一种面向中医文本的实体关系深度学习联合抽取方法

杨延云 杜建强* 聂 斌 罗计根 贺 佳
(江西中医药大学计算机学院 江西 南昌 330004)

摘 要 目前实体识别和关系抽取任务大多采用流水线方式,但该方法存在错误累积、忽略两个任务相关性和信息冗余等诸多问题。结合中医文本的特点,提出一种基于深度学习的中医实体关系联合抽取方法。该方法使用改进的序列标注策略,将中医的实体关系联合抽取转换成序列标注任务,词向量与字符向量并联拼接作为双向 LSTM-CRF 输入,利用双向 LSTM 神经网络强大的特征提取能力,以及 CRF 在序列标注上的突出优势,结合优化的抽取规则完成中医实体关系联合抽取。在中医语料库上的实验结果表明,实体关系联合抽取的 F1 值可以达到 80.42%,与传统流水线方法以及其他方法相比,实验效果更佳。

关键词 实体关系联合抽取 深度学习 字词向量拼接 中医文本

中图分类号 TP391 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2023.03.033

A JOINT EXTRACTION METHOD OF ENTITIES AND RELATIONS FOR TRADITIONAL CHINESE MEDICINE TEXT BASED ON DEEP LEARNING

Yang Yanyun Du Jianqiang* Nie Bin Luo Jigen He Jia
(*School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, Jiangxi, China*)

Abstract At present, the entity recognition and relationship extraction tasks mostly use the pipeline method, which has the problems of error accumulation, ignoring the relevance of two tasks, and information redundancy. Combining the characteristics of traditional Chinese medicine(TCM) text, this paper proposes a joint extraction method of entities and relations for traditional Chinese medicine text. The improved sequence labeling strategy was used to convert the joint extraction of entities and relations of TCM into a sequence labeling task. The word vector and char vector parallel splicing was used as the input of BiLSTM-CRF. Using the strong feature extraction ability of BiLSTM neural network and the prominent advantages of CRF in sequence labeling, the joint extraction of entities and relations of TCM was completed by combining the optimized extraction rules. The experimental results show that the F1 value of the entities and relations can reach 80.42%. Compared with the pipeline method and other methods, it has better experimental effect.

Keywords Joint extraction of entities and relations Deep learning Char vector and word vector splicing Traditional Chinese medicine text

0 引 言

为推进国家中医药信息化的发展,各种中医药信息化平台的建设接踵而至,例如,中医辅助诊疗系统、中医智能问答系统、中医电子病历系统等。中医文献作为中医传承载体,记录了证型、方剂、中药、病因、病机和治则治法等数据,且存在着大量实体重叠的问题。而实体和关系抽取作为底层最基础的任务,能够快速地从半结构化、非结构化的中医文本中提取出实体以及它们之间的语义关系,对中医文献数据的有效利用和中医药的信息化研究具有促进作用和重要意义。

收稿日期: 2020-08-12。国家重点研发计划项目(2019YFC1712301);国家自然科学基金项目(61762051,61562045);江西省自然科学基金项目(20202BAB202019);江西省教育厅科技项目(GJJ190863);江西省研究生创新专项资金项目(YC2019-S358)。
杨延云,硕士生,主研领域: 自然语言处理,数据挖掘。**杜建强**,教授。**聂斌**,副教授。**罗计根**,硕士。**贺佳**,硕士。

1 相关研究

1.1 流水线方法研究

实体关系抽取作为信息抽取的重要子任务[1]，处理该任务的方法主要可以分为流水线方法和实体关系联合抽取方法两类。流水线方法即将实体关系抽取任务分为命名实体识别[2]（Named Entity Recognition，NER）和关系抽取[3]（Relation Extraction，RE）两个子任务，即给定一段半结构化或非结构化文本，首先通过命名实体识别提取出文本中的实体，然后对每个候选实体对进行关系分类。

典型的命名实体识别方法主要分为三类：基于规则的方法；基于统计学习的方法和基于深度学习的方法。其中，基于规则的方法大多是利用语言学知识，通过语言规则识别实体；基于统计学习的方法主要有隐马尔可夫模型（Hidden Markov Models，HMM）[4]、最大熵模型（Maximum Entropy Model，MEM）、支持向量机（Support Vector Machine，SVM）[5]和条件随机场[6]等，该方法依赖复杂的特征工程。近几年，循环神经网络（Recurrent Neural Networks，RNN）、长短期记忆网络（Long Short-Term Memory）等神经网络模型被应用于实体识别任务，并展现出强大的优势。

关系抽取方法主要可以分为经典的关系抽取方法和基于深度学习的抽取方法。经典的关系抽取方法主要包括有监督、半监督、弱监督和无监督 4 种[7]，这几种方法存在特征提取误差传播的问题，很大程度上影响最终关系抽取的结果。基于深度学习的方法避免了人工特征提取，Zeng 等[8]于 2014 年首次使用 CNN 进行关系分类。Vu 等[9]采用深度循环神经网络（Deep Recurrent Neural Networks，DRNN）进行关系抽取。

流水线方法虽然在模型选择和实验操作比较灵活、简单，但是这种方法存在以下几个问题：① 导致错误累积；② 忽略了两个子任务间的相关性；③ 产生大量冗余信息。例如文本“方剂麻杏石甘汤是由麻黄、杏仁等多味中药组成”，采用流水线方法的具体流程如图 1 所示。假如在命名实体识别阶段模型没有识别出实体“麻杏石甘汤”，由于关系抽取完全依赖实体识别的结果，则所有包含“麻杏石甘汤”的三元组皆无法得到，因此导致错误累积；已知文本中存在“方剂/中药”这一关系，可以推理第一个实体的类别是“方剂”类，第二个实体的类型是“中药”类，而采用流水线方法无法利用该信息进行推理；关系抽取是对每个候选实体对进行关系分类，不属于预定义关系的实体组合就是冗余信息，如（麻黄，None，杏仁）。

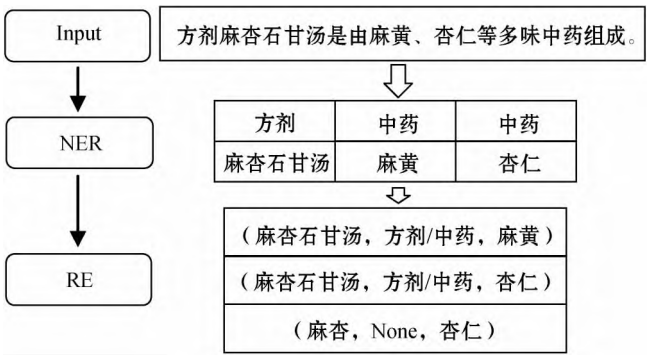

图 1 流水线方法流程

1.2 联合抽取方法研究

针对以上流水线方法存在的问题，实体关系联合抽取直接抽取给定文本中含有的实体和实体间语义关系的三元组（$Entity_1$，Relation，$Entity_2$），不仅能够充分考虑二者的相关性，将二者联合学习，还使两个子任务的性能得到了不同程度的提升。

Ren 等[10]提出 CoType 框架。Miwa 等[11]使用填表方法，将实体识别和关系抽取进行联合学习，但是都基于人工提取特征，依赖于复杂的特征工程，还需使用各种自然语言处理工具包。随着深度学习方法的兴起，Miwa 等[12]使用 BiLSTM 实现实体识别，通过共享输入层和 LSTM 编码层的参数，连用 Bi-TreeLSTM 结构实现关系抽取。Katiyar 等[13]针对 Miwa 等[12]利用依存树结构的缺点提出融合注意力机制的 RNN 方法实现实体关系联合抽取。Zheng 等[14]采用 BiLSTM 对输入层进行编码，选用 LSTM 进行解码，实现实体识别；通过共享 BiLSTM 编码器参数，利用 CNN 模块对编码层结果进行关系分类。文献[15]通过引入互反馈机制，反馈更新共享层的参数来提升联合抽取的效果。基于参数共享的实体和关系联合抽取方法增强了实体识别和关系抽取两个子任务的相关性，改善了传统流水线方法错误累积的不足。但是由于该方法都是利用共享底层模型参数来增强两者的相关性，实质上仍是先进行 NER，再利用 NER 的结果进行 RE，因此仍会产生不存在关系的实体对冗余信息，也存在错误传递。

Zheng 等[16]首次将实体关系联合抽取转化为序列标注问题，还设计了带有偏置损失函数的端到端模型，实现了真正意义上的实体关系联合抽取。但在最终三元组的抽取时采用就近距离策略，且规定一个实体只能存在一个三元组中，导致大量关系数据丢失，无法解决实体重叠问题。曹明宇等[17]借鉴 Zheng 等[16]的方法，改进标注策略，采用 BiLSTM-CRF 模型有效缓解了同一实体参与多个关系的重叠问题，在生物医学领域的药物实体关系数据集上取得了较好的效果。

鉴于传统流水线方法的不足和中医文本中存在大量实体重叠的问题,本文提出一种基于字词向量拼接的中医实体关系联合抽取方法。首先将字词向量拼接作为输入,再采用改进的序列标注策略在 BiLSTM-CRF(Bi-directional Long Short-Term Memory Conditional Random Fields,BiLSTM-CRF) 模型上对中医文本进行标注,最后通过自定义的抽取规则进行关系三元组提取。

2　中医实体及关系联合抽取方法

该方法使用改进的序列标注策略,将中医的实体关系联合抽取转换成序列标注任务,词向量与字符向量并联拼接作为双向 LSTM-CRF 输入,利用双向 LSTM 神经网络强大的特征提取能力,以及 CRF 在序列标注上的突出优势,结合优化的抽取规则完成中医实体关系联合抽取。整体方法流程如图 2 所示。

图 2　方法流程

该方法的整体流程为:

1) 对输入的文本句子利用 Word2vec 进行向量转化,分别生成字向量和词向量;

2) 将生成的向量以字为基本语义单元进行字词向量并联拼接;

3) 采用改进的标注策略,通过 BiLSTM-CRF 模型对每个句子进行序列标注;

4) 根据序列标注结果,结合自定义的抽取规则来抽取关系三元组。

2.1　模型输入

One-hot 编码得到的是稀疏向量,向量的维度完全取决于语料库的大小,且每个词的向量之间都是独立的,相近意思的词语也没有关联关系。相较于 One-hot 编码,Word2vec 得到的词向量降低了向量的维度,且语义相近的词语被映射在相近的位置。

本文训练向量所用语料来源于《中医证候鉴别诊断学》《中医 150 证候辨证论治辑要(何晓晖) 》和《中医药学概论》三本中医相关书籍。而采用分词工具得到的中文分词结果并非完全正确,且单独用词作为语义单元也忽略了词内字间的联系; 单独用字作为语义单元,又不能准确地表达当前的语境,因此本文采用了字词向量并联拼接作为模型输入,将字和词的信息有效地结合起来。中医语料利用 jieba 分词工具,并加载自定义的中医领域自定义词典进行分词,自定义词典主要包含大量的证型、方剂等信息,通过 Word2vec 训练得到 100 维词向量。中医语料使用 Word2vec 训练得到 100 维字向量。最终,本文采用以字向量为基本语义单元与该字所在词的词向量进行并联拼接得到 200 维字向量作为模型的输入,字词向量并联拼接丰富了词的语义信息,提取有效特征,如图 3 所示。例如文本: “四逆散中重用柴胡为君药”,则该句中作为模型输入“胡”的向量由“胡”的字向量与“胡”所在的词“柴胡”的词向量构成。

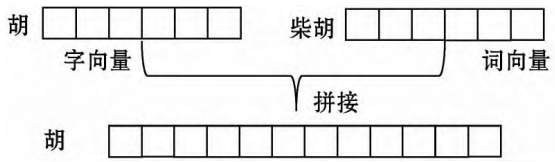

图 3　字词向量拼接

2.2　标注策略

本文在 Zheng 等[16] 提出的标注策略和曹明宇等[17] 的标注策略基础上进行改进,将实体关系联合抽取转化为序列标注的问题,对每个字符根据标注策略进行标注。如图 4 所示。

Input: 麻黄与桂枝相须为用以增发汗解表之力,如麻黄汤,用于风寒表实证。
Tags: B-方剂/中药-2　E-方剂/中药-2　O　B-方剂/中药-2　E-方剂/中药-2　OOOOOOOOOOOOOO　B-M-P　I-M-P　E-M-P　OOO　B-证型/方剂-1　I-证型/方剂-1　I-证型/方剂-1　I-证型/方剂-1　E-证型/方剂-1　O
Final results: (麻黄汤,方剂/中药,麻黄)　(麻黄汤,方剂/中药,桂枝)　(风寒表实证,证型/方剂,麻黄汤)

图 4　标注实例

其中“O”表示该字不与其他字构成实体,且在该句中与其他任何实体不存在预定义的关系; 此外,每个标签共包含三部分的内容: 该字在实体中的位置、关系类别、实体在三元组中的位置。该字在实体中的位置采用“BIES”策略进行表示, “B”代表实体开始, “I”代表实体中部, “E”代表实体末尾, “S”代表单个字构成实体; 关系类别是根据中医语料预先定义好的,本文共涉及 5 种关系,分别为方剂/中药、证型/方剂、证型/症状、病因/证型和 M,M 表示该实体与多个实体组成关系不同的三元组; 实体在三元组中的位置有 3 种: 1、2 和 P,其中 P 表示该实体与多个实体组成三元组且处于不同的位置。此种标注策略有效缓解了实体重叠的问题。

2.3　BiLSTM-CRF

本文采用 BiLSTM-CRF 模型进行中医文本的序列标注任务,具体模型结构如图 5 所示。

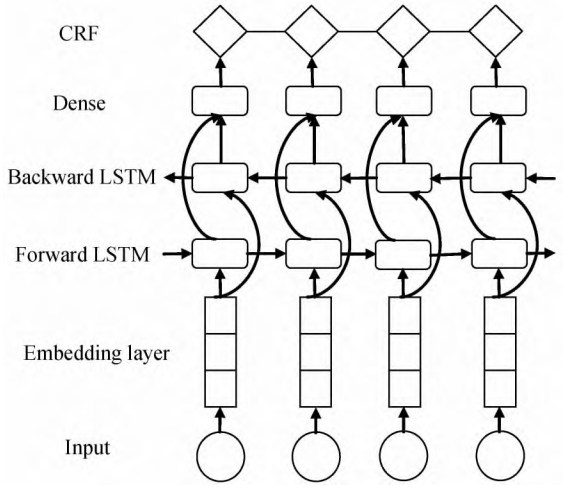

图 5 BiLSTM-CRF 模型结构图

LSTM 网络是 RNN 的一种变种[18]，引入了细胞状态概念，通过决定哪些信息需要被记忆，哪些需要被遗忘来解决 RNN 梯度爆炸和梯度消失的问题。LSTM 主要通过遗忘门、输入门和输出门来达到信息传递目的。具体计算公式如下:

$$\boldsymbol{f}_t = \text{sigmoid}(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \quad (1)$$

$$\boldsymbol{i}_t = \text{sigmoid}(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \quad (2)$$

$$\tilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_C[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_C) \quad (3)$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t * \boldsymbol{C}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{C}}_t \quad (4)$$

$$\boldsymbol{o}_t = \text{sigmoid}(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \quad (5)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{C}_t) \quad (6)$$

式中: $\boldsymbol{f}_t$表示遗忘门的值; $\boldsymbol{x}_t$表示当前时刻的输入词; $\tilde{\boldsymbol{C}}_t$表示临时细胞状态; $\boldsymbol{C}_t$表示细胞状态; i_t表示记忆门的值; $\boldsymbol{o}_t$表示输出门的值; $\boldsymbol{h}_t$表示隐藏状态; $\boldsymbol{W}$ 表示权值矩阵; $\boldsymbol{b}$ 表示偏置矩阵。

BiLSTM 由前向的 LSTM 与后向的 LSTM 结合而成，得到一个前向 t 时刻的隐藏层输出 $\overrightarrow{\boldsymbol{h}}_t$和一个后向 t 时刻的隐藏层输出 $\overleftarrow{\boldsymbol{h}}_t$拼接而成 $[\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t]$，充分了利用上下文信息。

BiLSTM 使用 softmax 进行归一化处理得到每个字对应每个标签的概率，然而每个标签并非独立存在，它们之间存在一定的约束，例如“E-方剂/中药-1”之前一定是“I-方剂/中药-1”，“B-方剂/中药-1”之后一定是“I-方剂/中药-1”。而 CRF 可以更好地学习各标签之间的依赖关系，进行全局优化，使标注处理更加准确和高效。

2.4 抽取规则

Zheng 等[16]默认一个实体只存在一个三元组中，关系抽取采取就近距离原则，这样便损失了大量实体关系信息，而中医文本中存在大量一个实体与多个实体构成关系三元组的情况。曹明宇等[17]在此基础上进行改进，取得了较好的效果，但在匹配最近实体时设置了匹配方向而导致一些三元组丢失。

依据上述分析以及中医文本的信息抽取需要，本文在采用就近原则抽取的基础上，自定义了以下 3 条抽取规则:

规则 1: 对于命名实体识别任务，当实体标签的三个部分信息均正确时进行抽取; 对于联合抽取任务，当组成三元组的实体 1、实体 2 和关系类别均正确时进行抽取。

规则 2: 组成三元组的关系类别约束。关系类别相同，或者其中一个或者两个实体的关系类别为 *M*，即本文预定义的 4 种关系类型可以与其相同的关系类型匹配也可以与 *M* 匹配。

规则 3: 组成三元组的实体位置约束: 实体位置分别为 1 和 2，或者其中一个或者两个实体的实体位置为 P，即 1 可以与 2 匹配，也可以与 P 匹配，2 和 P 同理。如图 2 样例所示，麻黄汤可与麻黄组成关系三元组(麻黄汤，方剂/中药，麻黄)，与桂枝组成关系三元组(麻黄汤，方剂/中药，桂枝)，与风寒表实证组成关系三元组(风寒表实证，证型/方剂，麻黄汤)。

3 实验及结果分析

3.1 实验数据集

本文使用的语料源于中医古籍、中医相关教材等整理的 2 968 个句子，均经人工按照本文的标注策略进行标注。该中医语料共包含方剂、中药、证型、症状和病因 5 类实体，方剂/中药、证型/方剂、证型/症状、病因/证型和 M 共 5 类关系，其中 M 表示该实体与多个实体组成关系不同的三元组。具体的占比见表 1，按照 7∶3的比例划分训练集和测试集。

表 1 各关系类型语料占比(%)

关系类型	占比
M	4.64
病因/证型	8.48
方剂/中药	31.42
证型/方剂	12.48
证型/症状	42.98

3.2 评价指标

实验采用的评价指标是准确率(Precision，P)、召回率(Recall，R)、F1 值。对于命名实体识别任务，当实体标签的三个部分信息均正确时认为其正确; 对于联合抽取任务，当组成三元组的实体 1、实体 2 和关系类别均正确时认为其正确。具体三元组的 P、R、F1 的

计算公式为:

$$P = \frac{n_{\text{predictright}}}{n_{\text{predict}}} \tag{7}$$

$$R = \frac{n_{\text{predictright}}}{n_{\text{right}}} \tag{8}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

式中: $n_{\text{predictright}}$ 表示预测得到且正确三元组的数目; n_{predict} 表示预测得到三元组的数目; n_{right} 表示实际三元组的数目。

3.3　实验设置

向量输入由 Word2vec 训练得到 100 维字向量和 100 维词向量拼接而成 200 维字向量。模型训练涉及的主要超参数: 学习率设置为 0.001; dropout 设置为 0.5; 优化器(optimizer) 设置为 Adam 等。

3.3.1　向量输入对比实验

为了验证字词向量拼接作为模型输入的有效性,通过多组不同维度的单独字向量作为输入和字词向量拼接作为输入对比实验进行验证,实验结果见表 2。

表 2　输入对比实验

向量输入类型	向量维度	模型	命名实体识别/%			三元组抽取/%		
			P	R	F1	P	R	F1
char	100	BiLSTM-CRF	85.00	86.79	85.88	83.85	72.93	78.01
char	200	BiLSTM-CRF	84.59	87.02	85.79	85.56	71.61	77.96
char	300	BiLSTM-CRF	84.14	84.48	84.31	82.97	67.21	74.26
char	400	BiLSTM-CRF	82.17	85.38	83.75	80.35	66.18	72.58
char	500	BiLSTM-CRF	82.09	84.31	83.19	83.69	65.97	73.78
char	600	BiLSTM-CRF	81.83	85.38	83.57	80.98	66.01	72.74
char_concat_word	50 + 50 = 100	BiLSTM-CRF	85.71	88.90	87.28	84.78	74.42	79.26
char_concat_word	100 + 100 = 200	BiLSTM-CRF	86.08	88.33	87.19	86.83	74.89	**80.42**
char_concat_word	150 + 150 = 300	BiLSTM-CRF	86.41	86.35	86.38	86.93	71.26	78.32
char_concat_word	200 + 200 = 400	BiLSTM-CRF	86.50	88.26	87.37	87.20	73.31	79.66
char_concat_word	250 + 250 = 500	BiLSTM-CRF	87.84	87.42	**87.63**	88.36	71.95	79.31
char_concat_word	300 + 300 = 600	BiLSTM-CRF	85.90	88.66	87.26	87.93	72.84	79.68

其中 char 表示字向量,char_concat_word 表示字词向量拼接。由表 2 可知,字词向量拼接作为输入的效果均优于单独字向量作为输入。本文最终目的是提取关系三元组,因此选用字向量 100 维和词向量 100 维并联拼接作为模型输入。

如表 2 所示,与实体识别相比,三元组抽取具有更高的精确率,但其召回结果低于实体识别任务,这意味着存在预测的实体并不能构成实体对,只找到了 Entity_1 而没有找到相应的 Entity_2,或者 Entity_2 而没有找到相应的 Entity_1。因此,实体对具有比单个实体更高的精度率和更低的召回率。

3.3.2　模型对比实验

将本文方法与两种流水线方法进行对比实验,方法一: BiLSTM-CRF 序列标注用于实体识别,在实体识别结果的基础上使用 SVM 进行关系抽取。方法二: BiLSTM-CRF 序列标注用于实体识别,在实体识别结果的基础上利用 LSTM 进行关系抽取。这两种方法所用语料均为中医语料,序列标注时采用“实体中字的位置-实体类别”的标注策略。

由表 3 实验结果可以得出,本文采用的联合抽取方法较传统的流水线方法 F1 值有较大的提升,较方法二(BiLSTM-CRF + LSTM) F1 值提升 4.49%,较方法一(BiLSTM-CRF + SVM) F1 值提升接近 10%,说明了本文方法的有效性。

表 3　模型对比实验(%)

方法类型	模型	P	R	F1
流水线方法	BiLSTM-CRF + SVM	69.18	71.79	70.46
流水线方法	BiLSTM-CRF + LSTM	76.63	75.24	75.93
本文方法	BiLSTM-CRF	86.83	74.89	**80.42**

3.3.3　抽取规则对比实验

采用本文提出的标注策略,字词向量拼接作为输入,通过 BiLSTM-CRF 模型进行序列标注,分别采用 Zheng 等[16]、曹明宇等[17]和本文的抽取规则进行三元组抽取对比实验,如表 4 所示。

表 4 抽取规则对比实验(%)

抽取规则	P	R	F1
Zheng 等[16]	89.51	28.78	43.55
曹明宇等[17]	86.99	67.98	76.32
本文方法	86.83	74.89	**80.42**

根据表 4 可知,使用本文的抽取规则实验效果整体更佳。前两种方法 P 值偏高的原因如下: Zheng 等[16] 默认一个实体只存在一个三元组,且在三元组抽取时采用就近原则; 曹明宇等[17] 在标注策略中增加了实体类别的信息,且在三元组抽取时规定实体位置 1 只能向后匹配,实体位置 2 只能向前匹配。为了进一步对比这 3 种方法的抽取结果,表 5 举例进行说明。

表 5 抽取规则对比

原句	例 1: 麻黄与桂枝相须为用以增发汗解表之力,如麻黄汤,用于风寒表实证	例 2: 逍遥散和真人养脏汤两方中都含有白术、白芍
正确三元组	(麻黄汤,方剂/中药,麻黄)(麻黄汤,方剂/中药,桂枝)(风寒表实证,证型/方剂,麻黄汤)	(逍遥散,方剂/中药,白术)(逍遥散,方剂/中药,白芍)(真人养脏汤,方剂/中药,白术)(真人养脏汤,方剂/中药,白芍)
Zheng 等[16]	(麻黄汤,方剂/中药,麻黄)	(逍遥散,方剂/中药,白术)(真人养脏汤,方剂/中药,白芍)
曹明宇等[17]	(麻黄汤,方剂/中药,桂枝)(风寒表实证,证型/方剂,麻黄汤)	(逍遥散,方剂/中药,白术)(真人养脏汤,方剂/中药,白术)(真人养脏汤,方剂/中药,白芍)
本文方法	(麻黄汤,方剂/中药,麻黄)(麻黄汤,方剂/中药,桂枝)(风寒表实证,证型/方剂,麻黄汤)	(逍遥散,方剂/中药,白术)(真人养脏汤,方剂/中药,白术)(真人养脏汤,方剂/中药,白芍)

由表 5 可知: 本文方法可以抽取到 Zheng 等[16] 和曹明宇等[17] 丢失的部分信息,改善了实体重叠的问题,但还是存在关系三元组损失的现象,仍需进一步改进。

4 结 语

本文使用改进的序列标注策略,将中医的实体关系联合抽取转换成序列标注任务,词向量与字符向量并联拼接作为 BiLSTM-CRF 输入,利用 BiLSTM 神经网络强大的特征提取能力,以及 CRF 在序列标注上的突出优势,结合优化的抽取规则完成中医实体关系联合抽取,不仅克服了传统流水线方法的弊端,很大程度地缓解了实体重叠的问题,并在中医语料上达到 80.42% 的 F1 值。

但是,本文的方法仍存在丢失三元组的现象。此外该方法依赖人工标注语料,而现实中存在大量无标签数据,可以借助远程监督的方法来缓解该问题。探究每个字符在句中的位置信息是否对实体关系联合抽取有促进作用是未来的工作。

参 考 文 献

[1] Golshan P N, Dashti H A R, Azizi S, et al. A study of recent contributions on information extraction [EB]. arXiv: 1803.05667, 2018.

[2] Grishman R, Sundheim B M. Message understanding conference-6: A brief history [C]//16th International Conference on Computational Linguistics, 1996.

[3] Kumar S. A survey of deep learning methods for relation extraction [EB]. arXiv: 1705.03645, 2017.

[4] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 473 - 480.

[5] Ma Y. Support Vector Machine(SVM) [M]. Berlin: Springer, 2008.

[6] Mccallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and weben-hanced lexicons [C]//Conference on Natural Language Learning at HLT-NAACL. ACM, 2003: 188 - 191.

[7] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述 [J]. 软件学报, 2019, 30(6): 1793 - 1818.

[8] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network [C]//2014 25th International Conference on Computational Linguistics, 2014.

[9] Vu N T, Adel H, Gupta P, et al. Combining recurrent and convolutional neural networks for relation classification [EB]. arXiv: 1605.07333, 2016.

[10] Ren X, Wu Z, He W, et al. CoType: Joint extraction of typed entities and relations with knowledge bases [C]//Proceedings of the 26th International Conference on World Wide Web. ACM, 2017: 1015 - 1024.

[11] Miwa M, Sasaki Y. Modeling joint entity and relation extraction with table representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), 2014: 1858 - 1869.

[12] Miwa M, Bansal M. End-to-end relation extraction using lstms on sequences and tree structures [EB]. arXiv: 1601.00770, 2016.

(下转第 234 页)

(b)

图 10 部分文字定位结果

4 结 语

本文采用改进的 ASHE 和 SWT 算法进行自然场景中低对比度图像文字定位。利用改进的自适应子直方图均衡算法提升图像对比度; 然后提取 MSER 候选区域,并结合启发式规则滤除大部分非文本区域; 最后应用改进 SWT 算法配合笔画宽度特征实现最终定位。实验结果表明,本文算法在性能上优于其他算法,准确率和综合性能均有所提高,证实了本文方法的可行性与有效性。

参 考 文 献

[1] 陈硕,郑建彬,詹恩奇,等. 基于笔画角度变换和宽度特征的自然场景文本检测[J]. 计算机应用研究,2019,36(4):1270 - 1274.

[2] Chen X, Yuille A L. Detecting and reading text in natural scenes[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2004: 366 - 373.

[3] 万燕,王晓华,卢达. 自然场景下中文文本定位关键技术的研究[J]. 计算机应用与软件,2018,35(7):243 - 249.

[4] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2010: 2963 - 2970.

[5] Yin X C, Yin X, Huang K, et al. Robust text detection in natural scene images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,36(5):970 - 983.

[6] 潘立,刘亮亮,张再跃. 自然场景中文本定位方法研究[J]. 计算机与数字工程,2019,47(6):1459 - 1465.

[7] 李东勤,徐勇,周万怀. 自然场景图像中的文本检测及定位算法研究——基于边缘信息与笔画特征[J]. 重庆科技学院学报(自然科学版),2019,21(3):81 - 83.

[8] 司飞. 自然场景图片中的文本检测和定位[J]. 电子技术与软件工程,2020(2):147 - 149.

[9] 丁畅,董丽丽,许文海. “直方图”均衡化图像增强技术研究综述[J]. 计算机工程与应用,2017,53(23):12 - 17.

[10] 陈子妍. 低对比度图像的清晰化与增强[J]. 科技视界,2019(14):74 - 75.

[11] 董丽丽,丁畅,许文海. 基于直方图均衡化图像增强的两种改进方法[J]. 电子学报,2018,46(10):2367 - 2375.

[12] 张国和,黄凯,张斌,等. 最大稳定极值区域与笔画宽度变换的自然场景文本提取方法[J]. 西安交通大学学报,2017,51(1):135 - 140.

[13] He Y, Dai S. Detecting the corners of light shape of automotive low-beam headlamp by inverse Hough transform[C]//International Conference on Industrial and Information Systems,2009: 309 - 311.

[14] 李鑫明,李俊芳,李大华,等. 基于笔画宽度特征和剪枝算法的自然场景标签检测[J]. 激光杂志,2020,41(1):65 - 70.

[15] 卢未来. 面向图像的场景文字识别技术研究[D]. 锦州:辽宁工业大学,2018.

[16] 刘功琴. 低对比度图像文字区域定位及文字识别算法的研究与实现[D]. 昆明: 云南大学,2018.

[17] 陈硕. 基于图像增强的多特征自然场景文本检测研究[D]. 武汉: 武汉理工大学,2018.

[18] Buta M, Neumann L, Matas J. FASText: Efficient unconstrained scene text detector[C]//IEEE International Conference on Computer Vision,2015: 1206 - 1214.

(上接第 222 页)

[13] Katiyar A, Cardie C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,2017: 917 - 928.

[14] Zheng S, Hao Y, Lu D, et al. Joint entity and relation extraction based on a hybrid neural network[J]. Neurocomputing, 2017, 257: 59 - 66.

[15] 马建红,李振振,朱怀忠,等. 反馈机制的实体及关系联合抽取方法[J]. 计算机科学,2019,46(12):242 - 249.

[16] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme[EB]. arXiv: 1706.05075, 2017.

[17] 曹明宇,杨志豪,罗凌,等. 基于神经网络的药物实体与关系联合抽取[J]. 计算机研究与发展,2019,56(7):1432 - 1440.

[18] Mikolov T. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111 - 3119.

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