基于领域知识图谱增强和 Lattice-LSTM 的中医药命名实体识别

牛天星[1,2] 郑小盈[2,3] 祝永新[2,3] 汪 辉[1,2,3*]

[1](上海科技大学信息科学与技术学院 上海 201210)

[2](中国科学院上海高等研究院 上海 201210)

[3](中国科学院大学 北京 101408)

摘 要 针对中医药领域命名实体识别任务中,现有的通过构造词典对实体识别模型进行增强的方法中存在的专业术语发现困难、构造词典效率低下和识别准确率不足等问题,提出一种基于领域知识图谱增强和 Lattice-LSTM 的领域命名实体识别模型。通过对已经构建完成的领域图谱使用嵌入算法,将其快速高效地转化为领域词典,并使用融合多粒度词汇信息的 Lattice-LSTM 将词典中的专业词汇编码到模型的输入中去,从而提高了模型在领域实体识别任务上的效果。采用中医药数据集进行实验,结果表明,所提模型的 F1 值高于传统实体识别模型,验证了模型的有效性。

关键词 领域知识图谱 中医药 命名实体识别 知识图谱嵌入

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2025.03.018

TRADITIONAL CHINESE MEDICINE NAMED ENTITY RECOGNITION BASED ON DOMAIN KNOWLEDGE GRAPH ENHANCEMENT AND LATTICE-LSTM

Niu Tianxing[1,2] Zheng Xiaoying[2,3] Zhu Yongxin[2,3] Wang Hui[1,2,3*]

[1](*School of Information Science and Technology, Shanghai Tech University, Shanghai 201210, China*)

[2](*Shanghai Advanced Research Institute, Chinese Academy of Science, Shanghai 201210, China*)

[3](*University of Chinese Academy of Sciences, Beijing 101408, China*)

Abstract Aiming at the problems in existing methods for enhancing entity recognition models through lexicon construction in Traditional Chinese Medicine (TCM) named entity recognition tasks—including difficulties in discovering domain-specific terms, inefficient dictionary construction, and insufficient recognition accuracy—this study proposes a domain-specific named entity recognition model based on domain knowledge graph enhancement and Lattice-LSTM. By applying embedding algorithms to a pre-constructed domain knowledge graph, we efficiently convert it into a domain-specific lexicon and incorporate multi-granularity lexical information through Lattice-LSTM to encode professional vocabulary in the dictionary into the model's input, thereby improving the model's effectiveness in domain-specific entity recognition tasks. Experiments on TCM datasets show that the F1-score of the proposed model is higher than that of traditional entity recognition models, verifying the model's validity.

Keywords Domain knowledge graph Traditional Chinese medicine Named entity recognition Knowledge graph embedding

0 引 言

近年来,随着知识图谱技术的发展,各个行业如医药、金融等都纷纷将领域知识与知识图谱技术深度融合形成领域知识图谱为行业下游任务进行赋能。而中医药学作为我国传统文化的瑰宝,包含了大量中医学理论、处方等知识信息。为了更好地推广与发展中医

收稿日期:2022-01-25。国家重点研发计划项目(2020SKA0120202);国家自然科学基金项目(U2032125)。**牛天星**,硕士生,主研领域:自然语言处理,知识图谱。**郑小盈**,副研究员。**祝永新**,研究员。**汪辉**,研究员。

药学,构建中医药知识图谱便显得尤为重要。通过构建好的中医药领域知识图谱,可以快速进行疾病筛查、处方开设,辅助医生快速为病人提供科学高效的中医服务。

在领域知识图谱构建流程中,领域命名实体识别是最核心的技术之一。命名实体识别是信息抽取下的核心子任务,旨在从非结构化文本中提取出潜在类型的实体,为后续关系抽取进行铺垫。例如,在句子“张三在十月二十日去了北京天安门”中,可以提取到实体有人名“张三”、时间“十月二十日”和地点“北京天安门”。再如,在句子“软坚散结,化瘀定痛,个别病人用药后有轻度头晕、恶心及局部灼痛”中,可以提取到的实体有中药功效“化瘀定痛”和症状“头晕”、“恶心”以及“灼痛”。通过对比两个句子可以发现,前者为通用领域的实体识别,识别的实体大多为常见词汇,而后者为中医药领域实体识别,其实体相对来说是比较少见的词汇,且样本数据集数量远小于通用领域数据集。同时,在领域中会遇到大量专业词汇,非专业人士十分难以区分,还可能存在嵌套性错误。如在“化瘀定痛”中,可能将“化瘀”识别为症状而忽略了整体词组为中药功效。

针对于领域命名实体识别任务中存在的数据集少、词汇专业性强等问题,越来越多的研究利用知识增强方法,通过不同方式构造领域词典来对传统命名实体识别模型进行增强,进而提升通用模型在领域任务中的效果。例如:Zhang 等[1]提出的 Lattice-LSTM 模型能够将输入的字符信息以及输入文本中潜在的所有能够与词典所匹配的词语信息一同编码输入到模型中,通过这种构造词典的方式,利用词语知识信息对模型进行知识增强,降低模型在特定单词识别时产生歧义的概率,进而提升了识别的准确率;赵耀全等[2]通过利用新词发现的方法,从大量医疗对话语料中获取了医疗领域的新词,并通过 Word2Vec 算法[3]构建了医疗领域词典对 Lattice-LSTM 模型进行增强,解决了原本 Lattice-LSTM 模型中词典不包含特定领域词汇信息的问题,验证了领域词汇知识信息对模型增强的有效性。

由于领域词典构造复杂,不论是通过人工标注还是利用新词发现或主题词发现[4]等算法从大量文本语料中提取领域词汇,都会耗费较多时间在人工收集、审核上。这使得构造领域词典的效率变慢,收集到的词汇信息依赖于文本语料,而且在使用 Word2Vec 生成词向量时还可能存在低频词丢失的情况[5]。

为了快速获得大量且准确的领域词汇,提高领域词典的构造效率,本文提出了基于领域知识图谱增强的 KGE-NER(Knowledge Graph Enhanced-Named Entity Recognition)模型组,该模型组充分利用已经构建好的包含大量专业词汇的领域知识图谱,通过知识图谱嵌入算法,将图谱快速转化为领域词典,并将词典中多种粒度的领域词汇作为特征,构建 Lattice-LSTM 模型来对单词进行建模。本文利用 KGE-NER 模型组在中医药领域任务上进行命名实体识别任务研究,使用 TransE[6]、TransD[7]两种不同知识图谱嵌入算法对中医药知识图谱进行嵌入操作后,生成中医药词典对基础 Lattice-LSTM 模型进行增强。通过实验结果发现,与原模型相比,引入领域知识图谱进行增强不仅能更好地提升实体识别的准确性,还可以有效地缩短构造词典的时间,提高词典中领域词汇的质量。最后,本文将算法流程进行抽象,提供了一种解耦式的知识增强方式,即快速利用不同的领域图谱来构建可“插卡式”使用、可快速迭代的领域词典对模型进行增强,使本文算法可以快速迁移至不同领域场景进行应用。

本文主要贡献如下:

(1) 提出了一种基于领域知识图谱增强和 Lattice-LSTM 的命名实体识别模型 KGE-NER,该模型在嵌入层使用知识图谱嵌入算法快速高效地将领域图谱转化为领域词典,在编码层使用融合多粒度词汇信息的 Lattice-LSTM,使模型能更好地利用领域词典中的词汇特征信息。同时,在中医药数据集上完成实验,验证了本文模型的有效性。

(2) 对比了嵌入层使用 TransE 和 TransD 两种不同知识图谱嵌入算法对模型效果的影响,验证了具有更高图表征能力的嵌入算法对模型具有一定提升效果。

(3) 提出了基于领域图谱增强的算法流程,使本文方法可被快速迁移至其他领域使用。

1 相关工作

传统的命名实体识别技术发展经历了四个阶段:规则阶段、序列标注阶段、深度学习阶段和基于预训练模型的阶段[8]。早期的规则阶段是设计一定的规则模板和词槽等方式,当遇到触发词后,自动将设定位置的实体切分出来,不同的触发词用来捕捉不同实体类型。如:在出现“邮箱:”触发词后,很容易将之后的邮箱实体抽取出来。随着文本信息的复杂程度提高,规则难以灵活适配更多的文本,研究者提出将实体识别任务转化为序列标注任务,从而使实体抽取更加灵活,不局限于固定的模板。例如:将“张三回家了”转换为序列标注“B-PER,I-PER,O,O,O”。其中“B-PER”与“I-PER”表示人名,“O”表示其他。如此对于任何长度

的复杂文本,都可以将其转化为标准的序列预测任务,并使用统计学习的方式,进行模型训练、预测。近几年,更多命名实体任务采用了深度学习的方法,使用卷积或循环神经网络作为编码器,提高了模型的特征捕获能力。尤其是在 Google 提出 Transformer 模型[9]之后,大规模预训练模型纷纷出现,BERT[10]和 ERNIE[11]等在大规模语料上训练出的模型代替传统神经网络作为实体识别任务的编码器后,通用领域的实体预测效果得到了巨大提升。但研究者很快意识到面向中文文档的专业领域命名实体识别与通用领域有着一定差异,英文中往往一个单词便可以划分为一个实体类型,而中文中经常会出现一个词语或多个字才能组成一个实体,而领域内的实体在中文实体的基础上还拥有更多复杂的词组。因此,更多中文的实体识别技术主要围绕在如何将中文的字、词汇、知识等信息引入到现有模型中,通过知识增强的方式,提升实体识别效果。目前最新的模型如 LatticeLSTM、FLAT[12]和 ERNIE 3.0 等都采用了知识增强的方式。LatticeLSTM 与 FLAT 模型都采用了构建中文词典的方式,在编码器阶段把已经训练好的中文词向量融入到神经网络的训练之中,区别是前者的编码器采用了 LSTM 而后者采用了改进的 Transformer 编码器。而 ERNIE 3.0 作为大规模预训练模型,它在预训练阶段,融入了大量通用知识图谱信息,通过 TransE 的知识图谱嵌入方式,把通用知识图谱中的词汇转化为词向量融入模型,极大提高了模型效果。

2　模型设计

本文所提出的中医药命名实体识别方法的流程分为两步。首先,利用知识图谱嵌入算法将已构建好的中医药领域知识图谱转化为一个中医药领域词典。然后,结合领域词典中潜在的多种词汇信息作为特征,构建基于词汇增强的 Lattice-LSTM 模型进行实体识别。

2.1　整体框架

图 1 为本文提出的 KGE-NER 模型框架。本文模型总共由三部分组成,分别为嵌入层的知识图谱嵌入模块、编码层的 Lattice-LSTM 模块和解码层的 CRF 模块。首先采用预训练好的字向量和经过知识图谱嵌入算法训练好的词向量作为 Embedding 输入;然后经过 Lattice-LSTM 进行编码,编码过程中会将嵌入层中得到的领域词汇作为特征进行输入;最后使用能捕捉标签间关系依赖的 CRF 模块进行解码,将隐含层的输出解码为出现概率最大的标记序列,得到最终的实体序列结果。

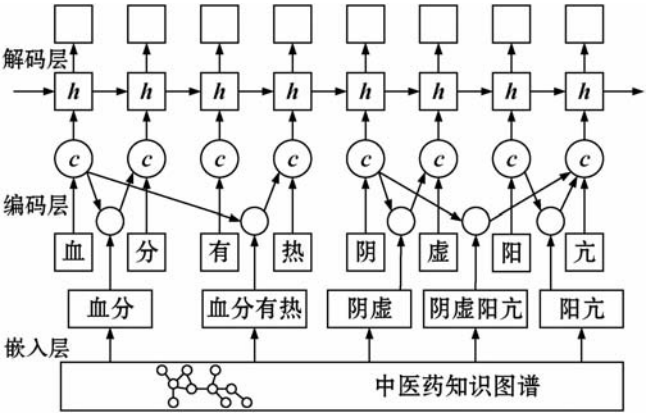

图 1　KGE-NER 模型框架

2.2　嵌入层:知识图谱嵌入算法

嵌入层主要是将输入的字、词进行向量映射,将离散字符映射为分布式表示。本文除了采用已经预训练的字向量外,最重要的是采用知识图谱嵌入算法对已构建的领域图谱进行向量化映射[13]。

传统知识图谱一般使用三元组(h,r,t)进行表示,其中 h 表示头实体,t 表示尾实体,r 表示关系。例如在三元组(阴虚,症状,血分有热)中,阴虚为头实体,血分有热为尾实体,关系为症状,即阴虚的症状为血分有热。为了知识图谱能在更多任务中,如预测、推理等,具有更强的可表达性与可扩展性,研究人员提出了知识图谱嵌入算法。此类算法将大量符号化的知识三元组映射到低维稠密的向量空间,同时保留数据中的结构信息,便于实体与关系之间的计算[14]。在得到的向量空间中,可以通过计算实体或关系之间的距离,进而推理出实体或关系的语义相似度。

本文采用经典的知识图谱嵌入算法 TransE 模型以及在此基础上进行改进的 TransD 模型对领域图谱进行映射。

TransE 算法属于翻译模型,其主要思想是将实体和关系映射为同一空间的向量,对于一个三元组的向量表示$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$,模型将其中的关系 $\boldsymbol{r}$ 视为头结点 $\boldsymbol{h}$ 到尾节点 $\boldsymbol{t}$ 的平移向量,即在向量空间中,头实体向量 $\boldsymbol{h}$ 加上关系向量 $\boldsymbol{r}$ 应该等于尾实体向量 $\boldsymbol{t}$。在 TransE 中,$\boldsymbol{h}+\boldsymbol{r}$ 与 $\boldsymbol{t}$ 的关系表示近似等价于向量相似度,采用计算欧氏距离的方式进行相似度计算,其得分函数为:

$$f(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}) = \|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_{L1/L2} \tag{1}$$

对于正样本三元组,$\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$,故得分函数值越小对正样本三元组越有利;对于负样本三元组,$\boldsymbol{h}+\boldsymbol{r}\neq\boldsymbol{t}$,故得分函数值越大对负样本三元组越有利。在模型训练过程中,不断更新参数,使 $\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}$ 的距离尽可能小。模型的损失函数为:

$$L = \sum_{(h,r,t)\in S}\sum_{(h',r',t')\in S'}[\gamma + f(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}) - f(\boldsymbol{h}',\boldsymbol{r}',\boldsymbol{t}')]_+ \tag{2}$$

式中：S 为正样本集合；S' 为负样本集合，负样本是通过大量随机替换正确三元组的头实体或尾实体产生；$[x]_+$ 表示函数值大于零时取值不变，小于零时则取零；γ 为正样本与负样本之间的间隔修正，值需要满足大于零的条件，一般设置为 1，γ 值越大对向量的修正越严格。模型目标为正确则三元组之间距离变小，错误三元组距离变大。最终通过不断训练，更新参数，获得知识图谱中每个实体与关系的向量化表示。

尽管 TransE 可以简单高效地自动捕捉推理特征，完成多种链接预测任务[15]，但它在知识图谱中存在的一对多、多对多等关系上，仍存在表达能力不足的问题。

TransD 算法在 TransE 的基础上进行设计，构造了映射矩阵。算法的主要思想是利用两个向量来表示符号对象（实体与关系），第一个表示一个实体（关系）的含义，另一个则动态构造映射矩阵。如此通过动态变化矩阵的方式，灵活地将实体表示投影到关系向量空间，表达了关系的多样性。其映射关系如图 2 所示。

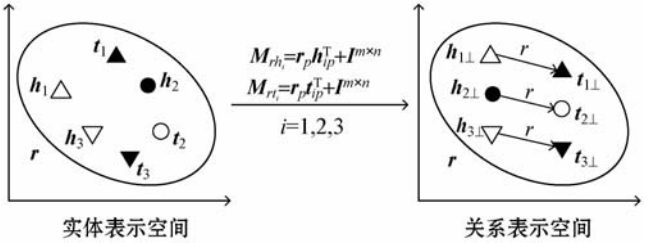

图 2 TransD 映射关系

TransD 的映射函数为：

$$\boldsymbol{M}_{rh} = \boldsymbol{r}_p \boldsymbol{h}_p^{\mathrm{T}} + \boldsymbol{I}^{m\times n} \tag{3}$$

$$\boldsymbol{M}_{rt} = \boldsymbol{r}_p t_p^{\mathrm{T}} + \boldsymbol{I}^{m\times n} \tag{4}$$

式中：头实体与尾实体利用 $\boldsymbol{M}_{rh}$ 和 $\boldsymbol{M}_{rt}$ 两个映射矩阵进行投影；$\boldsymbol{I}^{m\times n}$ 为单位矩阵；$\boldsymbol{h}_p^{\mathrm{T}}$ 和 $\boldsymbol{t}_p^{\mathrm{T}}$ 分别为头实体映射向量、尾实体映射向量；$\boldsymbol{r}_p$ 为关系映射向量，它与两种实体向量一同决定了映射矩阵。映射后的头、尾实体向量分别为：

$$\boldsymbol{h}_\perp = \boldsymbol{M}_{rh}\boldsymbol{h} \tag{5}$$

$$\boldsymbol{t}_\perp = \boldsymbol{M}_{rt}\boldsymbol{t} \tag{6}$$

故 TransD 模型的损失函数变为：

$$f_r(\boldsymbol{h}+\boldsymbol{t}) = \|\boldsymbol{h}_\perp + \boldsymbol{r} - \boldsymbol{t}_\perp\| \tag{7}$$

经实验验证，TransD 模型在处理复杂关系的图谱上有着很好的效果。它利用动态的关系投影矩阵增加了全局特征的捕捉能力，能更加充分地对知识图谱进行特征表示，为下游任务提供了更好的支持作用。

本文使用了 TransE 和 TransD 两种嵌入算法进行实验，将整理好的中医药领域知识图谱通过算法转化为词向量，构成领域词典，需要注意的是在选择图谱嵌入维度时需和模型中原有词典中词向量维度保持一致。最后经由人工筛选后，将领域词典与原有的词典进行融合，形成最终的词典供编码层使用。

2.3 编码层：Lattice-LSTM

LSTM 网络[16]由记忆单元来记录状态信息，通过输入门、遗忘门和输出门对模型的输入内容、记忆单元的存储内容以及模型的输出内容进行控制，形成对之前输入信息的记忆。

本文所使用的 Lattice-LSTM 模型是在基于字符的 LSTM-CRF 模型的基础上，提出了融合字、词的网格结构，如图 3 所示，通过增加词向量信息流，引入潜在的多粒度词语信息，增强了模型效果。

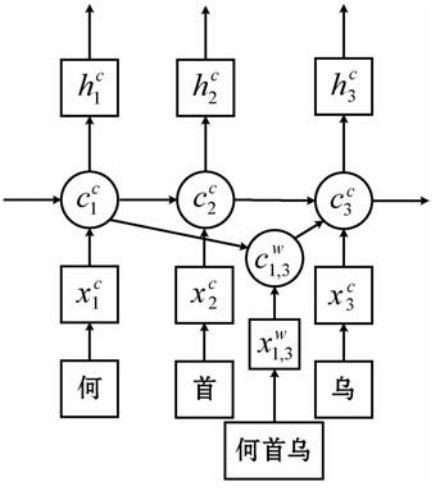

图 3 基于字词融合的 Lattice 结构

模型的输入由字序列和词序列两部分组成。用 $c_1, c_2, \cdots, c_m$ 表示输入的字符序列；用 $\boldsymbol{x}_j^{\mathrm{c}}$ 表示输入字序列中第 j 个字的字向量。

$$\boldsymbol{x}_j^{\mathrm{c}} = \boldsymbol{e}^{\mathrm{c}}(c_j) \tag{8}$$

式中：$\boldsymbol{e}^{\mathrm{c}}$ 表示嵌入层得到的字向量映射表；用 $\boldsymbol{c}_j^{\mathrm{c}}$ 表示字符序列信息流；用 $w_1, w_2, \cdots, w_m$ 表示输入的词序列；用 $x_{b,e}^{\mathrm{w}}$ 表示字序列中以 b 为起始位置，e 为结束位置的单词的词向量。

$$\boldsymbol{x}_{b,e}^{\mathrm{w}} = \boldsymbol{e}^{\mathrm{w}}(w_{b,e}^{\mathrm{d}}) \tag{9}$$

式中：$\boldsymbol{e}^{\mathrm{w}}$ 表示嵌入层得到的词向量映射表，$w_{b,e}^{\mathrm{d}}$ 表示词典中的单词；用 $\boldsymbol{c}_{b,e}^{\mathrm{w}}$ 表示词序列信息流；用 $\boldsymbol{h}_j^{\mathrm{c}}$ 表示 LSTM 的输出。

例如，在图 3 中 $\boldsymbol{x}_{1,3}^{\mathrm{w}}$ 表示“何首乌”的词向量；$\boldsymbol{c}_{1,3}^{\mathrm{w}}$ 表示“何首乌”单词的状态信息，它将作为词特征往字状态信息 $\boldsymbol{c}_3^{\mathrm{c}}$ 进行融入、增强。

对于字符序列信息流，采用基本 LSTM 公式进行计算：

$$\begin{bmatrix} \boldsymbol{i}_j^{\mathrm{c}} \\ \boldsymbol{o}_j^{\mathrm{c}} \\ \boldsymbol{f}_j^{\mathrm{c}} \\ \tilde{\boldsymbol{c}}_j^{\mathrm{c}} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\boldsymbol{w}^{\mathrm{c}\mathrm{T}} \begin{bmatrix} \boldsymbol{x}_j^{\mathrm{c}} \\ \boldsymbol{h}_{j-1}^{\mathrm{c}} \end{bmatrix} + \boldsymbol{b}^{\mathrm{c}} \right) \tag{10}$$

$$\boldsymbol{c}_j^{\mathrm{c}} = \boldsymbol{f}_j^{\mathrm{c}} \odot \boldsymbol{c}_{j-1}^{\mathrm{c}} + \boldsymbol{i}_j^{\mathrm{c}} \odot \tilde{\boldsymbol{c}}_j^{\mathrm{c}} \tag{11}$$

$$\boldsymbol{h}_j^{\mathrm{c}} = \boldsymbol{o}_j^{\mathrm{c}} \odot \tanh(\boldsymbol{c}_j^{\mathrm{c}}) \tag{12}$$

式中:$\boldsymbol{i}_j^{\mathrm{c}}$、$\boldsymbol{o}_j^{\mathrm{c}}$和 $\boldsymbol{f}_j^{\mathrm{c}}$ 分别为输入门、输出门和遗忘门;$\boldsymbol{w}^{\mathrm{c}\mathrm{T}}$ 和 $\boldsymbol{b}_c$ 为模型参数;$\sigma(\,)$代表 sigmoid 函数。

对于词序列信息流,并没有设置输出门。模型本质上是将词信息作为特征融入字符序列信息流中,使得字符信息的记忆单元 $\boldsymbol{c}_j^{\mathrm{c}}$ 计算时会受到多条路径上的信息流影响。如图 3 中 $\boldsymbol{c}_3^{\mathrm{c}}$ 会受到 $\boldsymbol{x}_3^{\mathrm{c}}$、$\boldsymbol{c}_{1,3}^{\mathrm{w}}$以及上一个时刻状态 $\boldsymbol{c}_2^{\mathrm{c}}$ 同时影响。而在有多个单词以同一个字符为结尾时,会将所有 $\boldsymbol{c}_{b,e}^{\mathrm{w}}$与 $\boldsymbol{c}_e^{\mathrm{c}}$ 连接起来,使用一个额外的输入门 $\boldsymbol{i}_{b,e}^{\mathrm{c}}$来控制 $\boldsymbol{c}_{b,e}^{\mathrm{w}}$到 $\boldsymbol{c}_{b,e}^{\mathrm{c}}$的信息流。其计算公式为:

$$\begin{bmatrix} \boldsymbol{i}_{b,e}^{\mathrm{w}} \\ \boldsymbol{f}_{b,e}^{\mathrm{w}} \\ \tilde{\boldsymbol{c}}_{b,e}^{\mathrm{w}} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\boldsymbol{w}^{\mathrm{w}\mathrm{T}} \begin{bmatrix} \boldsymbol{x}_{b,e}^{\mathrm{w}} \\ \boldsymbol{h}_b^{\mathrm{c}} \end{bmatrix} + \boldsymbol{b}^{\mathrm{w}} \right) \tag{13}$$

$$\boldsymbol{c}_{b,e}^{\mathrm{w}} = \boldsymbol{f}_{b,e}^{\mathrm{w}} \odot \boldsymbol{c}_b^{\mathrm{c}} + \boldsymbol{i}_{b,e}^{\mathrm{w}} \odot \tilde{\boldsymbol{c}}_{b,e}^{\mathrm{w}} \tag{14}$$

$$\boldsymbol{i}_{b,e}^{\mathrm{c}} = \sigma\left(\boldsymbol{w}^{l\mathrm{T}} \begin{bmatrix} \boldsymbol{x}_e^{\mathrm{c}} \\ \boldsymbol{c}_{b,e}^{\mathrm{w}} \end{bmatrix} + \boldsymbol{b}^{l} \right) \tag{15}$$

$$\boldsymbol{c}_j^{\mathrm{c}} = \sum_{b \in \{b' \mid w_{b',j}^{d} \in D\}} \alpha_{b,j}^{\mathrm{c}} \odot \boldsymbol{c}_{b,j}^{\mathrm{w}} + \alpha_j^{\mathrm{c}} \odot \tilde{\boldsymbol{c}}_j^{\mathrm{c}} \tag{16}$$

式中:$\alpha_{b,j}^{\mathrm{c}}$和 α_j^{c} 是由 $\boldsymbol{i}_{b,j}^{\mathrm{c}}$和 $\boldsymbol{i}_j^{\mathrm{c}}$ 归一化而来,故所有的 $\alpha_{b,j}^{\mathrm{c}}$ 的和为 1,所有的 α_j^{c} 的和也为 1,归一化公式为:

$$\alpha_{b,j}^{\mathrm{c}} = \frac{\exp(\boldsymbol{i}_{b,j}^{\mathrm{c}})}{\exp(\boldsymbol{i}_j^{\mathrm{c}}) + \sum_{b' \in \{b'' \mid w_{b'',j}^{d} \in D\}} \exp(\boldsymbol{i}_{b',j}^{\mathrm{c}})} \tag{17}$$

$$\alpha_j^{\mathrm{c}} = \frac{\exp(\boldsymbol{i}_j^{\mathrm{c}})}{\exp(\boldsymbol{i}_j^{\mathrm{c}}) + \sum_{b' \in \{b'' \mid w_{b'',j}^{d} \in D\}} \exp(\boldsymbol{i}_{b',j}^{\mathrm{c}})} \tag{18}$$

隐含层的输出 $\boldsymbol{h}_j^{\mathrm{c}}$ 的计算公式为:

$$\boldsymbol{h}_j^{\mathrm{c}} = \boldsymbol{o}_j^{\mathrm{c}} \odot \tanh(\boldsymbol{c}_j^{\mathrm{c}}) \tag{19}$$

式中:$\boldsymbol{w}^{\mathrm{w}}$、$\boldsymbol{b}^{\mathrm{w}}$、$\boldsymbol{w}^{\mathrm{l}}$ 和 $\boldsymbol{b}^{\mathrm{l}}$ 为模型参数。

2.4　解码层:CRF

编码层在对输入序列编码完成后,得到所有隐含层的输出信息 $\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n$,在对隐含层输出使用 CRF 后,得到标签序列 $y = l_1,l_2,\cdots,l_n$ 的概率为:

$$P(y \mid s) = \exp \frac{\left(\sum_i \left(\boldsymbol{w}_{\mathrm{CRF}}^{l_i} \boldsymbol{h}_i + b_{\mathrm{CRF}}^{(l_{i-1},l_i)} \right) \right)}{\sum_{y'} \exp\left(\sum_i \left(\boldsymbol{w}_{\mathrm{CRF}}^{l'_i} \boldsymbol{h}_i + b_{\mathrm{CRF}}^{(l'_{i-1},l'_i)} \right) \right)} \tag{20}$$

式中:y'表示任意的标签序列;$\boldsymbol{w}_{\mathrm{CRF}}^{l_i}$是针对于特定 l_i的参数;$b_{\mathrm{CRF}}^{(l_{i-1},l_i)}$是针对于 l_{i-1}到 l_i 的偏差。

模型采用带有 L_2 正则化的句子级别对数似然损失函数进行训练:

$$L = \sum_{i=1}^{N} \log(P(y_i \mid s_i)) + \frac{\lambda}{2} \|\theta\|^2 \tag{21}$$

式中:λ 是 L_2正则化参数;θ 表示整个模型的参数集合。

本文使用维特比解码算法从最终输出的标签序列中寻找概率最高的路径,得到的结果就是最终所需的标注序列。

2.5　基于领域图谱增强的算法流程及应用

本文对所提的中医药命名实体识别算法进行抽象,提出了基于领域图谱增强的算法流程,便于快速迁移至其他领域使用。算法旨在通过利用额外的领域知识信息增强多类构建图谱子任务的效果,最终除了得到任务所需结果外,通过不断的迭代过程,同时可以产生、完善相关领域的知识图谱。

算法步骤如下:

步骤 1　分析需求,确定任务领域,确定任务类型。

步骤 2　收集相关领域知识,寻找开源或已构建完成的相关领域图谱,将收集到的知识人工审核后添加入图谱之中,进行完善。

步骤 3　通过分析已构建好的领域图谱结构,选择合适的知识图谱嵌入算法,生成领域词典。

步骤 4　通过任务类型,选择该任务中具有更好的融合知识信息的模型,如实体识别领域的 FLAT 模型等。利用模型生成任务结果。同时,任务结果中一般包含了更新的领域知识信息,对这些信息进行整理、收集。

步骤 5　判断任务效果是否达到要求,或已构建的领域图谱是否满足使用条件,若否,则将已得到的任务结果人工审核后,跳转到步骤 2,持续更新完善图谱。若符合要求,跳转到步骤 6。

步骤 6　最终得到任务结果,同时得到符合使用条件的领域知识图谱。

经过验证,该算法流程在分词、文本分类等领域均可起到一定增强效果,可被快速迁移至多领域应用。同时,随着构建图谱的不断完善,所起到的增强效果不断提升。

3　实　验

3.1　数据集

在中医药命名实体识别任务中,本文实验所用数据来自于医疗网站上中医药科医生与患者对话数据和

部分中药说明书以及药品使用指南。通过收集、汇总相关数据后,由具有中医药学背景的相关人员进行标注,数据中的实体采用 BIO 的标注方式。标注后再由具有临床经验医生进行审核,保证了数据的标注质量。同时本文所用的中医药领域知识图谱是在专业网站中收集、整理中医药知识信息后由专业人员构建生成。本文将数据中的中医药实体标注为五种类型,分别为临床症状、药品效果、疾病名称、适合人群和药品名称。不同类型实体的具体数目如表 1 所示。最后将全部数据大致按照 6∶2∶2的比例划分为训练数据、验证数据和测试数据。

表 1 实体类型及数目

数据	临床症状	药品效果	疾病名称	适合人群	药品名称
训练数据	2 735	2 026	1 110	849	648
验证数据	698	592	314	205	148
测试数据	840	498	247	231	129

3.2 评价指标

本文使用命名实体识别领域的 F1 值来对模型性能进行评测。F1 值是由标准准确率(P)、召回率(R)计算得出。其计算公式为:

$$P = \frac{T_P}{T_P + F_P} \tag{22}$$

$$R = \frac{T_P}{T_P + F_N} \tag{23}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{24}$$

式中:T_P 为真正例,代表真实数据为正例同时预测值也为正例,所谓预测值正确是指实体的开始位置、中间位置和结束位置同时预测正确;F_P 为假正例,代表真实数据为反例同时预测值为正例;F_N 为假反例,代表真实数据为正例同时预测值为反例;P 代表预测正确的结果占所有预测结果的比例、R 代表预测正确的结果占所有数据的比例;F_1 值结合了两种指标,通常作为命名实体识别任务的综合测评指标使用。

3.3 实验超参数设置

本文模型的超参数设置如表 2 所示。

表 2 模型超参数设置

参数	值	参数含义
char emb size	50	字符嵌入层维度
lattice emb size	50	lattice 嵌入层维度
char dropout	0.5	嵌入层丢弃率,防止过拟合
LSTM layer	1	LSTM 网络层数
LSTM hidden	200	LSTM 隐藏层维度
lattice dropout	0.5	lattice 层丢弃率,防止过拟合
init lr	0.015	模型初始学习率
decay rate	0.05	学习率衰减速率
regularization	1e-8	正则化参数

本文使用 SGD(随机梯度下降算法)对模型进行优化,每轮学习率不断更新,由初始学习率 *init_lr*、学习率衰减速率 *decay_rate* 和训练轮次 *epoch* 三个参数决定,参数的值在表 2 中已给出。学习率 *lr* 更新公式为:

$$lr = init_lr \times ((1 - decay_{rate}) \times epoch) \tag{25}$$

同时本文实验环境均采用 Python 语言,基于 PyTorch 框架进行搭建。

3.4 结果分析

实验使用本文所提的 KGE-NER 模型组在中医药数据集上与主流命名实体识别模型进行了综合对比。主要从以下几方面进行对比分析:对比分析了通过领域知识图谱构建的词典与传统词典的区别,以及构建词典对识别效果的影响;对比了 KGE-NER 模型组与传统模型识别结果的 F1 值,验证了本文模型的有效性;对比分析了通过 TransE 算法进行知识图谱嵌入的 KGTransE-NER 模型与通过 TransD 算法嵌入的 KGTransD-NER 模型的识别效果;最后分析了基于领域知识图谱构建词典的方式与传统构建方式的构建效率及人工成本。

基于中医药领域知识图谱转化形成的领域词典与传统方式形成的字典相比,包含更多高质量的中医药专业词汇,为提高模型效果提供帮助。例如,在测试集中有如下句子:“凡阴虚,血分有热,胃火炽盛,肺有痰热,外感热病者慎服。”其中包含一个疾病名称实体“阴虚”以及三个症状实体“血分有热”“胃火炽盛”和“肺有痰热”。传统模型与本文模型的识别结果如表 3 所示。

表 3 识别结果示例

模型	识别正确实体
Lattice-LSTM[1]	阴虚
KGTransD-NER	阴虚、血分有热、胃火炽盛

通过对比识别结果示例可以发现,本文模型在复杂领域实体上的识别效果优于传统模型。其原因是:在构建领域词典时所用到的中医药图谱中包含如下一

些三元组:(阴虚,病因,不能滋润)、(阴虚,症状,头晕)、(阴虚,症状,血分有热)、(阴虚,症状,胃火炽盛)和(阴虚,症状,多梦)等。故由图谱转化而来的领域词典中包含如下专业词汇:阴虚、不能滋润、头晕、血分有热、胃火炽盛、多梦等。可以发现领域词典中的词汇包含了测试集示例中的疾病实体“阴虚”以及部分症状实体“血分有热”、“胃火炽盛”,故模型可以准确地识别出复杂的领域实体。而传统词典中不包含数据集示例中的相关实体词汇,所以很难识别到这些实体。此外,通过全面对比传统模型与本文模型所生成的词典可以发现,由领域图谱转化而来的词典对数据集中实体词汇的覆盖率要高于传统词典,这为后续利用词汇特征信息增强模型效果做了很好的铺垫,使得本文模型在领域实体识别的准确率上高于传统模型。同时,还有一些方法会利用新词发现、关键词提取等算法来构建领域词典,这些方法可以有效地提高对实体词汇的覆盖率,但是整体的覆盖率仍低于本文所提方式,这是因为在领域图谱的构建过程中,本身就会使用到类似算法来抽取知识信息,再经由专家审核后将信息添加进图谱之中,故图谱本身就是由各类算法以及人工参与集成而来的产物,包含更多的知识信息,所以对于某一领域专业知识的覆盖面较为完全。但是,仍可以发现,文中示例中的部分症状实体词汇在领域词典中仍未被包含,导致模型未能识别到症状实体“肺有痰热”。其原因是所使用的领域图谱具有一定的局限性与时效性,即本身图谱的构建过程中,不可能包含所有的领域知识,并且知识具有更新的过程,领域内随时都有新的知识产生,而这些知识可能未被包含在已构建好的图谱内,故可以通过不断迭代更新已构建好的图谱,来增加图谱的覆盖面与时效性。同时,存在的另一个问题是在一些专业词汇上存在多种近义词表达相同含义的情况,例如:将测试集中“胃火炽盛”替换为“胃热炽盛”后,模型存在预测错误的情况。为了解决类似问题,领域内仍需不断探讨知识定义,统一领域知识标准,为构建更规范、更高质量的领域图谱做好铺垫,如此才能更好地促进领域的发展。

本文实验对比了所提基于领域知识图谱增强的 Lattice-LSTM 模型与传统命名实体识别模型在中医药数据集上的实验结果。用来对比的传统模型包括基于字符的 CRF、BiLSTM-CRF 和未经过领域图谱增强的 Lattice-LSTM 模型。对比结果如表 4 所示,通过对比不同模型识别结果的 F1 值来验证模型识别的准确率。其中 KGTransE-NER 和 KGTransD-NER 为本文所提的基于不同的图谱嵌入方式的实体识别模型。与传统方法相比,本文算法的 F1 值得到了很大提升,同时与未经过领域图谱增强的 Lattice-LSTM 算法相比,F1 值最高也从 0.887 1 提升到 0.911 0。

表 4　实验结果对比

模型	F1
CRF[17]	0.836 4
BiLSTM-CRF[18]	0.860 8
Lattice-LSTM[19]	0.887 1
KGTransE-NER	0.901 7
KGTransD-NER	0.911 0

通过对实验结果对比分析,由 CRF 到 BiLSTM-CRF 上 F1 值的提升可以验证深度学习在实体识别效果上的优势;由 Lattice-LSTM 到 Bi-LSTM 上 F1 值的提升验证了引入多粒度词汇信息能对实体识别模型起到一定的帮助作用;由 Lattice-LSTM 到 KGTransE-NER 上 F1 值的提升验证了引入领域知识图谱算法的有效性;最后对比 KGTransD-NER 相较于 KGTransE-NER 上 F1 值的提升,可以验证更合适的图谱嵌入算法对模型的识别效果有着更大的提升。

本文实验虽然验证了使用具有更强大图表征能力的图谱嵌入算法 TransD 相较于普通的嵌入算法 TransE 会有更好的结果,F1 值由 0.901 7 提升至 0.911 0。但是,与两种嵌入算法在知识图谱预测任务中 TransD 比 TransE 大幅提升的效果相比,更好的图表征算法在本文模型中起到的作用并没有特别明显。原因可能有:在传统知识图谱预测任务中,所得到的嵌入结果直接应用在任务之上,而本文中所得到的嵌入结果仍需输入到编码层中使用,效果的提升依赖于模型结构对额外输入知识信息的利用能力,故探究更加强大的融合知识信息的模型结构是提升识别效果的一个关键;同时,TransD 与 TransE 相比,增加了投影矩阵,可以更好地表达复杂关系,而在本文所使用的领域图谱上,由于知识特性,并未存在大量的复杂关系与同义词组。所以其算法优势没有得到充分体现,而在其他具有复杂关系的图谱上,可能具有更好的效果,如金融领域、新媒体领域。故在使用嵌入算法前,对所用图谱的结构与特性需进行更好的探索。选择更符合的嵌入算法同样对提升模型效果起到了关键作用。

本文分析了领域词典构建的流程,相较于利用新词发现等算法,通过领域知识图谱转化的方式效率更高。因为本身是通过直接导入图谱进行操作,而不是人工筛选所发现的新词。知识图谱本身的构建过程已经消耗了人工成本,故直接利用可以节省一定的时间。同时在领域词典构建过程无须操作人员了解行业知

识,直接使用图谱嵌入算法即可。但是因为领域图谱本身不可能包含所有的领域词汇信息,需要一个更迭升级的过程。所以通常在构建流程中,可以先以已有的知识图谱快速生成字典后,同时也使用新词发现等算法更加快速的更迭升级字典,达到更好的识别效果。

4 结 语

本文在现有的通过构建领域词典对命名实体识别模型进行增强方式的基础上,针对传统方式构建词典效率低下、识别准确率不足等问题,提出了基于领域知识图谱增强和 Lattice-LSTM 的 KGE-NER 模型。通过使用知识图谱嵌入算法,将领域知识图谱快速高效的转化为领域词典,然后使用 Lattice-LSTM 模型将词典中的词汇特征融入到模型之中进行增强,从而在中医药领域实体识别中取得了更好的效果,降低了构建词典所花费的时间。同时,还提出了一种基于领域图谱增强的算法流程,便于快速迁移至其他领域使用。但是,在整个算法流程中仍存在以下问题:如何设计具有更好图表征能力的嵌入算法,以及如何设计具有更高的知识信息利用率的编码器结构。接下来我们会探索嵌入算法的效果与图谱结构的关系,尝试引入图神经网络的方式,利用其强大的结构表征能力,来设计嵌入算法与编码器结构,并应用在本文所提的算法流程中去。

参 考 文 献

[1] Zhang Y, Yang J. Chinese NER using lattice LSTM[C]// 56th Annual Meeting of the Association for Computational Linguistics. ACL,2018:1554 - 1564.

[2] 赵耀全,车超,张强. 基于新词发现和 Lattice-LSTM 的中文医疗命名实体识别[J]. 计算机应用与软件,2021,38(1):161 - 165,249.

[3] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB]. arXiv:1301.3781,2013.

[4] 李文坤,张仰森,陈若愚. 基于词内部结合度和边界自由度的新词发现[J]. 计算机应用研究,2015,32(8):2302 - 2304,2342.

[5] 朱向其,张忠林,李林川,等. 基于改进词性信息和 ACBiLSTM 的短文本分类[J]. 计算机应用与软件,2021,38(12):179 - 186.

[6] Bordes A, Usunier N, Garcia-Durán A, et al. Translating embeddings for modeling multi-relational data[C]//27th International Conference on Neural Information Processing Systems. ACM,2013:2787 - 2795.

[7] Ji G, He S, Xu L, et al. knowledge graph embedding via dynamic mapping matrix[C]//53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. ACL,2015:687 - 696.

[8] 蒋翔,马建霞,袁慧. 基于 BiLSTM-IDCNN-CRF 模型的生态治理技术领域命名实体识别[J]. 计算机应用与软件,2021,38(3):134 - 141.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB]. arXiv:1706.03762,2017.

[10] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB]. arXiv:1810.04805,2018.

[11] Sun Y, Wang S, Li Y, et al. ERNIE: Enhanced representation through knowledge integration[EB]. arXiv:1904.09223,2019.

[12] Li X, Yan H, Qiu X, et al. FLAT: Chinese NER Using flat-lattice transformer[EB]. arXiv:2004.11795,2020.

[13] 文峰,曹雄,黄海新,等. 基于知识图谱的推荐算法研究[J]. 沈阳理工大学学报,2021,40(6):13 - 17.

[14] 张正航,钱育蓉,行艳妮,等. 基于 TransE 的表示学习方法研究综述[J]. 计算机应用研究,2021,38(3):656 - 663.

[15] 余传明,张贞港,孔令格. 面向链接预测的知识图谱表示模型对比研究[J]. 数据分析与知识发现,2021,5(11):29 - 44.

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

[17] Lafferty J, Mccallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//18th International Conference on Machine Learning. ACM,2002:282 - 289.

[18] Huang Z, Wei X, Kai Y. Bidirectional LSTM-CRF models for sequence tagging[EB]. arXiv:1508.01991,2015.

[19] 张笑天. 基于 Lattice LSTM 的医学文本中文命名实体识别研究与实现[D]. 成都:电子科技大学,2019.

(上接第 67 页)

[18] Lillo I, Soto A, Niebles J. Discriminative hierarchical modeling of spatio-temporally composable human activities[C]// IEEE Conference on Computer Vision and Pattern Recognition,2014:812 - 819.

[19] Bulling A, Blanke U, Schiele B. A tutorial on human activity recognition using body-worn inertial sensors[J]. ACM Computing Surveys,2014,46(3):1 - 33.

[20] Raftery A E, Lewis S. How many iterations in theGibbs sampler? [C]//4th Valencia International Meeting,1992:1931 - 1989.

[21] Peebles P Z. Probability, random variable, and random signal principles[M]. 4th ed. New York: McGraw Hill,2001:100 - 110.