量 $\boldsymbol{\text{CLS}}_i=(\text{cls}_1,\text{cls}_2,\text{cls}_3,\cdots,\text{cls}_i,\cdots,\text{cls}_n)$；然后，利用 Softmax 函数结合训练的最佳权重和偏置（$W1_i$ 和 $b1_i$）将 $\boldsymbol{\text{CLS}}_i$ 转换为概率向量 $\boldsymbol{P}_i=(p_1,p_2,p_3,\cdots,p_l)$，其中，$p_i$ 为疾病描述文本映射到候选疾病名称的概率；最后，利用交叉熵函数计算该任务损失，具体公式为

$$\boldsymbol{P}1_i=\text{Softmax}(W1_i\boldsymbol{\text{CLS}}+b1_i)$$

$$\text{Loss}_{\text{task1}}=-\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}y_{i,j}\log(p1_i^j)\tag{1}$$

2）任务二

任务二基于静态语义向量进行特征挖掘和疾病描述映射。首先，文本 S_i 经过数据准备阶段转化为字符向量矩阵 $\boldsymbol{C}_i=(c_1,c_2,c_3,\cdots,c_i,\cdots,c_n)$ 输入到该任务，c_i 可与多特征融合嵌入层建立唯一映射关系；其次，利用 GRU 训练向量矩阵，增强输入文本序列间语义关系，计算得到向量矩阵 $\boldsymbol{H}_i$；接着，利用 CNN 提取该向量矩阵中的重要信息，保留输入文本的关键语义特征，经过卷积池化后得到向量矩阵 $\boldsymbol{F}_i$；然后，利用 Softmax 函数结合训练得到的最佳权重和偏置（$W2_i$ 和 $b2_i$）将 $\boldsymbol{F}_i$ 转换为概率向量 $\boldsymbol{P}_i=(p_1,p_2,p_3,\cdots,p_l)$，其中 p_i 为疾病描述文本映射到候选疾病名称的概率；最后，利用交叉熵函数计算该任务损失，具体公式为

$$\boldsymbol{P}2_i=\text{Softmax}(W2_i\boldsymbol{F}+b2_i)$$

$$\text{Loss}_{\text{task2}}=-\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}y_{i,j}\log(p2_i^j)\tag{2}$$

3）辅助任务

辅助任务可提取任务二中输入文本的关键词注意力权重。首先，将任意输入文本 $\boldsymbol{T}_i=(t_1,t_2,t_3,\cdots,t_i,\cdots,t_n)$ 与注意力权重词典建立映射；其次，当输入文本的词汇在注意力权重词典出现时，将该位置标记为 1，否则标记为 0，得到一个注意力矩阵 $\boldsymbol{\text{AT}}_i=(\text{at}_1,\text{at}_2,\text{at}_3,\cdots,\text{at}_i,\cdots,\text{at}_n)$，其中 $\text{at}_i=0,1$；再次，将该矩阵 $\boldsymbol{\text{AT}}_i$ 与任务二中的 $\boldsymbol{C}_i$ 矩阵相乘计算得到 $\boldsymbol{\text{C\_AT}}_i$，该向量经任务二特征提取得到向量矩阵 $\boldsymbol{\text{F\_AT}}_i$；最后，计算融入注意力权重后的概率向量 $\boldsymbol{P}2_i$，具体公式为

$$\boldsymbol{P}2_i=\text{Softmax}(W2_i\boldsymbol{\text{F\_AT}}+b2_i)\tag{3}$$

4）共享参数

多任务学习中，多个关联任务间通过损失函数相互调节以共享信息，并优化参数，分别反馈到每

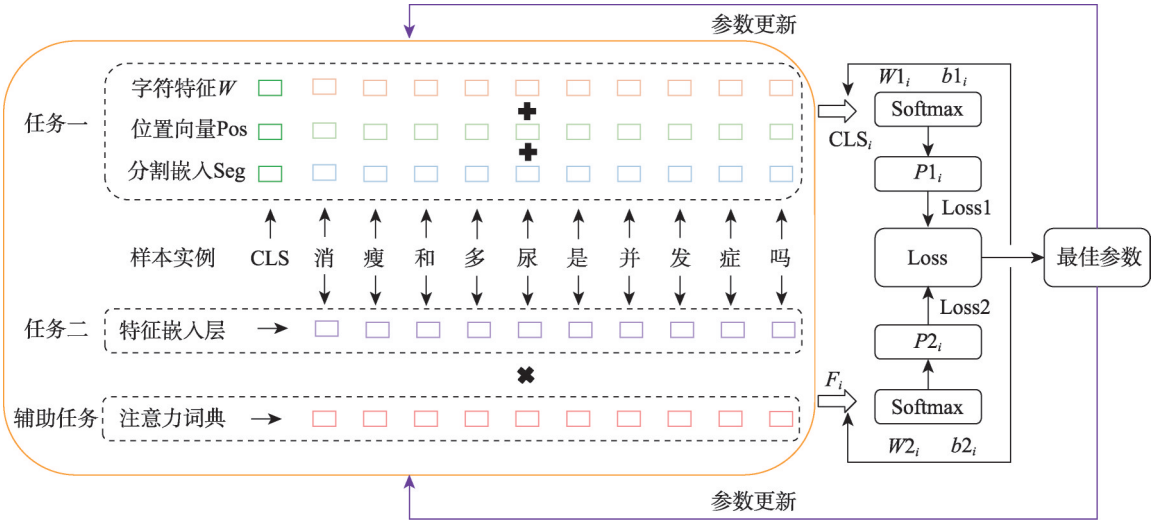

图 2　多任务共享参数流程

个子任务以提高模型效果。其中，共享损失函数为

$$\text{Loss} = -\sum_{i=0}^{m-1}\sum_{j=0}^{l-1} y_{i,j}(\log(p1_i^j) + \log(p2_i^j)) \quad (4)$$

具体共享参数流程如图2所示。

4 实验分析

4.1 实验设计

本文实验目的如下。

（1）验证基准模型在中文疾病名称归一化任务上的效果。

（2）验证引入语义关系对中文疾病名称归一化实验的影响。

（3）验证引入多任务学习对中文疾病名称归一化实验的影响。

基于上述实验目的，本文共设计了三组对照实验。每组实验均采用五折交叉验证，按7∶2∶1划分为训练集、验证集和测试集，具体设计如下。

实验一：对比分析CNN-WRv（CNN中嵌入词级随机向量）、CNN-CRv（CNN中嵌入字级随机向量）、CNN-WGv（CNN中嵌入外部语义特征）以及BERT-Base（基于预训练BERT进行微调）的实验效果。

实验二：在实验一中实验效果最佳CNN的基础上，分别引入GRU、LSTM、BiGRU和BiLSTM训练语义关系，分析语义训练后不同特征向量对中文疾病名称归一化的影响。

实验三：基于多任务学习，将实验一和实验二中表现最优的模型相结合，验证多任务学习对中文疾病名称归一化的效果，并在此基础上引入计算注意力权重的辅助任务，分析调节向量权重后模型对实验的影响。

具体实验思路如图3所示。

4.2 实验环境

本实验环境是一台内存20 GB、CPU型号为Intel(R) Core i5-7600K CPU、频率3.80 GHz、GPU为型号Nvidia GeForce RTX 2080 Ti、显存11 GB、操作系统为Windows 10的服务器。此外，实验中还使用了jieba分词库、哈工大LTP语言云、word2vec和Glove词向量训练工具、BERT和Tensorflow框架。开发环境为python 3.6、Tensorflow 1.13、keras 2.2.4、cuda10.0、cudnn 7.3.1。

4.3 实验参数

本实验中的具体参数设置如表2所示。

表2 模型参数设置

模型参数	CNN	GRU/LSTM/BiGRU/BiLSTM	BERT
输入句向量维度	100	100	768
卷积核窗口数	4	—	—
神经元	128	128	—
输入样本数	20	20	16
迭代次数	10	20	10
学习速率	0.001	0.001	2×10^{-5}
Dropout机制	0.5		
Softmax层数	归一化疾病名称数(407)		
优化器	Adam		

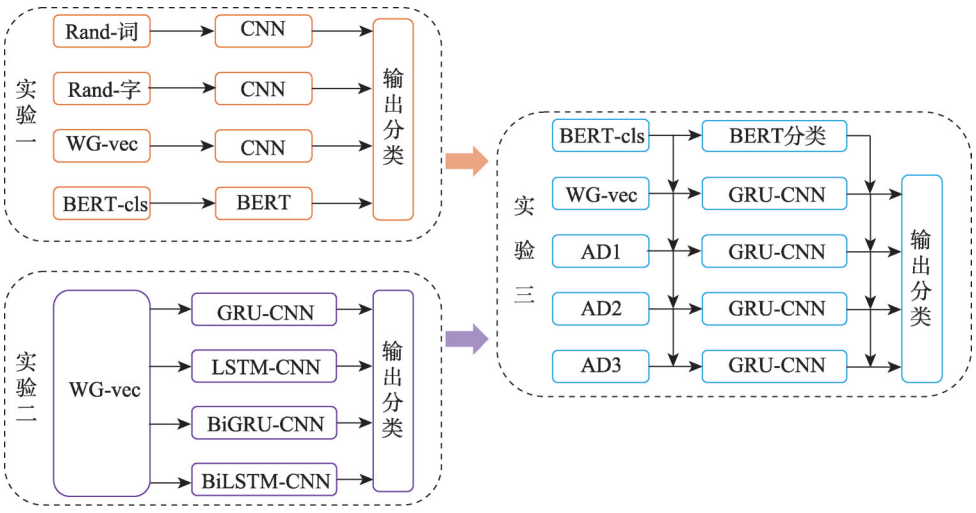

图3 实验思路

4.4　评价指标

参照已有研究[14,46-47]，本实验采用准确率（Accuracy）指标进行归一化评价，利用 Accuracy@k 评估疾病名称归一化效果，Accuracy@k 表示前 k 个预测疾病中正确结果的占比。分别取排名前 1、5 和 10 个疾病作为预测疾病，计算 Accuracy@1、Accuracy@5 和 Accuracy@10。由于多分类任务中难以计算负样本对结果的影响，本实验的归一化评价指标为

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

其中，TP 为判断为正确的疾病指称；FP 为判断为错误的疾病指称。

4.5　实验结果与分析

4.5.1　基准模型实验

为验证基准模型在中文疾病名称归一化任务上的效果，分别利用 CNN 和 BERT-Base 进行实验，结果如表 3 所示。

表 3　基准模型实验结果　%

模型	Accuracy@1	Accuracy@5	Accuracy@10
CNN-WRv	60.98	74.89	76.64
CNN-CRv	70.06	83.09	84.48
CNN-WGv	71.05	83.95	85.48
BERT-Base	74.46	85.41	88.02

由表 3 可知，在中文疾病名称归一化中，字级 CNN 效果优于词级 CNN；引入外部语义特征对模型效果的提升并不明显；BERT 微调后的效果较好，较 CNN 有明显提升。

（1）字级 CNN 效果优于词级 CNN。CNN-CRv 在 Accuracy@1、Accuracy@5 和 Accuracy@10 上较 CNN-WRv 分别提升了 9.08%、8.20% 和 7.84%，提升幅度较为明显。通过分析可知，这是由于在线医疗健康文本中医学词汇和口语化表述经常混杂出现，导致分词质量难以保证，从而影响到词级向量；而通过分字生成的字级向量可独立表示字符语义，因此在实验中表现出更好的效果。

（2）引入外部语义特征对实验的影响并不明显。CNN-WGv 在 Accuracy@1、Accuracy@5 和 Accuracy@10 上较 CNN-CRv 分别提升了 0.99%、0.86% 和 1.00%，提升幅度较小，表明词向量嵌入层中语义特征对 CNN 的影响较小，这是由于随机向量和外部语义特征均为唯一表示，不影响特征分布，但引入外部语义特征能够丰富特征语义，对模型效果有小幅提升。

（3）BERT 预训练模型微调后的效果较好。BERT-Base 在 Accuracy@1、Accuracy@5 和 Accuracy@10 上较最优的基线模型 CNN-WGv 分别提升了 3.41%、1.46% 和 2.54%，提升效果较为明显。这验证了 BERT 能够进一步提升疾病名称归一化效果，且显著优于其他基线模型，表明 BERT 能够更充分地捕获文本深层特征。

4.5.2　引入语义关系的 CNN 实验

通过实验一可知，字级向量在 CNN 上有较高的准确率，在此基础上，实验二分别引入 GRU、LSTM、BiGRU 和 BiLSTM 验证语义关系训练对实验结果的影响，具体如表 4 所示。

表 4　基于语义关系的 CNN 实验结果　%

模型	Accuracy@1	Accuracy@5	Accuracy@10
GRU-CNN	74.00	85.31	86.60
LSTM-CNN	73.77	85.12	86.34
BiGRU-CNN	73.13	84.17	85.47
BiLSTM-CNN	72.72	84.00	85.22

由表 4 可知，在 CNN 上引入 GRU、LSTM、BiGRU 和 BiLSTM 捕获文本间语义关系后的模型效果较表 3 中 CNN-WGv 有较大提升。其中，GRU-CNN 效果最优，在 Accuracy@1、Accuracy@5 和 Accuracy@10 上较引入外部语义特征的 CNN 分别提升了 2.95%、1.36% 和 1.12%。该结果表明，通过引入文本向量间语义关系可提高向量质量，在 CNN 中可提取更关键特征以进一步提升模型效果。

研究分析发现，引入 GRU 和 BiGRU 的效果优于 LSTM 和 BiLSTM，这是由于文本中大量的非医疗领域信息会影响模型学习疾病特征的语义质量，GRU 网络结构较 LSTM 更为简洁，可减少因大量非医疗领域信息计算而出现过拟合的影响。此外，引入 BiGRU 和 BiLSTM 的实验效果低于 GRU 和 LSTM，这是由于医疗健康文本的语序对语义关系影响不大，而 BiGRU 和 BiLSTM 因同时学习文本正负向语义关系造成过拟合，反而降低了文本语义关系的表达质量。

4.5.3　多任务学习实验

根据表 3 和表 4 可知，BERT-Base 和 GRU-CNN 两模型的表现最优，因此，在两模型基础上构建了 MT-BERT-GCNN 模型，用于验证多任务学习对中文

疾病名称归一化的影响。为了提高输入向量质量，进一步引入注意力权重词典来调节任务的特征输入，构建MTAD-BERT-GCNN模型以提升实验效果。多任务学习实验结果如表5所示。

表5 多任务学习实验结果 %

模型	Accuracy@1	Accuracy@5	Accuracy@10
MT-BERT-GCNN	74.97	85.92	88.79
MTAD-BERT-GCNN-THUOCL	75.32	86.23	89.03
MTAD-BERT-GCNN-CCKS	75.09	86.12	88.91
MTAD-BERT-GCNN-ICD10	75.39	86.66	89.60

由表5可知，基于多任务学习构建的MT-BERT-GCNN效果较BERT和GRU-CNN均有小幅提升，在Accuracy@1、Accuracy@5和Accuracy@10上，较GRU-CNN分别提升了0.97%、0.61%和2.19%，较BERT-Base分别提升了0.55%、0.51%和0.77%。这表明MT-BERT-GCNN的效果提升并非简单线性效果相加，而是能够利用多任务学习共享子任务参数，通过并行训练学习更多特征信息可提升当前主任务学习性能，从而获得更多代表性特征以提高疾病名称归一化的准确率。

进一步分析发现，引入计算注意力矩阵的辅助任务后，MTAD-BERT-GCNN效果较MT-BERT-GCNN得到了进一步提升，表明引入辅助任务调节特征输入可以筛选疾病的关键特征，对模型特征提取具有辅助作用。其中，MTAD-BERT-GCNN-ICD10的效果最佳，在Accuracy@1、Accuracy@5和Accuracy@10上，较MT-BERT-GCNN分别提升了0.42%、0.74%和0.81%，均略高于引入其他注意力权重词典的模型。引入注意力权重词典后，ICD10提升效果最佳，THUOCL次之。通过分析可知，ICD-10中包含了更多的专业医学术语，因而能够更充分地表示疾病特征；而CCKS中用词规范性略低于专业医学词典，在筛选特征时出现了部分非医学术语在辅助任务中权重分配错误的情况。

为了直观地呈现模型组合及多任务学习在中文疾病名称归一化上的效果，图4给出了三组对照实验结果。可以发现，在Accuracy@1、Accuracy@5和Accuracy@10上，MTAD-BERT-GCNN-ICD10较词级CNN基准模型分别提高了14.41%、11.77%和12.96%，较字级CNN基准模型分别提高了5.33%、3.57%和5.12%，这表明本文所提出的MTAD-BERT-GCNN可以在中文疾病名称归一化任务上取得最优效果。通过各模型汇总分析，实验结果可归纳为MTAD-BERT-GCNN>MT-BERT-GCNN>BERT-Base>引入语义关系的CNN>字级CNN>词级CNN。

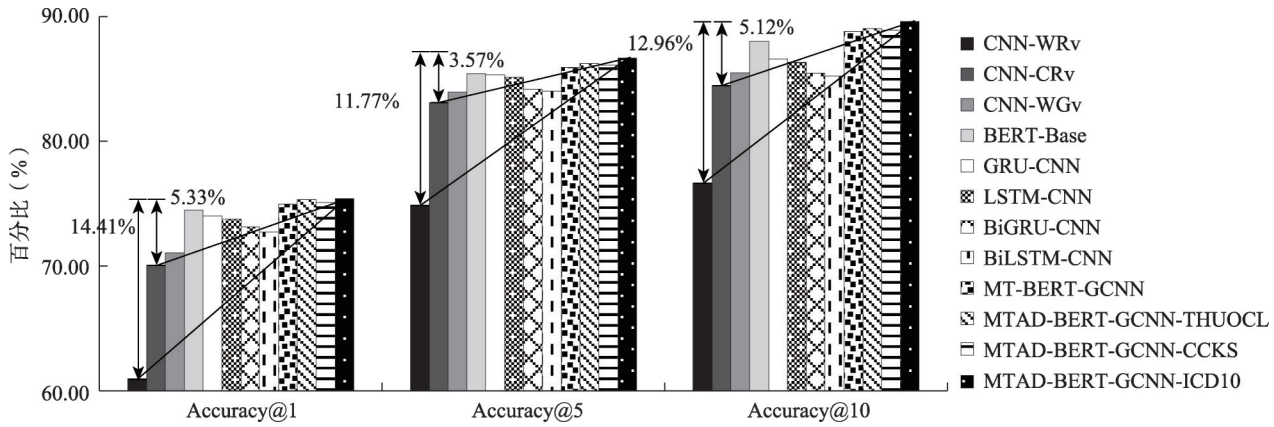

图4 实验数据对比分析

5 结 论

本文基于多任务学习和多态语义特征提出了中文疾病名称归一化模型MTAD-BERT-GCNN，该模型能够更好地利用多任务学习捕获多态语义信息，通过共享多任务间权重参数以深度挖掘文本信息，从而达到最优效果。研究结果发现，在中文疾病名称归一化中：①字级CNN效果优于词级CNN，引入外部语义特征对实验效果有小幅提升，BERT-Base较其他基准模型有大幅提升；②在CNN上融入GRU、LSTM、BiGRU和BiLSTM可捕获文本语义关系，进而提升中文疾病名称归一化效果；③基于多任务学习思路构建的MT-BERT-GCNN结合不同子任务的特点，通过优化任务间的共享参数，可

进一步提升实验效果，并且引入辅助任务筛选特征构建的MTAD-BERT-GCNN可使中文疾病名称归一化效果达到最优，最终在Accuracy@1、Accuracy@5和Accuracy@10上的准确率分别达到了75.39%、86.66%和89.60%，在Accuracy@10上较词级CNN和字级CNN分别提高了12.96%和5.12%。本研究将多任务学习思路应用于中文疾病名称归一化任务，并在中文数据集上验证了模型效果，为中文疾病名称归一化研究提供了可借鉴的思路。

尽管国外对疾病名称标准化和归一化的研究较多，但中文领域疾病名称归一化研究尚未得到充分重视。在后续研究中，一方面，将考虑结合文本、图片、语音和视频等多模态信息，从多维度进行疾病归一化研究；另一方面，将深入挖掘文本细微特征，以进一步推动中文疾病名称归一化研究进展。

参　考　文　献

[1] Magumba M A, Nabende P, Mwebaze E. Ontology boosted deep learning for disease name extraction from Twitter messages[J]. Journal of Big Data, 2018, 5(1): 1-19.

[2] 陈美杉, 夏晨曦. 肝癌患者在线提问的命名实体识别研究: 一种基于迁移学习的方法[J]. 数据分析与知识发现, 2019, 3(12): 61-69.

[3] Grover S, Aujla G S. Prediction model for influenza epidemic based on Twitter data[J]. International Journal of Advanced Research in Computer and Communication Engineering, 2014, 3(7): 7541-7545.

[4] 王萍, 牟冬梅, 高和璇, 等. 基于传染病监测数据的危机探测研究[J]. 情报学报, 2019, 38(5): 492-499.

[5] Chen L T, Baird A, Straub D. Fostering participant health knowledge and attitudes: an econometric study of a chronic disease-focused online health community[J]. Journal of Management Information Systems, 2019, 36(1): 194-229.

[6] Thelwall M, Buckley K. Topic-based sentiment analysis for the social web: the role of mood and issue-related words[J]. Journal of the American Society for Information Science and Technology, 2013, 64(8): 1608-1617.

[7] Li S, Yu C H, Wang Y C, et al. Exploring adverse drug reactions of diabetes medicine using social media analytics and interactive visualizations[J]. International Journal of Information Management, 2019, 48: 228-237.

[8] Karimi S, Metke-Jimenez A, Kemp M, et al. CADEC: a corpus of adverse drug event annotations[J]. Journal of Biomedical Informatics, 2015, 55: 73-81.

[9] Ching T, Himmelstein D S, Beaulieu-Jones B K, et al. Opportunities and obstacles for deep learning in biology and medicine[J]. Journal of the Royal Society Interface, 2018, 15(141): 20170387.

[10] Leaman R, Islamaj Doğan R, Lu Z Y. DNorm: disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29(22): 2909-2917.

[11] 韩普, 马健, 张嘉明, 等. 基于多数据源融合的医疗知识图谱框架构建研究[J]. 现代情报, 2019, 39(6): 81-90.

[12] 林泽斐, 欧石燕. 多特征融合的中文命名实体链接方法研究[J]. 情报学报, 2019, 38(1): 68-78.

[13] Luo Y, Song G J, Li P Y, et al. Multi-task medical concept normalization using multi-view convolutional neural network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018.

[14] Zhang Y Z, Ma X J, Song G J. Chinese medical concept normalization by using text and comorbidity network embedding[C]// Proceedings of the 2018 IEEE International Conference on Data Mining. IEEE, 2018: 777-786.

[15] Zhou S J, Li X. Feature engineering vs. deep learning for paper section identification: toward applications in Chinese medical literature[J]. Information Processing & Management, 2020, 57(3): 102206.

[16] Ristad E S, Yianilos P N. Learning string-edit distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522-532.

[17] Aronson A R. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program[J]. Proceedings of the AMIA Symposium, 2001: 17-21.

[18] Tsuruoka Y, McNaught J, Tsujii J, et al. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression[J]. Bioinformatics, 2007, 23(20): 2768-2774.

[19] Yang H. Automatic extraction of medication information from medical discharge summaries[J]. Journal of the American Medical Informatics Association, 2010, 17(5): 545-548.

[20] Khare R, Li J, Lu Z Y. LabeledIn: cataloging labeled indications for human drugs[J]. Journal of Biomedical Informatics, 2014, 52: 448-456.

[21] Kate R J. Normalizing clinical terms using learned edit distance patterns[J]. Journal of the American Medical Informatics Association, 2015, 23(2): 380-386.

[22] Jonnagaddala J, Jue T R, Chang N W, et al. Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion[J]. Database, 2016, 2016: baw112.

[23] Shi H R, Xie P T, Hu Z T, et al. Towards automated ICD coding using deep learning[OL]. (2017-11-30). https://arxiv.org/pdf/1711.04075.pdf.

[24] Liu H W, Xu Y. A deep learning way for disease name representation and normalization[C]// Proceedings of the 8th National CCF Conference on Natural Language Processing and Chinese Com-

puting. Cham: Springer, 2017: 151-157.

[25] Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 1014-1023.

[26] Li H D, Chen Q C, Tang B Z, et al. CNN-based ranking for biomedical entity normalization[J]. BMC Bioinformatics, 2017, 18 (Suppl 11): 385.

[27] Tutubalina E, Miftahutdinov Z, Nikolenko S, et al. Sequence learning with RNNs for medical concept normalization in user-generated texts[OL]. (2018-11-29). https://arxiv.org/pdf/1811.11523.

[28] Huang J M, Osorio C, Sy L W. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes[J]. Computer Methods and Programs in Biomedicine, 2019, 177: 141-153.

[29] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]// Proceedings of the 25th International Conference on Machine Learning. New York: ACM Press, 2008: 160-167.

[30] Liu P F, Qiu X P, Huang X J. Recurrent neural network for text classification with multi-task learning[OL]. (2016-05-17). https://arxiv.org/pdf/1605.05101.

[31] Liu P F, Qiu X P, Huang X J. Adversarial multi-task learning for text classification[OL]. (2017-04-19). https://arxiv.org/pdf/1704.05742.

[32] Yang J L, Liu Y N, Qian M H, et al. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding[J]. Applied Sciences, 2019, 9(18): 3658.

[33] Niu J H, Yang Y H, Zhang S H, et al. Multi-task character-level attentional networks for medical concept normalization[J]. Neural Processing Letters, 2019, 49(3): 1239-1256.

[34] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[OL]. (2019-05-24). https://arxiv.org/pdf/1810.04805.

[35] 陆伟, 李鹏程, 张国标, 等. 学术文本词汇功能识别——基于BERT向量化表示的关键词自动分类研究[J]. 情报学报, 2020, 39(12): 1320-1329.

[36] 吴俊, 程垚, 郝瀚, 等. 基于BERT嵌入BiLSTM-CRF模型的中文专业术语抽取研究[J]. 情报学报, 2020, 39(4): 409-418.

[37] Li F, Jin Y H, Liu W S, et al. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study[J]. JMIR Medical Informatics, 2019, 7(3): e14830.

[38] Xu D F, Gopale M, Zhang J C, et al. Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)-based ranking for concept normalization[J]. Journal of the American Medical Informatics Association, 2020, 27(10): 1510-1519.

[39] Ji Z C, Wei Q, Xu H. BERT-based ranking for biomedical entity normalization[OL]. (2019-08-09). https://arxiv.org/ftp/arxiv/papers/1908/1908.03548.pdf.

[40] Kalyan K S, Sangeetha S. BertMCN: mapping colloquial phrases to standard medical concepts using BERT and highway network [J]. Artificial Intelligence in Medicine, 2021, 112: 102008.

[41] Lee K, Hasan S A, Farri O, et al. Medical concept normalization for online user-generated texts[C]// Proceedings of the IEEE International Conference on Healthcare Informatics. IEEE, 2017: 462-469.

[42] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[43] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[44] Kim Y. Convolutional neural networks for sentence classification [OL]. (2014-09-03). https://arxiv.org/pdf/1408.5882.

[45] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [OL]. (2017-12-06). https://arxiv.org/pdf/1706.03762.

[46] Dogan R I, Lu Z. An inference method for disease name normalization[C]// Proceedings of the AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text. Palo Alto: AAAI Press, 2012: 8-13.

[47] Karadeniz İ, Özgür A. Linking entities through an ontology using word embeddings and syntactic re-ranking[J]. BMC Bioinformatics, 2019, 20(1): 156.

（责任编辑 车 尧）