doi:10.3772/j.issn.2095-915x.2019.01.010

开放科学（资源服务）标识码（OSID）

基于深度学习的中医典籍命名实体识别研究

高甦[1] 金佩[2] 张德政[2]

1. 北京师范大学医院 北京 100875；
2. 北京科技大学计算机与通信工程学院 北京 100083

摘要：本文针对中医典籍存在的知识体系复杂、分词困难等难点以及传统方法人工构建特征不准确等问题，提出了一种基于深度学习的中医典籍命名实体识别方法。根据中医典籍的语料特征及主流的深度学习模型特点，以中医典籍的字向量为输入，采用基于双向长短时记忆神经网络和条件随机场（BiLSTM-CRF）的实体识别模型，对《黄帝内经》中的中医认识方法、中医生理、中医病理、中医自然、治则治法等 5 种实体进行识别，精确率为 85.44%，召回率为 85.19%，F1 值为 85.32%。在相同的中医典籍语料上做了大量对比分析实验，验证了该方法的有效性。结果证明：该方法有效提高了中医典籍的实体识别准确率，是深度学习在特殊语料处理领域的一次较有价值的尝试，具有一定的实践意义。

关键词：命名实体识别；深度学习；中医；黄帝内经
中图分类号：G35 TP182 R221

Research on Named Entity Recognition of TCM Classics Based on Deep Learning

GAO Su[1] JIN Pei[2] ZHANG Dezheng[2]

1. Beijing Normal University Hospital, Beijing 100875, China;
2. School of Computer and Communication Engineering University of Science and Technology Beijing, Beijing 100083, China

Abstract: Aiming at the problems of complex knowledge system, difficult word segmentation and inaccurate artificial construction of the Chinese medical classics, a method of named entity recognition of Chinese medical classics based on deep learning is

基金项目：国家重点研发计划云计算和大数据专项“大数据驱动的中医智能辅助诊断服务系统”（2017YFB1002300）。
作者简介：高甦（1978-），硕士，研究方向：中西医结合临床专业，E-mail: gaosu_su_78@163.com；金佩（1993-），硕士，研究方向：知识图谱自动构建、知识工程；张德政（1964-），博士，教授，研究方向：数据挖掘、知识工程。

doi:10.3772/j.issn.2095-915x.2019.01.010

proposed. According to the characteristics of Chinese medical classics and mainstream deep learning model, taking the character vectors of Chinese medical classics as input, a column labeling model based on Bidirectional Long Short Term Memory neural network and conditional random field (BiLSTM-CRF) is adopted to recognize the cognitive methods, physiology, pathology, nature and therapy in *Huangdi Neijing*. The recognition accuracy is 85.44%, the recall rate is 85.19%, and the F1 value is 85.32%. Besides, a large number of comparative analysis experiments have been done on the same corpus of TCM classics to verify the effectiveness of the method. The results show that the method significantly improves the recognition accuracy of entities in TCM classics, and makes a valuable attempt in the field of special corpus processing, which has a certain practical significance.

Keywords: Named entity recognition; deep learning; traditional Chinese Medicine; *Huangdi Neijing*

引言

在智慧医疗的大背景下，促进中医药发展已上升为国家发展战略。中医典籍是中医药学的精髓，如何从海量、异构的中医典籍中获取可理解、可应用的经验知识，辅助医生临床决策，已成为中医数字化的重要环节。而命名实体识别（named entity recognition, NER）[1,2]能从文本中自动识别专有名称并加以归类，是知识获取的基础任务。在医学领域主要从医案、病历和文献中识别病名、药方等常见医学术语[3,4]，以供进一步查询或分析，一直以来都是研究重点。

传统的医学领域命名实体识别方法主要分为两类：1）基于规则[5]的方法需针对特定领域人工构造规则，需耗费大量人力，可扩展性较差，不适用于大数据。2）基于统计的方法主要包括隐马尔科夫模型（Hidden Markov Model, HMM）[6]、条件随机场模型（Conditional Random Fields, CRF）[7]和支持向量机（Support Vector Machine, SVM）[8]等。中医领域主要利用 CRF 模型进行实验，均获得了不错效果，如王世昆、张五辈[9,10]等人分别从明清古医案、《名医类案》中的进行了病症、方剂、药材等中医术语抽取，F1 值均达到了 80% 以上；孟洪宇[11]等人利用 CRF 从中医典籍《伤寒论》中进行术语抽取，F1 值为 75.56%。该方法虽无需构建特有模式，但严重依赖人工构建特征的准确度，需大量标注语料，且受分词效果制约。随后，Wu[12,13]等人利用深度学习方法进行实体识别，证明了其效果优于最好的 CRF 模型，其突破了传统方法的局限性，有很强的记忆和学习能力，随后被广泛应用于在中医药学领域。例如，张帆[14]从四类疾病语料中抽取了病名、检查等五类医学实体；薛天竹[15]对电子病历进行术语抽取；步君昭[16]等人抽取生物医学文献中的药物名。目前，双向长短期记忆神经网络和条件随机场（BiLSTM-CRF）模型是命名实体识别的主流方法。自 Guillaume[17]等人首次将其应用于实体识别任务后，大量实验证明了它的优越性，如 Jagannatha[18]等人验证该模型对英文电子病历的实体识别效果明显优于传统深度学习方法。

可见，中医领域识别的实体种类仍局限于症候、病名、穴位和药方等简单实体，且对比深度学习方法在公共数据集上 F1 值达到 90% 以上的效果[19]，仍存在较大差距。针对中医典

籍的实体识别更具挑战性，因其主要存在以下难点：1）术语多采用嵌套结构，且存在大量通假字、生僻字、一词多义的情况；2）句式多为复合长句，结构复杂，采用大量的修饰手法；3）很多中医典籍没有标点和分隔符，导致分词不准确。4）中医源于经验积累，名老中医对中医知识的总结阐述没有统一的标准。此外，深度学习方法在中医领域的实验多基于现代医案和电子病历，针对中医典籍的命名实体识别还鲜有研究。因此，本文在研究《黄帝内经》语言特点的基础上，总结扩展了中医典籍的知识体系及可抽取的实体类别。针对中医领域实体识别的难点，提出结合字向量的 BiLSTM-CRF 模型，探究其在中医典籍中的识别效果，最终实现了对典籍中生理、病理、诊法等更全面的多实体识别。

1 相关技术

1.1 条件随机场

条件随机场最早由 Lafferty[20] 等人提出，可看作是一种无向图模型或马尔科夫随机场，用于标记和切分结构化数据的统计框架模型。它够进行序列概率的全局归一化，且能自由设定序列的特征函数标记序列，避免对输出序列做条件独立性假设，很好地解决了标注偏置问题，被广泛应用于如分词、词性标记、命名实体识别等自然语言处理的相关任务中。条件随机场模型是在给定随机变量 X 的条件下，随机输出变量 Y，目标是构建条件概率模型 P(Y|X)，满足马尔科夫性：

$$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v)$$ （公式 1）

式中 $w\sim v$ 表示无向图中与结点 v 有边连接的所有结点 w，表示结点 v 以外的所有结点，Y_v，Y_u 与 Y_w 为结点 v，u 与 w 对应的随机变量。

命名实体识别任务中常用线性链条件随机场（linear chain CRF）[21]，给定观测序列 $X=\{X_1, X_2, X_3....X_T\}$ 及对应的标记序列 $Y=\{Y_1, Y_2, Y_3.... Y_T\}$，$Y$ 的条件概率分布 $P(Y|X)$ 构成条件随机场：

$$P(Y_i|X,Y_1,Y_2,\ldots,Y_{i-1},Y_{i+1},\ldots,Y_T)=P(Y_i|X,Y_{i-1},Y_{i+1}),i=1,2,\ldots,T$$ （公式 2）

对于线性链条件随机场，随机变量 Y 取值为 y 的条件概率有如下形式：

$$P(Y\mid X)=\frac{1}{Z(X)}\exp(\sum_{i=1}^{T}\sum_{k=1}^{K}w_k f_k(t,Y_t,Y_{t-1},X))$$ （公式 3）

其中，

$$Z(X)=\sum_{Y}\exp(\sum_{i=1}^{T}\sum_{k=1}^{K}w_k f_k(t,Y_t,Y_{t-1},X))$$ （公式 4）

$f_k(t,Y_t,Y_{t-1},X)$ 表示当给定输入序列中的位置 t 和输入 X，当前位置的标记 Y_t 和前一个位置的标记 Y_{t-1} 时的第 k 个特征值，w_k 为特征权重，$Z(X)$ 为归一化因子，在所有可能的输出序列上进行求和。条件随机场模型利用前后向算法进行不同序列位置的条件概率和特征期望，使用拟牛顿法等极大化似然估计求解模型参数，利用 Viterbi 算法进行动态规划解码测试序列数据。

1.2 长短时记忆神经网络

由于循环神经网络（Recurrent Neural Network, RNN）[22] 存在“梯度消失”问题，Hochreiter [23] 等人提出了长短时记忆神经网络（Long Short-Term Memory, LSTM），它由记忆单元 c_t、输入门 i_t、输出门 o_t、遗忘门 f_t 构成，从输入 (x_t, h_{t-1}) 到输出 h_t 的一条线为细胞状态。

doi:10.3772/j.issn.2095-915x.2019.01.010

在模型学习和训练过程中，LSTM 通过门单元调整信息衰减、更新和去留的程度，由一个 Sigmoid 神经网络层和一个成对乘法操作组成。该层的输出是一个介于 0 到 1 的数，0 表示完全不允许通过，1 表示允许完全通过。LSTM 的结构流程如下：

首先，决定从细胞状态中丢弃什么信息。遗忘门 f_t 判断过去记忆 c_{t-1} 的重要程度，进而判断让过去的记忆内容多大的程度参与新记忆的生成。

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad （公式 5）$$

下一步，确定什么样的新信息被存放进细胞状态中。输入门 i_t 通过 Sigmoid 来判断当前的单词的重要程度，进而判断让它以何种程度参与生成新的记忆。同时，用一个 tanh 层用来生成新的候补记忆单元 c_in_t，把这两部分产生的值结合来进行更新。

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad （公式 6）$$

$$c\_in_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad （公式 7）$$

接下来，进行细胞状态的更新，得到当前时刻记忆单元 c_t。

$$c_t = f_t * c_{t-1} + i_t * c\_in_t \quad （公式 8）$$

最后一步，决定模型的输出。首先通过输出门 o_t 来确定细胞状态的哪个部分将输出出去。接着把当前时刻细胞状态通过 tanh 进行处理，并将两者综合考虑，仅仅输出确定的那部分，得到隐藏层最终输出 h_t。

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad （公式 9）$$

$$h_t = o_t * \tanh(c_t) \quad （公式 10）$$

但这种网络只考虑了过去序列对当前的影响，忽略了后文信息的作用，导致对模型的效果造成负面影响。因此引入双向长短时神经网络（Bidirectional LSTM, BiLSTM）模型，它能够联结了上文和下文两个方向的 LSTM 单元在同一时刻的输出并给出最终包含上下文信息的隐含层输出，进而提升整体模型的性能。

2 模型与方法

2.1 模型构建

双向长短时记忆神经网络和条件随机（BiLSTM-CRF）模型一共可分为四层，分别是输入层、Embedding 层、BiLSTM 层和 CRF 层。模型的详细结构如图 1 所示：

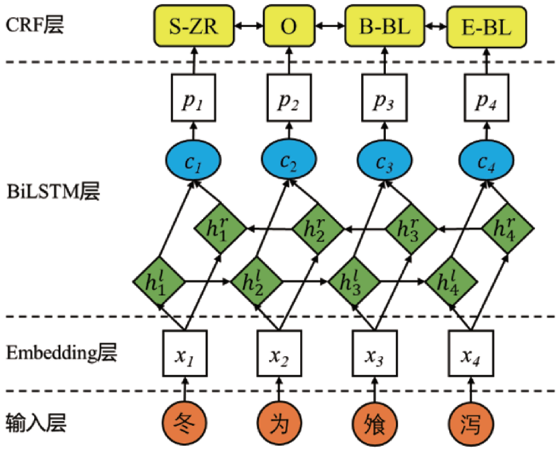

图 1 基于字的 **BiLSTM-CRF** 模型结构

2.1.1 输入层

目前大多模型的输入都是词语，导致实体识别的效果严重依赖于分词的效果。而中医典籍中的“字”就包含了大量语言信息，因此该模型以“字”作为初始输入，规避了分词效果不佳带来的错误累积。将一个包含 n 个字的句子记作 $W=(w_1,w_2,w_3,\ldots,w_n)$，构成一个字典，其中 w_i 是句子的第 i 个字在字典中的 id，维数是字典大小，即字的个数。

2.1.2 Embedding 层

为解决深度神经网络模型训练数据集规模与需训练参数不匹配问题，用高质量的预训练结果进行参数初始化，得到更好的效果。将输入的各个字 w_i 利用预训练的 Embedding 矩阵映射为新的低维稠密的字向量 $X=(x_1,x_2,x_3,\cdots,x_n)$，传递给 BiLSTM，其中 x_i 就是预训练中维度指定为 d 的向量。为防止过拟合，加入了一层 Dropout 正则化机制。

2.1.3 BiLSTM 层

该层用于自动提取句子特征。将字向量 $X=(x_1,x_2,x_3,\cdots,x_n)$ 作为 BiLSTM 各个时间步的输入，经过前向 LSTM 得到了左侧每个字的输出隐状态 $H^L=(h_1^l, h_2^l, h_3^l, \ldots, h_n^l)$。同理，经过后向 LSTM 得到了右侧的输出隐状态 $H^R=(h_1^l, h_2^l, h_3^l, \ldots, h_n^l)$。拼接得到各个位置输出的隐状态 $c_i=[(h_i^r, h_i^r)]$，最终得到完整的隐状态序列 $C=(c_1, c_2, c_3, \cdots, c_n)$。在加入 Dropout 后利用一个全连接层 (U,b)，将隐状态向量映射到 k 维，k 是标注集的标签数，从而得到自动提取的句子特征，即 $P=(p_1, p_2, \ldots, p_n)$，$p_i$ 的每一维 $p_{i,j}$ 都可看成是将字 w_i 分类到第 j 个标签的分值。

2.1.4 CRF 层

该层进行句子级的序列标注，保证在全局上生成最优标注序列。BiLSTM 层输出的 p_i 是相互独立的，忽略了前后标注结果之间具有强依赖性。利用 CRF 层可从训练数据中自动获得一些约束性规则，从而对整个句子进行联合建模，降低非法序列出现的概率，提升了标签序列预测的准确率。考虑到需在句子首部添加一个起始状态，在句子尾部添加一个终止状态，因此 CRF 层的参数是一个 (k+2)×(k+2) 的状态转移矩阵 A。$A_{i,j}$ 表示的是从第 i 个标签到第 j 个标签的转移得分。假设输入一个句子 $W=(w_1,w_2,w_3,\cdots,w_n)$，得到一个预测标签序列 $y=(y_1, y_2, \ldots, y_n)$，那么定义该预测的得分为：

$$s(W,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \quad （公式 11）$$

其中 P_{i,y_i} 为第 i 个位置 BiLSTM 输出为 y_i 的概率，$A_{y_i,y_{i+1}}$ 为从 y_i 到 y_{i+1} 的转移概率，整个序列得分等于各位置的得分之和，每个位置的得分由 BiLSTM 输出的 p_i 和 CRF 的转移矩阵 A 共同决定。例如假设 BiLSTM 输出的最有可能序列为 BBEO，但是转移概率矩阵中 B->B 的概率很小甚至为负，s 得分就会降低，那么就排除了这个不合理的预测序列。

对于每个训练样本 W，利用 Viterbi 算法求出所有可能的标注序列 y 的得分 s(W,y)，然后利用 Softmax 层对所有得分实现归一化，最终得到序列 y 的概率为：

$$p(y\mid W)=\frac{e^{s(W,y)}}{\sum_{\bar{y}\in Y_W}e^{s(W,\bar{y})}} \quad （公式 12）$$

模型训练时，对于句子输入序列 X，损失函数设置为对目标真实标记序列 Y 的概率取对数。为了使真实标记序列对应的概率最大化，采用取负值然后最小化的方法，引入梯度下降算法来求解参数，最大化 log 似然函数：

$$\log(p(Y\mid X))=s(X,Y)-\log(\sum_{\bar{Y}\in Y_x}e^{s(X,\bar{Y})})$$
$$=s(X\mid Y)-\log(\sum_{\bar{Y}\in Y_x}s(X,\bar{Y})) \quad （公式 13）$$

在预测过程时，根据训练好的参数求出所有可能的 y 序列对应的 s 得分，使用动态规划的 Viterbi 算法来求解最优路径，预测结果记为 Y*：

doi:10.3772/j.issn.2095-915x.2019.01.010

$$Y^* = \arg\max_{\overline{Y} \in Y_X} (s(X, \overline{Y})) \quad （公式 14）$$

2.2 框架设计

2.2.1 特征提取

中医是一种“分类医学”[24]，通过对阴阳五行、人体自然的认识，对人体、病因、病机、病症、治法、方药等建立了独特的分类体系。本文选取《黄帝内经》为实验语料，以张德政教授[25]提出的基于本体的中医知识体系为指导，根据对文本的理解与分析，将中医典籍实体类别划分为中医认识方法、中医生理、中医病理、中医自然、治则治法五大类。其中，中医认识方法包括阴阳、五行学说、天干地支、数字等概括总结形成的术语；中医生理包括脏腑、形体、官窍、经络穴位、气血、津液等概念；中医病理包括病名、病因、病机、症状等概念；中医自然包括了季节、方位、时间、颜色、味道、动植物等实体；治则治法包括治则、治法、方剂名、中草药名等概念。

数据集采用 BIOES 标注方式，即 B 表示实体的开始，I 表示实体的中间部分，E 表示实体的结尾，S 表示单个字符的实体，非实体部分用 O 表示。同时 FF 表示中医认识方法，ZR 表示中医自然，SL 表示中医生理，BL 表示中医病理，ZF 表示治则治法。具体的标注如表 1 所示：

表 1 《黄帝内经》实体标注 **BIOES** 标签表

实体类别	单字符实体	多字符实体		
	标记	开始标记	中间标记	结束标记
中医认识方法	S-FF	B-FF	I-FF	E-FF
中医自然	S-ZR	B-ZR	I-ZR	E-ZR
中医生理	S-SL	B-SL	I-SL	E-SL
中医病理	S-BL	B-BL	I-BL	E-BL
治则治法	S-ZF	B-ZF	I-ZF	E-ZF
非实体标记	O	O	O	O

2.2.2 实验流程

中医典籍命名实体识别实验的整体流程是一个迭代寻找最优模型的过程，如图 2 所示。

实验以《黄帝内经》为语料，从网络中爬取全文后，进行数据预处理，包括清除乱码、统一标点符号、将繁体字全部转换为简体字、统一通假字的转换，如“四支”统一为“四肢”、“五藏”统一为“五脏”等。此外，为获得 Embedding 层所需的预训练字向量，爬取了 701 本中医典籍，融合语料库中拥有的 30 万份名老中医医案和《中华历代名医医案全库》中 1 万多份古医案，对其进行合并以及数据预处理，随后拆分成字，共得到 3.84G 的预训练语料。使用 Google 开源的词向量生成工具 gensim Word2Vec 进行字向量的训练，将 Skip-gram 模型窗口大小设为 10，得到 200 维的字

向量。

为减少人工标注语料的工作量及难度，采用自定义词表匹配方式对数据集进行自动 BIOES 标注，并辅以人工校对方式快速构建所需的实验语料。该数据集共标注了 27642 个样本，将其中的 60% 作为训练集，用于迭代生成模型，20% 作为验证集来选择最优训练模型，20% 作为测试集，查看当前模型的识别效果，防止过拟合。各类别标注数量如表 2 所示：

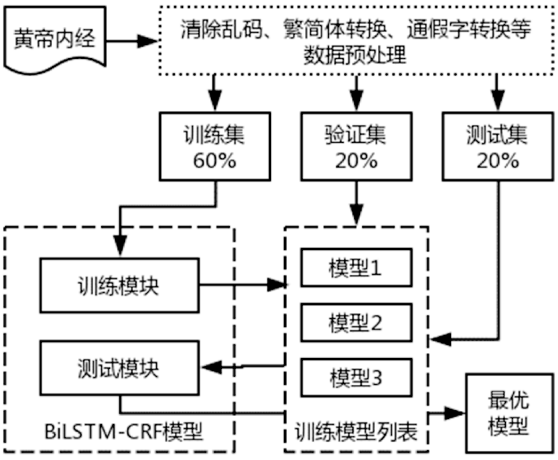

图 2 《黄帝内经》中医实体识别流程图

表 2 《黄帝内经》实体标注实验语料表

数据集	标注数量	中医认识方法	中医自然	中医生理	中医病理	治则治法
训练集	16585	2515	2862	8449	2252	507
验证集	5428	859	903	2816	712	138
测试集	5629	818	1005	2817	791	198

2.2.3 评价指标

针对基于深度学习的中医典籍实体识别性能，采用以下 3 个指标进行评估：

1) 精确率（Precision, P）是正确被检索的实体数占所有实际被检索到的实体数的比例。

$$P(\%)=\frac{\text{正确的实体数}}{\text{抽取出的实体数}}\times 100 \quad \text{（公式 15）}$$

2) 召回率（Recall, R）是所有正确被检索实体数占所有应该被检索到的实体数的比例。

$$R(\%)=\frac{\text{正确的实体数}}{\text{语料中包含的实体数}}\times 100 \quad \text{（公式 16）}$$

3) F1 值（F1 Score）是对精确率和召回率的调和均值，它表示对精确率和召回率的综合考量。

$$F_1(\%)=\frac{2P\bullet R}{P+R}\times 100 \quad \text{（公式 17）}$$

3 实验与对比

3.1 实验实现

3.1.1 环境与参数设置

本文实验环境为：1) 操作系统为 Red Hat 4.8.5；2) 内存 120G；3) 硬盘 500G；4) CPU 为至强 Xeon E5-2600V2，8 核；5) 显卡为华硕 GTX1080TI，11G。

模型中，LSTM 的神经元数量设置为 50 个。为加快模型训练速度，将中医典籍中的长句按照最长 20 个字分割。为解决梯度爆炸问题，采用 clip gradients，将梯度阈值 clip 设置为 5。为防止过拟合现象，引入 Dropout 正则化机制，设定 Dropout 层的概率为 0.5。模型采用反向传播算法拟合训练数据，针对每个训练样例更新参数。为改善模型训练方式，优化调整模型更

doi:10.3772/j.issn.2095-915x.2019.01.010

新权重和偏差参数的方式，采用 Adam 梯度下降算法。此外，还有一些超参数的设置，如：学习率、迭代次数、批大小等。

表 3 命名实体识别模型超参设置表

参数	值	参数	值
Segmentation dimension	20	Learn rate	0.001
Hidden units number	50	Batch_size	20
Clip gradients	5	Max_epoch	200
Dropout rate	0.5	Steps_check	50

3.1.2 实验结果

实验结果为运行 3 次得到的最优模型各项分值的平均值，如表 4 所示：精确率 85.44%，召回率 85.19%，F1 值 85.32%。结合实验结果和标注语料分析，可知实体识别效果受标注量及实体区分度的影响。中医认识方法实体的标注量较多且区分度最高，因此识别效果最好；中医生理的标注量至少为其他类别标注量的 3 倍，而治则治法训练样本虽少，但实体区分度很高，因此这两类的效果也较好。中医自然的实体标注量虽较多，但与其他类别的冲突较多，如“水”有的属于五行，有的属于自然，而属于五行的概率远高于自然，降低了识别准确率。而中医病理的词语大多边界不清，大部分是中医生理与病症的结合，嵌套现象严重，极大程度影响了实体识别的精度，因此识别效果最差。

表 4 各个实体类别实验结果

实体	Precision (%)	Recall (%)	F1 (%)
模型均值	85.44	85.19	85.32
中医病理（BL）	68.47	62.81	65.52
认识方法（FF）	93.86	86.99	90.30
中医生理（SL）	88.35	87.05	87.70
中医自然（ZR）	81.58	84.94	83.19
治则治法（ZF）	86.36	70.37	77.55

3.2 验证对比

为验证本文所提出的模型和使用的参数在中医典籍命名实体识别中的有效性，设计从不同模型效果、向量维度效果、组件参数效果进行对比实验。

3.2.1 不同模型效果对比

以传统方法中最常用 CRF 模型对中医典籍的识别结果为 baseline，对比原始 LSTM 模型、双向 LSTM 模型、LSTM 与 CRF 的结合模型以及 BiLSTM-CRF 模型的效果。同时，为充分利用 GPU 的处理性能，引入迭代空洞卷积神经网络和 CRF 结合（IDCNN-CRF）模型作为对比。在相同的语料和实体分类的情况下进行实验。

表 5 实体识别模型不同模型实验结果

序号	模型	参数组合	F1 (%)
1	CRF	字符、词边界、类别标签	75.56
2	LSTM	字向量、pretrain、dropout	82.29
3	BiLSTM	字向量、pretrain、dropout	82.48
4	LSTM-CRF	字向量、pretrain、dropout	84.67
5	BiLSTM -CRF	字向量、pretrain、dropout	85.32
6	IDCNN-CRF	字向量、pretrain、dropout	85.01

由表 5 可知，实验 1 效果最差，说明深度学习方法提取长句特征能力优于 CRF 模型。实验 3 优于实验 2，说明结合过去和未来特征的双向 LSTM 确实可提升效果。实验 4、5、6 说明加入 CRF 层后考虑了标签之间的依赖关系，效果提升明显。同时实验 5 效果最好验证了本文提出的模型确实较好地解决了中医典籍的实体识别问题，提升了标记的准确性，减少了对词嵌入的依赖。而 IDCNN 虽然网络结构比 LSTM 复杂，参数较多，但是处理速度远高于

LSTM。

3.2.2 字向量维度效果对比

为对比不同字向量维度对实体识别的影响，使用 Word2Vec 针对相同训练语料预训练得到 50 维至 400 维的字向量进行实验。实验结果如表 6 所示，是一个先增后减的过程。50 维、100 维不能对字进行充分表征。随字向量维度提高，模型训练过程中与实验数据的拟合情况逐渐变好，在 200 维时达到最优平衡状态。若维度继续增大，会因训练语料规模的限制，无法支持所需的参数得到充分训练，导致效果下降。实验证明，采用的 200 维字向量是针对中医典籍实体识别的最佳维度。

3.2.3 不同参数效果对比

为对比模型中组件的组合对实验结果的影响，设计对比试验：如选择预训练 pretrain、在 BiLSTM 输入端添加一个 L-dropout、在输出端添加一个 R-dropout。结果证明，pretrain 字向量预训练可更好的表征字的特征，更好地初始化 embedding 层的参数，效果优于随机初始化矩阵。而 dropout 也会影响实验效果，两种 embedding 方式下均为单独使用 R-dropout 效果更好。实验证明，采用预训练字向量机制，并且引入 dropout 防止过拟合，确实能够提升模型的识别效果。

表 6 实体识别模型不同字向量维度实验结果

维度	Precision (%)	Recall (%)	F1 (%)
50	83.66	84.11	83.88
100	84.76	85.67	85.21
200	85.44	85.19	85.32
300	85.43	84.46	84.94
400	85.42	84.24	84.82

表 7 实体识别模型不同参数组合实验结果

embedding 组件	dropout 组件	Precision (%)	Recall (%)	F1 (%)
Pretrain	L-dropout+R-dropout	85.91	85.07	85.32
	L-dropout	85.37	86.04	85.64
	R-dropout	85.47	86.21	85.75
	-	84.76	85.29	85.02
Random	L-dropout+R-dropout	83.21	85.47	84.33
	L-dropout	84.66	85.28	84.96
	R-dropout	84.52	85.42	84.97

4 结论

针对中医典籍语言表述复杂、难以理解，实体识别存在着分词和人工构建特征不准确的问题，根据文本特点和以往研究，归纳了中医典籍的 5 类实体进行特征提取，进一步提出了一种结合字向量及深度学习神经网络模型的中医典籍命名实体识别方法，以《黄帝内经》为语料的基础上获得了不错的识别效果。该方法弥补了深度学习方法在中医典籍领域实体识别研究的空白。此外，多角度设计对比实验，结果表明 : 与其他方法相比，该方法的 F1 值更高，泛化能力和鲁棒性更强，有效弥补了传统方法的不足，是中医典籍命名实体识别的理想方法。

doi:10.3772/j.issn.2095-915x.2019.01.010

综合识别结果来看，目前该方法仍有进一步优化的空间，其识别效果主要受两个因素制约：1) 训练数据规模，现有的参数规模无法很好地支撑模型所需的参数，影响模型学习效果；2) 实体标注粒度。现有的实体类别使得中医病理类的嵌套现象严重，譬如“肝痹”可以进一步拆分为“肝”和“痹”，分别属于生理和病理，极大程度影响了实体识别的精度。未来随着迁移学习、无监督学习技术的发展以及比字向量更细粒度的部首向量特征表示方式，必将在较小数据规模和较少标注的情况下取得更好的识别效果。同时，粗细粒度两种方式联合标注，配合实体对齐、语义消歧技术，也能解决中医典籍的一词多义问题，提高识别准确率，识别有层次概念的中医实体。

参考文献

[1] Sundheim B, Sundheim B. Message Understanding Conference-6: a brief history[C]. Conference on Computational Linguistics. Association for Computational Linguistics, 1996: 466-471.

[2] Chinchor N. MUC-7 Named Entity Task Definition[C]. In: Proceedings of the 7th Message Understand ing Conference, Virginia. 1998.

[3] 胡双 , 陆涛 , 胡建华 . 文本挖掘技术在药物研究中的应用 [J]. 医学信息学杂志 , 2013, 34(8): 49-53.

[4] 范岩 . 基于条件随机场模型的中医文献知识发现方法研究 [D]. 北京 : 北京交通大学 , 2009.

[5] Fukuda K, Tsunoda T, Tamura A, et al. Towards information extraction: identifying protein names from biological papers[C]. Proc. Pacific Symposium on Biocomputing, 1998(3): 707-718.

[6] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]. Conference on Applied Natural Language Processing. 1997.

[7] Mccallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]. Conference on Natural Language Learning at Hlt-Naacl. Association for Computational Linguistics, 2003: 188-191.

[8] Asahara M, Matsumoto Y. Japanese Named Entity extraction with redundant morphological analysis[C]. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003: 8-15.

[9] 王世昆 , 李绍滋 , 陈彤生 . 基于条件随机场的中医命名实体识别 [J]. 厦门大学学报 (自然科学版), 2009, 48(3): 359-364.

[10] 张五辈 , 白宇 , 王裴岩 , 等 . 一种中医名词术语自动抽取方法 [J]. 沈阳航空工业学院学报 , 2011, 28(1): 72-75.

[11] 孟洪宇 , 谢晴宇 , 常虹 , 等 . 基于条件随机场的《伤寒论》中医术语自动识别 [J]. 北京中医药大学学报 , 2015, 38(9): 587-590.

[12] Hiton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786): 504-507.

[13] Wu Y, Jiang M, Lei J, et al. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network[J]. Study Health Technology Information, 2015(216): 624-628.

[14] 张帆 , 王敏 . 基于深度学习的医疗命名实体识别 [J]. 计算技术与自动化 , 2017, 36(1): 123-127.

[15] 薛天竹 . 面向医疗领域的中文命名实体识别 [D]. 哈尔滨 : 哈尔滨工业大学 , 2017.

[16] 步君昭 . 生物医学文献中的药物名抽取方法研究 [D]. 哈尔滨 : 哈尔滨工业大学 , 2016.

[17] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J]. 2016: 260-270.

[18] Jagannatha A, Yu Hong. Structured prediction models for RNN based sequence labeling in clinical text[C].

Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 856-865.

[19] Zhiheng Huang, Wei X, Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv, 2015, 1508.01991.

[20] Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, 2001: 282-289.

[21] Getoor L, Taskar B. An Introduction to Conditional Random Fields for Relational Learning[J]. Foundations & Trends® in Machine Learning, 2010, 4(4): 267-373.

[22] Mikolov T, Kombrink S, Burget L, et al. Extensions of recurrent neural network language model[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2011: 5528-5531.

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

[24] 李崇超 . 论中医知识体系中的分类 [J]. 中医杂志 , 2015, 56(21): 1804-1807.

[25] 张德政 , 谢永红 , 李曼 , 等 . 基于本体的中医知识图谱构建 [J]. 情报工程 , 2017, 3(1): 35-42.