

【引文格式】孙超, 谢晴宇. 中医病历术语识别方法探讨[J]. 中国中医药图书情报杂志, 2020, 44(2): 1-5.

• 信息技术与中医药 •

## 中医病历术语识别方法探讨

孙超<sup>1</sup>, 谢晴宇<sup>2\*</sup>

1. 首都医科大学中医药学院, 北京 100069; 2. 中国中医科学院中医临床基础医学研究所, 北京 100700

**摘要:** **目的** 探索中医领域利用少量标注语料进行电子病历中医学实体信息的命名实体识别(NER)研究工作, 为更复杂的中医电子病历信息处理及深度学习方法在中医领域内的运用提供参考。**方法** 分析中医电子病历词汇术语与一般的NER任务相比较的特殊性, 对比了目前3种NER技术的优缺点, 找寻适合中医电子病历医学术语的NER技术。**结果** 长短时记忆神经网络(LSTM)是一种无监督学习模型, 能有效利用序列数据中远距离依赖信息, 特别适合处理文本序列数据; 还可以和条件随机场(CRF)模型相结合, 解决中医NER的难点。长短时记忆神经网络联合条件随机场模型(LSTM-CRF)可以在未标记的病历文本语料上无监督学习词语特征, 不依赖于人工设计特征模板而达到自动提取患者症状、疾病、诱因等命名实体的目的。**结论** 中医电子病历术语识别应利用多种命名实体识别技术, 充分发挥这些技术的优势, 提高模型识别准确性。

**关键词:** 命名实体识别; 长短时记忆神经网络; 条件随机场; 中医电子病历

中图分类号: R241; TP391.1 文献标识码: A 文章编号: 2095-5707(2020)02-0001-05

DOI: 10.3969/j.issn.2095-5707.2020.02.001

开放科学(资源服务)标识码(OSID):



### Discussion on Methods of Terminology Recognition in TCM Medical Records

SUN Chao<sup>1</sup>, XIE Qing-yu<sup>2\*</sup>

(1. School of Traditional Chinese Medicine, Capital Medical University, Beijing 100069, China; 2. Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China)

**Abstract: Objective** To explore how to use the small amount of labeled corpora in the field of TCM to conduct research on named entity recognition (NER) of medical entity information in electronic medical records (EMR); To provide references for the application of more complex information processing of TCM EMR and in-depth learning methods in the field of TCM. **Methods** Specificity of vocabulary and terminology of TCM EMR compared to general NER tasks was analyzed, and the advantages and disadvantages of the current three NER technologies were compared, so as to find the named entity recognition technologies suitable for medical terminology of TCM EMR. **Results** As an unsupervised learning model, long and short-term memory (LSTM) neural network could effectively utilize long-distance dependent information in sequential data, especially suitable for processing text sequence data. It could also be combined with conditional random field model (CRF) to solve the difficulty of NER in TCM. LSTM-CRF model could learn word features in unsupervised condition in unmarked medical record text corpus, and could automatically extract named entities such as symptoms, diseases and causes of patients without relying on the artificial design of feature templates. **Conclusion** TCM EMR should be applied to multiple NER technologies, making full use of the advantages of these technologies

基金项目: 北京中医药“薪火传承3+3工程”崔锡章中医文化传承工作室

第一作者: 孙超, E-mail: sunchaotcm@ccmu.edu.cn

\*通讯作者: 谢晴宇, E-mail: xieqingyu@vip.126.com

and improving the accuracy of model recognition.

**Key words:** named entity recognition (NER); long and short-term memory (LSTM); conditional random fields; TCM electronic medical records (EMR)

随着医院信息化建设的发展,针对电子病历信息开展的后结构化研究已成为主流趋势。通过集成平台的后结构化策略,有效推动生产系统业务逻辑的改进,是真实世界平台建设领域的核心议题。目前,中医电子病历领域研究关键词有“数据挖掘”“远程医疗”“人工智能”等<sup>[1]</sup>,这些粗颗粒度的关键词显然没有将研究热点聚焦在更深入的解决方案上。电子病历中记载着患者的症状、体征、辨证和用药治疗的全部过程,利用这些记录,我们可以对医生的诊疗经验进行总结,为后人学习参考使用;同样,若将这些记录信息提供给计算机作为“学习”的文本素材,在充分“学习”大量事实性的数据之后,利用人工智能技术,理论上计算机就可以模拟人的思维进行诊断和治疗。计算机辅助诊疗能够帮助临床医生进行诊断,选择合适的治疗方式,进行风险预测,减少医疗错误的发生,最终达到协助临床决策的目的<sup>[2]</sup>。目前,已有多篇人工智能模拟医生进行诊断的研究报道<sup>[3-6]</sup>。实现计算机辅助诊疗的思路为:首先识别提取出病历中的症状、疾病、病机、诱因等具有特定含义的医学实体信息,再利用数据挖掘相关技术,发掘这些术语信息之间的联系。可见,医学实体信息的识别提取是实现计算机辅助诊疗的首要环节。

将电子病历信息作为文本语料,利用自然语言处理相关技术,从电子病历中进行医学实体信息的提取成为目前医学领域的重要任务之一<sup>[7]</sup>。术语识别工作是一项重要而关键的基础性步骤,它可以为中医药人工智能辅助临床决策服务,有很大的理论研究价值和应用研究价值。

### 1 中医病历命名实体识别研究的特殊性

利用计算机自动提取病历中的实体信息的难点在于:虽然医学术语的表述方式有一定的规范,但它还是一种自由化的文本表述,不同的医生在表达同一种意思时使用的中医术语往往会有不同的表达方式,对于这种情况,医生可以很容易判断出它们是否表达了同一意义,例如,医生可以很迅速地反应出纳差、不能食、食少、不知饥饿、饥不欲食、不思饮食、食欲不振等均表达“纳呆”之意,而计算机想判断出这一点却并不容易。在实现让计算机

理解的过程中,我们显然无法找到一本包含各种表述的词典,采用“字-字”匹配的模式来让计算机进行理解。此外,我们还希望计算机能够对识别出的中医术语进行分类,把属于症状的归属到症状术语里,属于病因的归属到病因术语里,属于方药的归属到方药术语里,以便进一步的挖掘分析。

实现病历文本语料术语识别的自然语言处理技术为命名实体识别(Named Entity Recognition, NER)技术,它最早由美国纽约大学学者 R Grishman 和 B Sundheim 于 1996 年在 MUC-6 (Message Understanding Conference 6)会议上提出,目的是从自然语言文本中识别出实体指称及其类别<sup>[8]</sup>。传统的 NER 任务包括识别人名、地名、组织机构名称等实体指称。尽管目前也有许多从文本中提取实体术语的模型,但是将这些模型应用于医学实体识别还是具有挑战性的,因为标准的自然语言处理工具不是为医学领域专门设计的,因此需要研究特定针对中医电子病历的 NER 办法<sup>[9]</sup>。

中医学领域与传统自然语言领域中识别人名、地名、组织机构名称等实体指称的不同点有 3 条。首先,传统的识别任务中,人名具有较固定的姓氏,地名、组织机构名称之后有固定的后缀用词;而中医命名实体往往没有一套严格的命名系统,有时表述还会带有古汉语的特点,如“纳可,寐佳”,命名实体特征性复杂,难以总结其中的规律性。其次,中医领域缺乏大规模、统一的标注语料集,这使得从大量标注好的语料中学习识别实体特征的监督学习算法实行起来人力时间成本较大,我们最好能寻求到半监督或无监督的学习算法。第三,中医命名实体长度不确定,实体内还会出现子实体或 2 个并列实体同时出现的嵌套情况,如“外感风寒”为四字术语,而“下元不足,元气升腾于上”则较长。实体名称越长,需要识别上下文信息范围就越广,识别难度越大。嵌套现象如“风热郁于胆络,兼脾有湿痰壅热”这里“风热”“胆络”“湿痰壅热”都是命名实体,而它们又共同组成了“风热郁于胆络,兼脾有湿痰壅热”这样的病机表述,嵌套现象的存在使得各类中医术语的识别工作是相互交织而非孤立的。这些特殊性决定了中医病历 NER 工

作要比一般领域的更加复杂多变,技术难度更大。

## 2 3种命名实体识别技术分析

NER技术大体可以归纳为:基于规则的方法(Rule-Based Model),基于统计模型的方法(Statistic-Based Model)和基于深度学习的方法(Deep Learning Method)。

### 2.1 基于规则的方法

基于规则的方法是在已有符号处理系统和规则下,由专家知识构造大量规则集,形成有限状态机,推理出可能的命名实体词组。规则表达易于理解,推理过程直观明了。但是,中医病历的语言缺乏一套严格的命名系统,有时候还会带有文言文的色彩,难以总结其中的规律性,单纯使用基于规则的方法难度较大,一般都将其与基于统计模型的方法联合使用。

### 2.2 基于统计模型的方法

基于统计模型的方法有隐马尔可夫模型(Hidden Markov Model, HMM)、条件随机场(Conditional Random Fields, CRF)模型、最大熵马尔可夫模型(Maximum Entropy Markov Model, MEMM)等。目前应用于中医病历术语识别工作最多的就是基于统计模型方法中的CRF模型。CRF能够在给定需要标注的观察序列条件下,计算整个标注序列的联合概率分布,并在整个观察值序列上求解一个最优的标注序列,具有很强的推理能力,能够使用复杂的、有重叠性的、非独立的特征进行训练,充分利用上下文信息,有效避免了HMM模型条件独立假设、MEMM模型标注偏置等诸多问题,识别效率也通常优于其他统计模型。

Wang Y Q等<sup>[10]</sup>对比分析了HMM、MEMM、CRF模型在中医临床记录中对症状实体的识别,结果发现CRF模型更适合中医临床病历的命名实体抽取。刘凯等<sup>[11]</sup>建立以词位、状态转移、上下文窗口、指示词、词典、构词模式为特征的CRF模型,对中医临床病历进行命名实体抽取。叶辉等<sup>[12]</sup>使用多特征CRF对《金匱要略》的症状、药物进行信息抽取研究,准确率达到84.5%,召回率70.9%, $F$ 值77.1%,有效抽取了中医古籍中所含的症状、药物文本实体信息。孟洪宇等<sup>[13]</sup>对《伤寒论》文本中的症状、病名、脉象、方剂等中医术语进行自动识别,采用CRF建立以字本身、词边界、词性、类别标签为特征组合的中医术语识别模型,模型准确率85.00%,召回率68.00%, $F$ 值75.56%。

但是,CRF模型非常依赖于特征工程,特征质

量的好坏直接影响到识别的准确率。特征选取需要基于大量的语言学知识、领域知识和专家经验,反复试验筛选调整。而中医语言表述抽象,加之缺乏统一标注的大规模标注语料集可供训练,使用CRF模型等监督学习模型人力时间成本投入较大。如何在少量人工干预前提下有效利用无标注语料,实现特征自动提取是我们应当关注的问题。

### 2.3 基于深度学习的方法

基于深度学习的方法是近年来逐渐占主流地位的方法,它通过多个神经元组成神经层,再由神经层逐层连接形成多层的神经网络结构<sup>[14]</sup>,模拟人的大脑思维过程进行分析学习。多隐层的结构使得每一层都能将原始输入进行线性或非线性的转换,从而放大其中与学习目标相关的部分,减小不相关的部分,数据规模更大,模型更复杂,刻画能力更强,识别效率更高。更值得注意的是,深度学习是一种无监督的学习,通过构建多隐层模型,自主抽取样本的特征,具有自动学习特征的能力,在一定程度上很好地替代一般的特征提取方法,减少了人工制定特征的工作量。自加拿大蒙特利尔大学学者Bengio Y将深度学习方法用于自然语言处理后<sup>[15]</sup>,越来越多的自然语言处理领域开始使用深度学习方法。“深度学习将会在自然语言理解领域产生巨大影响”<sup>[14]</sup>,可以预见,深度学习的下一个主战场就是自然语言处理领域。

深度学习模型的自主学习能力恰好可以解决CRF模型需要大量依赖人工制定特征工程的弊端,使得今后在抽取中医术语时,即使没有语言学专家的加入,也可以完成术语抽取工作。因此,应当对深度学习方法进行专门研究,以找寻适用于中医病历术语识别工作的深度学习模型。

## 3 深度学习模型在中医病历术语识别中的应用

### 3.1 中医病历术语识别属于NER序列标注问题

中医病历术语识别属于NER序列标注问题<sup>[16]</sup>。所谓序列标注,是指把输入句子文本看作由词语组成的序列 $X=(x_1, x_2, \dots, x_i, \dots, x_n)$ ,如 $X$ 为现病史文本中“发作时伴有反酸,嗝气,无呕吐”这一句话, $x_i$ 表示经过分词处理后的文本词语,即“发作/时/伴有/反酸/嗝气/无呕吐/”,序列标注就是给句子中每个词语打上标签集合中的某个标签 $Y=(y_1, y_2, \dots, y_i, \dots, y_n)$ 。使用BIEOS标记方法<sup>[17]</sup>,其中B为实体标记的开始,I为实体标记的其他部分,E为实体标记的结尾,O为不属于命名实体,

S 为单字即构成症状术语。例如，“发作时伴有反酸，嗝气，无呕吐”可被标识为“发/O 作/O 时/O 伴/O 有/O 反/B 酸/E，/O 嗝/B 气/E，/O 无/B 呕/I 吐/E。/O”。适用于序列标注问题的深度学习模型是递归神经网络（Recursive Neural Network, RNN），所谓“递归”是指它们的反馈回路结构，即在模型的隐层中加入了自连接和互连接，通过重现矩阵传播延迟信号，这样反馈回路就能把上一个时间标注的输出信息作为下一个时间的输入信息来处理，对前面的信息进行记忆并应用于当前的输出计算中，从而实现了对上下文信息的记录保存和利用。正因为 RNN 具有这样的特点和优势，使它特别适用于语音识别、机器翻译等需要根据上下文预测下一个单词、下一个语音的序列标注问题。

### 3.2 中医病历术语识别适用的深度学习模型为长短时记忆神经网络模型

RNN 在学习训练过程中需要将递归项反绕解开，它最大的弱点是需克服神经网络层数过多带来的参数训练时学习梯度消失的问题，RNN 在理论上虽然可以对任何长度的序列数据进行处理，但在实际应用中，特别是进行长程依赖的学习时，若某一项会受到很远处的标记影响，普通 RNN 表现往往不佳<sup>[18]</sup>。而中医病历中命名实体往往较长，需要识别上下文信息范围广，普通的 RNN 模型识别不佳。

为有效克服普通 RNN 梯度消失的问题，由德国慕尼黑大学学者 Hochreiter S 和 Schmidhuber J 提出<sup>[19]</sup>、后经改进的长短时记忆神经网络（Long Short-Term Memory, LSTM）结构<sup>[20-21]</sup>，可以看成是对 RNN 模型的改进。LSTM 包括 1 个用于保存信息的记忆单元（memory cell），3 组自适应的元素门进行控制更新，即控制网络输入的输入门（input gate），控制网络输出的输出门（output gate），控制记忆单元的忘记门（forget gate），共同组成记忆存储块（block）的结构，从而解决 RNN 梯度消失的问题。LSTM 既可以保存很久之前的信息，达到利用较远处的上下文信息的效果，有效克服梯度消失的问题；又可以避免无关紧要的内容进入记忆，通过训练学习达到对信息自动筛选的目的。且 LSTM 模型是一种数据驱动的方法，它不依赖特征工程，是一种端到端的训练过程，可以减少传统统计方法 CRF 模型需要大量制定特征模板的人工干预过程。近年来，LSTM 在自然语言处理领域发挥了重大作用。Lample G 等<sup>[22]</sup>将 LSTM 与 CRF 模型结

合，以词和字符为特征，加入 dropout 策略，进行 NER 标注。Ma X Z 等<sup>[23]</sup>利用双向 LSTM 合并卷积神经网络和 CRF 模型，得到 97.55% 的词性标注准确率和 91.21% 的 NER 准确率。

由于文本句子中词语和词语之间不是独立的，是有语义关系的，因此词语归属的标签也不是独立的，打标签时需要利用前面或后面的信息。当前的预测标签不仅与当前的输入词语有关，还与之前的预测标签相关，即预测标签序列之间是有强相互依赖关系的，有的命名实体标记之间互相是不能搭配的。若仅依靠 LSTM 得到某词属于某命名实体标记的概率，则可能预测出非法的标签序列。例如，使用 BIEOS 进行命名实体标注时，正确的标签序列中标签 O 后面是不会接标签 I 的。而此问题通过 CRF 模型可以得到解决，因为 CRF 模型的目标函数不仅考虑输入的状态特征函数，而且还包含了标签转移特征函数，可以在 LSTM 输出端将 softmax 函数分类器与 CRF 结合起来进行 NER 的标注<sup>[22,24]</sup>，使用 LSTM 解决提取序列特征的问题，使用 CRF 有效利用句子级别的标记信息，更好地进行 NER 工作。

张艺品等<sup>[25]</sup>以《备急千金要方》《千金翼方》《神农本草经》作为语料，应用 LSTM-CRF 模型，识别其中的病症、方剂、中草药等实体，准确率 95.47%，召回率 95.21%，F 值 95.34%，高于 HMM、CRF 模型。高甦等<sup>[26]</sup>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，准确率为 85.44%，召回率为 85.19%，F 值 85.32%。这些研究均证实了 LSTM 结合 CRF 技术适用中医文本的特点，模型泛化能力和鲁棒性更强。

## 4 小结

针对中医病历命名实体识别研究的特殊性，我们认为中医病历 NER 工作的解决流程为：首先，借助中医词典等规则知识对病历文本进行过滤；其次，对于中医词典无法识别的中医术语，使用 LSTM，利用其记忆存储块的结构，控制信息的存储和遗忘，从而实现对梯度信息选择性地读取和覆盖。LSTM 模型善于处理长范围的上下文信息问题，有效解决中医领域命名实体过长的难题；LSTM 模型作为深层非线性网络是一种无监督的学习过程，可以在原始字符集上提取特征，减少人工特征

制定的工作量,解决标注语料集匮乏的问题。此外,LSTM模型还可与CRF模型等线性方法相结合,解决中医病历文本数据量可能过小的问题,更好地利用NER标记上下文信息。LSTM-CRF可以在未标记的病历文本语料上无监督地学习词语特征,不依赖于人工设计特征模板,达到中医病历NER的目的。

#### 参考文献

- [1] 荣光,谢晴宇,孟庆刚. 中医电子病历研究领域科学知识图谱分析[J]. 中国中医药信息杂志, 2017, 24(1):99-104.
- [2] HE J, BAXTER S L, XU J, et al. The practical implementation of artificial intelligence technologies in medicine[J]. Nat Med, 2019, 25(1):30-36.
- [3] GULSHAN V, PENG L, CORAM M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs[J]. JAMA, 2016, 316(22):2402-2410.
- [4] KERMANY D S, GOLDBAUM M, CAI W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning[J]. Cell, 2018, 172:1122-1131.
- [5] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639):115-118.
- [6] CHENG J Z, NI D, CHOU Y H, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans[J]. Sci Rep, 2016(6):24454.
- [7] FORD E, CARROLL J A, SMITH H E, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review[J]. Journal of the American Medical Informatics Association, 2016, 23(5):1007-1015.
- [8] GRISHMAN R, SUNDHEIM B. Message Understanding Conference 6: A Brief History[C]// Proceedings of the 16th conference on Computational linguistics - Volume 1. Association for Computational Linguistics, 1996:466-471.
- [9] CHOWDHURY S, DONG X, QIAN L, et al. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records[J]. BMC Bioinformatics, 2018, 19(17):499.
- [10] WANG Y Q, YU Z H, CHEN L, et al. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study[J]. Journal of Biomedical Informatics, 2014, 47:91-104.
- [11] 刘凯,周雪忠,于剑,等. 基于条件随机场的中医临床病历命名实体抽取[J]. 计算机工程, 2014, 40(9):312-316.
- [12] 叶辉,姬东鸿. 基于多特征条件随机场的《金匮要略》症状药物信息抽取研究[J]. 中国中医药图书情报杂志, 2016, 40(5):14-17.
- [13] 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.
- [14] LECUN Y, BENGIO Y, HINTON G. Deep Learning[J]. Nature, 2015, 521(7553):436-444.
- [15] BENGIO Y, SCHWENK H, SENECALE J S, et al. Neural probabilistic language models[M]. Innovations in Machine Learning. Springer, 2006:137-186.
- [16] FINKEL J R, GRENAGER T, MANNING C. Incorporating non-local information into information extraction systems by gibbs sampling[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005:363-370.
- [17] 王蕾,谢云,周俊生,等. 基于神经网络的片段级中文命名实体识别[J]. 中文信息学报, 2018, 32(3):84-90, 100.
- [18] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2):157-166.
- [19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.
- [20] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: Continual prediction with LSTM[J]. Neural computation, 2000, 12(10):2451-2471.
- [21] GRAVES A. Supervised sequence labelling with recurrent neural networks[M]. Springer, 2012:37-45.
- [22] LAMPIE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[J]. Proceedings of NAACL-HLT, 2016:260-270.
- [23] MA X Z, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016:1064-1074.
- [24] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.0199(2015).
- [25] 张艺品,关贝,吕荫润,等. 深度学习基础上的中医实体抽取方法研究[J]. 医学信息学杂志, 2019, 40(2):58-63.
- [26] 高甦,金佩,张德政. 基于深度学习的中医典籍命名实体识别研究[J]. 情报工程, 2019, 5(1):113-123.

(收稿日期: 2019-10-15)

(修回日期: 2019-11-13; 编辑: 魏氏)