深度学习基础上的中医实体抽取方法研究*

张艺品 关 贝 吕荫润 王 翀

（1 中国科学院软件研究所协同创新中心 北京 100190）
（2 中国科学院大学 北京 100049）
（1 中国科学院软件研究所协同创新中心 北京 100190）
（2 中国科学院大学 北京 100049）
（3 中国科学院软件研究所计算机科学国家重点实验室 北京 100190）

吴炳潮 王永吉

（1 中国科学院软件研究所协同创新中心 北京 100190）
（2 中国科学院大学 北京 100049）
（1 中国科学院软件研究所协同创新中心 北京 100190）
（2 中国科学院大学 北京 100049）
（3 中国科学院软件研究所计算机科学国家重点实验室 北京 100190）

毕诗旋

（北京工业大学 北京 100190）

〔摘要〕 介绍命名实体识别及模型应用研究情况，以中医典籍作为数据源，采用深度学习方法，进行中医疾病、方剂、中草药等实体抽取，设计 BiLSTM－CRF 序列标注模型，构建中医典籍实验语料进行实验，结果表明该模型算法具有高度准确性。

〔关键词〕 知识图谱；实体抽取；中医；深度学习
〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2019.02.012

Study on the Entity Extraction Method of Traditional Chinese Medicine on the Basis of Deep Learning *ZHANG Yipin, 1X－Lab Institute of Software, Chinese Academy of Sciences, Beijing 100190, China; 2University of Chinese Academy of Sciences, Beijing 100049; GUAN Bei, LV Yinrun, WANG Chong, 1X－Lab, Institute of Software, Chinese Academy of Sciences, Beijing 100190; 2University of Chinese Academy of Sciences, Beijing 100049; 3State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China; WU Bingchao, 1X－Lab, Institute of Software, Chinese Academy of Sciences, Beijing 100190; 2University of Chinese Academy of Sciences, Beijing 100049, China; WANG Yongji, 1X－Lab, Institute of Software, Chinese Academy*

〔修回日期〕 2018－09－28
〔作者简介〕 张艺品，硕士研究生；通讯作者：关贝，助理研究员。
〔基金项目〕 科技部国家重点研发计划重点专项（项目编号：2017YFB1002300）；大数据驱动的中医智能辅助诊断服务系统课题一“多模态异构中医药大数据高效获取与资源库建设”（项目编号：2017YFB1002301）和课题三“基于深度学习的中医多尺度认知方法和辩证论治分析模型”（项目编号：2017YFB1002303）。

of Sciences, Beijing 100190; 2University of Chinese Academy of Sciences, Beijing 100049; 3State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China; BI Shixuan, Beijing University of Technology, Beijing 100124, China

〔**Abstract**〕 The paper introduces the study on named entity recognition and model application, conducts extraction of entities such as Traditional Chinese Medicine (TCM) diseases, prescription, Chinese herbal medicine, etc., by adoption of deep learning method and taking TCM classics as data sources, designs the model for sequencing tagging - BiLSTM - CRF. It also conducts experiments by building corpus of experiments in TCM classics. The result shows that the aforesaid model algorithm is of high accuracy.

〔**Keywords**〕 knowledge graph; entity extraction; Traditional Chinese Medicine (TCM); deep learning

1 引言

中医知识图谱是对中医知识内容的结构化存储、语义化表示，有助于实现中医知识的检索、推送与可视化，也是中医辅助诊断系统的底层数据核心。知识抽取是知识图谱构建研究的重要问题，其分为两个模块，即抽取实体与实体间关系，本文在中医领域中就实体抽取问题进行研究。

中医典籍是中医知识的重要来源，从中医典籍中抽取知识是中医知识图谱规模化和知识全面化的必然要求。中医典籍作为一种非结构化的自然语言，其文法和词法与白话语言多有不同。以《备急千金方》的一段方剂为例，观察发现中医典籍中有大量的生僻词，且语法与现在普通汉语有极大的不同，分词困难。此外中医方剂往往参照其主药命名，命名实体间易存在嵌套关系，如“泽兰汤”与“泽兰”、“茯神煮散”与“茯神”、“半夏补心汤”与“半夏”等。另外不同医书对于中医内容的阐述习惯不同，没有统一标准。

传统实体抽取方法需要手动构建特征，如词性标注、解析语法树等，这些方法难以适用于中文文言文形式的中医典籍。为适应中医典籍的特点需手动构建更为复杂的特征，这无疑会在源数据中引入大量额外内容增加计算负荷。此外特征构建需要详细的设计，耗时耗力且迁移性较差。为解决上述问题，本文采用深度学习方法来实现知识获取模块的实体抽取。根据中医典籍文本特点，设计字级别的基于 BiLSTM - CRF 的实体抽取模型。通过字向量来表征字所携带的语义信息，利用双向的长短期记忆神经网络获取字的上下文表征向量。最后将上下文表征向量作为条件随机场的直接输入，利用其求取全局最优标注序列的特性，完成整个句子的序列标注。

2 相关研究

2.1 命名实体识别

实体抽取是知识图谱自动化构建的核心，决定知识图谱的质量与规模。实体抽取又称为命名实体识别（Named Entity Recognition，NER），是自然语言处理（Natural Language Processing，NLP）中的一项基础任务，是指从文本中识别出命名性指称项，如人名、地名和组织机构。在不同领域中其定义的实体类型也不同，本研究抽取的实体是中医领域的中药材、方剂、病症 3 类实体。命名实体抽取的技术与方法主要有 3 种：基于规则和知识的方法、基于统计的方法以及两者混合的方法。基于规则和知识的方法通过观察文本结构特征，人工设计规则，利用正则表达式等方式抽取实体，对于简单的命名实体识别任务十分简捷高效[1]，但对于复杂的任务要避免规则间的相互冲突，制定规则需要消耗大量的时间和精力，可移植性差。基于统计学习的方法将命名实体识别看成序列标注问题[2]，采用隐马尔可夫模型（Hidden Markov Model，HMM）、最大熵模型、条件随机场等机器学习序列标注模型[3-6]。Zhou 和 Su 等[7]设计 4 种不同的特征，提高隐马尔可夫模型在实体抽取任务中的效果。Borthwick 等[8]在最大熵模型中通过引入额外的知识集合提高标注的准确性。Lafferty 等[9]提出将条件随机场用于序列标注任务，后来 McCallum 和 Li[10]提出特征自动感应法，将条件随机场应用于命名实体识别任务中。

2.2 模型应用

近年来随着深度学习的兴起，将深度神经网络

模型应用到 NER 任务中取得较好的效果。Collobert 等[11]于 2011 年提出统一的神经网络架构机器学习算法，将窗口方法与句子方法两种网络结构用于 NER，对 NN/CNN - CRF 模型进行对比试验。2013 年 Zheng 等[12]在大规模未标记数据集上改进中文词语的内在表示形式，使用深度学习模型发现词语的深层特征以解决中文分词和词性标记问题。借鉴 Collobert 的思路 2016 年左右出现一系列使用循环神经网络（Recurrent Neural Network，RNN）结构并结合 CRF 层进行 NER 的工作[13-15]，模型主要由 Embedding 层，双向 RNN 层以及最后的 CRF 层构成。2016 年 Rei M 等[16]在 RNN - CRF 模型结构的基础上重点改进词向量与字符向量的拼接方法，使用 Attention 机制将原始的字符向量和词向量拼接改进为权重求和。模型通过双层神经网络学习 Attention 的参数，动态地利用词向量和字符向量信息，实验表明比原始的拼接方法效果更好。之后 AkashBharadwaj 等[17]在原始 BiLSTM - CRF 模型的基础上加入音韵特征，在字符向量上使用 Attention 机制来学习关注更有效的字符。2017 年 Yang Z 等[18]采用迁移学习的方式在小量标注数据集进行实体识别任务。同年 Matthew E. Peters 等[19]使用海量无标注语料库训练双向神经网络语言模型，通过这个模型来获取当前要标注词的语言模型向量，然后将该向量作为特征加入到原始的双向 RNN - CRF 模型中。实验结果表明在少量标注数据上加入这个语言模型向量能够大幅度提高 NER 效果，即使在大量的标注训练数据上加入这个语言模型向量仍能提高原始 RNN - CRF 模型效果。本文采用基于 BiLSTM - CRF 的方法，利用双向 RNN 累积获取上下文信息的特性，条件随机场对全部特征全局归一化，获取全局最优解。

3 基于深度学习的中医实体抽取

中医典籍存在大量通假字、专业术语与古语句式，加大分词的难度，难以获取好的分词文本。低质量的分词无疑会直接影响命名实体识别的效果。语言中语素是最小的有意义单位，词是最小的能够独立使用单位。如词"production" 由两个语素"product" 和名词词缀" - tion" 组成，其中"product" 也是一个词可以独立使用，而" - tion" 只是一个语素，不能独立使用。汉语中类似于英语单词这样的形式并不是现成的，字不仅是一个音节单位，同时也可以是承载意义的单位，即语素。依照汉语中字的信息携带特性提出字级别的模型，见图 1。同时可以规避分词效果不佳带来的错误累积问题。所以模型的第 1 层以中文汉字的序列作为初始输入。第 2 层为 embedding 层，旨在将第 1 层输入的字序列中各个字映射为相应的向量。在深度神经网络模型中往往有大量的参数需要学习，但是训练数据集的大小是有限的。为解决这个问题，基于无监督或自监督的学习方法得到广泛使用。模型使用优质的预训练结果进行参数的初始化可以取得更好的效果。使用 word2vec 进行预训练得到词向量，作为 embedding 层的初始化参数。对于初始输入的字序列 $W = (w_1, w_2, \cdots, w_m)$，经过 embedding 层得到 $X = (x_1, x_2, \cdots, x_n)$，其中 x_i 是维度指定为 d 的向量。第 3 层为双向的长短期记忆（Long Short Term Memory，LSTM）网络结构，通过双向结构在对当前位置进行标注时双向循环神经网络用于能充分考虑上下文的信息，得到当前字的上下文表征向量。对于从 embedding 输入的 $X = (x_1, x_2, \cdots, x_n)$，经过前向 LSTM 得到左侧每个字的表征上文 $H^l = (h_1^l, h_2^l, h_3^l, \cdots, h_n^l)$。同理经过后向 LSTM 得到右侧的表征下文为 $H^R = (h_1^r, h_2^r, h_3^r, \cdots, h_n^r)$。最终得到上下文表征向量序列为 $C = (c_1, c_2, \cdots, c_n)$，其中 $c_i = [h_i^l, h_i^r]$。

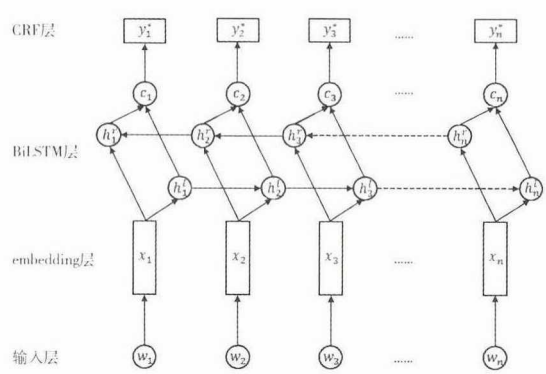

图 1 基于字的 BiLSTM - CRF 模型结构

最后为 CRF 层，目标是依照上下文表征向量序列得到标注序列。可以依照每个词的上下文表征向量 c_i，使用分类算法如 softmax 直接得出每个字标注

目标的概率。但是在命名实体识别任务中前后的标注结果之间具有强依赖性。如药方实体的起始（B－PRE），接下来一定不可能是中药材实体的内部内容（I－MED）。因此用 CRF 对整个句子进行联合建模，以保证在全局上生成最优标注序列。对于双向 LSTM 输出的 C，通过全连接层（W，b），将其转换为 $n \times k$ 的矩阵 P，其中 k 为目标标签的种类数量。$p_{i,j}$ 表示第 i 个字标记为第 j 个标签的得分。对于给定预测序列 Y =（y_1，y_2，…y_n），其得分为：

$$s(X,Y) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \tag{1}$$

其中 A 为条件随机场的状态转移矩阵，是 $k+2$ 阶的方阵，$A_{i,j}$ 表示从标签 i 转移到标签 j 的得分。之后接入 softmax 层实现归一化，对于序列 Y，其概率为：

$$p(Y \mid X) = \frac{e^{s(X,Y)}}{\sum_{\bar{Y} \in Y_x} e^{s(X,\bar{Y})}} \tag{2}$$

在训练阶段，对于 X 的正确标注序列 $T = (t_1, t_2, \cdots, t_n)$，用梯度下降算法最大化 log 似然函数：

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

$$= s(X,Y) - \log \sum_{\bar{Y} \in Y_x} s(X,\bar{Y}) \tag{3}$$

在预测阶段，预测结果记为 Y*：

$$Y^* = argmax(S(X,\bar{Y}))$$

$$\bar{Y} \in Y_x \tag{4}$$

通常通过维特比算法等动态规划方法得到最优解。

4 实验结果与分析

4.1 Word2Vec

模型中 embedding 层使用 Word2vec 训练的字向量初始化参数，在模型训练时进行参数微调。使用 Google 开源的 gensium[20] 工具包，其提供 python 版本的 Word2vec。语料方面，使用维基百科提供的 Latest Chinese Article 语料[21] 与中医典籍语料的拼接作为 Word2vec 的训练语料。在经过 opencc 繁简体转换、移除 non－utf8 字符、统一不同类型标点符号、空格处理、分字得到 2.01G 的未标注语料。字向量的训练使用 CBoW 模型，窗口大小设为 10，字向量维度设为 300。

4.2 数据集和评价指标

实验选用中医典籍《备急千金方》、《千金翼方》、《神农本草经》作为语料。在实体抽取算法实验中，为减小模型训练的计算复杂度，加快训练速度，将长句分割为不超过 50 字的短句。通过字典匹配与人工校对方式构建试验语料，选用万方中医药知识库的病症方剂中药材字典。训练数据集共 39 100 条标注样本，其中 32 000 条为训练集，7 100条为测试集，共设置 3 类命名，即中药材、方剂、病症实体。数据集采用 BIO 标注体系，即 B 表示实体的起始，I 表示实体的中间内容，O 表示非实体内容，具体的标注方法与语料中各类实体，见表 1。

表 1 实体标注标签

实体项	标注序列	实体数量
中药材	B－MED I－MED	25 092
方剂	B－PRE I－PRE	7 855
病症	B－DES I－DES	3 063
其他	O	—

注：B－MED 中的 B 代表药材实体的起始，MED 代表药材实体，依此类推

实验使用精确率（Precision，P），召回率（Recall，R）以及 F－Score（F）来评价算法性能。精确率 P 在本文表示预测为正的样本中有多少是真正的正样本，召回率 R 表示样本中的正例有多少被预测正确。F－Score 是对精确率和召回率的调和均值，表示对精确率和召回率的综合考量。F－Score 计算公式如下：

$$F_\alpha = \frac{(\alpha^2+1)P \cdot R}{\alpha^2 P + R} \tag{5}$$

其中 α 用于度量精确率与召回率的权重，当 $\alpha > 1$ 时，召回率对 F 值的影响更大，当 $\alpha < 1$ 时精确率对 F 值得影响更大，当 $\alpha = 1$ 时，此时的 F 值即为 F_1－Score：

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{6}$$

4.3 基于 BiLSTM－CRF 的实体抽取算法实验

将 LSTM 的维度设定为 embedding 层输出维度

的一半。为防止过拟合，在训练阶段模型中引入 dropout 正则化机制，将其置于 BiLSTM 层的输出端，设定 dropout 层概率为 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。具体采用 Adam 梯度下降算法，学习率为 0.001，共训练 50 轮，batch_ size 为 16。LSTM 基本单元选用无窥视孔，遗忘门初始化偏置为 1，独立的遗忘和输入门限。为验证模型的有效性选取 CRF、HMM 进行效果比较。实验结果，见表 2。HMM 模型在 3 种模型中效果最差，F_1 值仅有 68.01。CRF 模型采用窗口尺寸为 3 的上下文特征。相比于 HMM 模型，CRF 在 F_1 等比增长 23.71。HMM 仅利用上一个字的信息，信息搜集跨度较小，而 CRF 采用取［-3，+3］跨度的局部特征，可以更好地搜集上下文信息，对训练数据的拟合性更好。此外 HMM 还存在标注偏置问题，对特征的融合能力也较弱，而 CRF 模型能够解决标注偏置与特征融合问题。本文提出的 BiLSTM-CRF 模型在中医实体抽取上效果最好，F_1 值较 CRF 增长 3.62。BiLSTM-CRF 模型通过 BiLSTM 层获取字的上下文信息。相比于 CRF 手工定义的窗口尺寸的语义跨度，BiLTM 覆盖整个目标语句，可以更完整地接收上下文特征。此外深度神经网络具有拟合非线性能力，单独 CRF 更趋向于对局部特征的线性加权拟合，而对训练数据的拟合能力，深度神经网络更优于 CRF。

表 2 不同模型实验结果

模型	精确率（%）	召回率（%）	F_1（%）
HMM	78.09	60.23	68.01
CRF	91.96	91.49	91.72
BiLSTM-CRF	95.47	95.21	95.34

5 结语

互联网上存在大量非结构化的中医医疗知识，自动化构建中医医疗图谱方法尚不成熟，存在着需要大量中医领域专家参与、耗时耗力且难以形成规模等问题，本文针对知识图谱自动化构建中的实体抽取模块，采用深度学习方法，在以字为基本输入元素的基础上提出基于 BiLSTM-CRF 的序列标注模型，实验结果显示本文提出的算法具有高度准确性。该算法将抽取的目标实体限定为中药材、方剂和病症 3 类实体。但是在中医体系中实体种类并非如此简明，拓充中医实体的种类是下一步重点工作方向。

参考文献

1 周昆. 基于规则的命名实体识别研究［D］. 合肥：合肥工业大学，2010.

2 Finkel J R, Grenager T, Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling［C］. Michigan: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 363-370.

3 阚琪. 基于条件随机场的命名实体识别及实体关系识别的研究与应用［D］. 北京：北京交通大学，2015.

4 冯元勇，孙乐，张大鲲，等. 基于小规模尾字特征的中文命名实体识别研究［J］. 电子学报，2008，36（9）：1833-1838.

5 钟志农，刘方驰，吴烨，等. 主动学习与自学习的中文命名实体识别［J］. 国防科技大学学报，2014，36（4）：82-88.

6 怀宝兴，宝腾飞，祝恒书，等. 一种基于概率主题模型的命名实体链接方法［J］. 软件学报，2014（9）：2076-2087.

7 Zhou G D, Su J. Named Entity Recognition Using An HMM-based Chunk Tagger［C］. California: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 473-480.

8 Brothwick. Maximum Entropy Approach to Named Entity Recognition［J］. PhD Dissertation, NewYorkUniversity, 1999, 12（1）: 18-25.

9 Lafferty J, McCallum A, Pereira F C N. Conditional Random Fields: probabilistic models for segmenting and labeling sequence data［C］. Williamstoun: Proceedings of 18th Internationl conference on Nachine Learning. ICML, 2001: 282-289.

10 Mccallum A, Li W. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons［C］. Sapporo: Conference on Natural Language Learning at Hlt-Naacl. Association for Computational Linguistics, 2003: 188-191.

11 Collobert, Ronan. Natural Language Processing (Almost)

from Scratch [J]. Journal of Machine Learning Research, 2011, 12 (1): 2493 - 2537.

12 Zheng Xiaoqing, Chen Hanyang, Xu Tiayu. Deep Learning for Chinese Word Segmentation and POS Tagging [C]. Seattle: EMNLP, 2013: 647 - 657.

13 Huang Z, Xu W, Yu K. Bidirectional LSTM - CRF Models for Sequence Tagging [EB/OL]. [2018 - 08 - 09]. https: //arxiv. org/abs/1508. 01991.

14 Ma X, Hovy E. End - to - end Sequence Labeling via Bi - directional Lstm - cnns - crf [EB/OL]. [2018 - 05 - 04]. https: //arxiv. org/abs/1603. 01354.

15 Nichols J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM - CNNs [EB/OL]. [2018 - 05 - 26]. https: //arxiv. org/abs/1511. 08308.

16 Rei M, Crichton G K O, Pyysalo S. Attending to Characters in Neural Sequence Labeling Models [EB/OL]. [2018 - 05 - 14]. https: //arxiv. org/abs/1611. 04361.

17 AkashBharadwaj, David Mortensen, Chris Dyer, et al. Phonologically Aware Neural Model for Named Entity Recognition in Low Resource Transfer Settings [C]. Austin: EMNLP, 2016: 1462 - 1472.

18 Yang Z, Salakhutdinov R, Cohen W W. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks [C]. Toulon: ICLR, 2017.

19 Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, et al. Semi - supervised Sequence Tagging with Bidirectional Language Models [C]. Vancouver: ACL, 2017.

20 Radim Řehůřek, Petr Sojka. Software Framework for Topic Modelling with Large Corpora [C]. Valletta: LREC, 2010: 45 - 50.

21 Wikipedia. 中文词条语料 [EB/OL]. [2018 - 07 - 20]. https: //dumps. wikimedia. org/zhwiki/.

2019 年《医学信息学杂志》编辑出版重点选题计划

2019 年本刊将继续以“学术性、前瞻性、实践性”为特色，及时追踪并深入报道国内外医学信息学领域前沿热点，反映学科研究动态，展示学科应用成果，引领学科发展方向。现对 2019 年度编辑出版重点选题策划如下：

一、医药卫生体制改革与医药卫生信息化

1“互联网 + 医疗健康”支撑体系、服务体系建设；2 医药卫生信息化发展规划与战略；3 信息化助力医疗服务、公共卫生服务、医疗保障体系建设的技术方案与典型案例；4 医疗卫生信息标准化与规范化建设现状和应用实践；5 医疗卫生信息化相关法律法规；6 智慧医院及智慧医疗服务模式建设目标、发展规划、解决方案。

二、医学信息技术

1 医疗人工智能及健康智能设备研究与应用；2 健康医疗大数据的管理及应用创新；3 家庭医生签约智能化平台建设及网上签约服务；4 精准医学与个性化医疗技术研究与应用；5 物联网、远程医疗服务与健康管理；6 医疗云平台功能、技术、系统架构及基础设施构建；7 基于互联网技术的医疗联合体建设与信息互通共享；8 网络安全体系建设与风险评估。

三、医学信息研究

1 医学信息学基础理论及方法研究；2 医学科技创新体系和发展战略；3 公民健康素养培养及健康促进；4 医学智库研究与智库服务；5 医药卫生数据分析、挖掘与知识发现技术。

四、医学信息组织与利用

1“互联网 +”环境下医学图书馆的创新举措；2 人工智能技术及其在医学图书馆中的应用；3 数字资源建设与学科服务模式演化与机制；4 区域医疗卫生信息资源整合。

五、医学信息教育

1“互联网 +”环境下医学信息专科、本科、研究生教育及继续教育面临的挑战、改革与实践创新；2 医学信息素养教育；3 网络化、数字化医疗健康教育培训平台及在线课程；4 基于互联网的健康科普知识精准教育；5 国外医学信息学教育的先进理念综述。

（《医学信息学杂志》编辑部）