·药事信息·

命名实体识别在中药名词和方剂名词识别中的应用

龚德山，梁文昱，张冰珠，马星光*（北京中医药大学，北京 100029）

摘要　目的：利用命名实体识别（Named Entity Recognition）技术识别文本中出现的中药名词和方剂名词，并比较两种命名实体识别方法在识别中药名词和方剂名词时的表现。**方法：**方法一为利用现有的分词工具（如“结巴”中文分词工具等）对文本进行分词，之后使用分词后的结果进行中药名词和方剂名词的匹配。方法二为搭建并训练用于中药名词和方剂名词识别的双向长短期记忆（Bidirectional Long Short Term Memory，BLSTM）神经网络模型。首先，采用两种可行的方法实现命名实体识别。其次，比较这两种方法的表现。**结果：**现有分词工具对中药名词和方剂名词的分词不准确，因此，会导致接下来的匹配阶段出现错误。而通过 BLSTM 神经网络模型进行命名实体识别，不但可以避免分词错误，而且在实验中表现出较强的歧义处理能力。**结论：**在应用命名实体识别技术于识别中药名词和方剂名词时，相比使用分词工具先分词后识别，通过训练神经网络模型对中药名词和方剂名词直接识别的方法更合适。

关键词： 自然语言处理；命名实体识别；BLSTM 神经网络；中文分词

中图分类号：TP314　**文献标识码：**A　**文章编号：**1002-7777(2019)06-0710-07
doi:10.16153/j.1002-7777.2019.06.016

Application of Named Entity Recognition in the Recognition of Words for Chinese Traditional Medicines and Chinese Medicine Formulae

Gong Deshan, Liang Wenyu, Zhang Bingzhu, Ma Xingguang* (Beijing University of Chinese Medicine, Beijing 100029, China)

Abstract　Objective: To identify words of Chinese traditional medicines, and Chinese medicine formulae by using Named Entity Recognition (NER) and compare the performance of two NER methods. **Methods:** The first method was to use the off-the-shelf programming modules, like “Jieba” Chinese word segmentation module, to segment sentences into words, and then to recognize the target keywords through word-matching. The second method was to build and train a neural network model-- Bidirectional Long Short-Term Memory (BLSTM) specially for recognizing the words of the Chinese traditional medicines, and the Chinese medicine formulae. The two possible methods were used to implement NER. Then, the performance of these two methods was compared. **Results:** The current off-the-shelf programming modules for Chinese word segmentation were unable to segment the words of the Chinese traditional medicines, and the Chinese medicine formulae accurately, which led to inaccurate word matching accordingly. By contrast, the trained BLSTM not only avoided the possibility of

基金项目：中央高校基本科研业务费专项资金（编号 2018-JYB-XSCXCY47）
作者简介：龚德山，研究生；E-mail：deshan.gong@bath.edu
通信作者：马星光，硕士，副教授，硕士生导师，研究方向：机器学习与人工智能；E-mail：himxg@126.com

inaccurate word segmentation, but also surprisingly exhibited better capability in dealing with the ambiguity of words. **Conclusion:** When NER was applied to identifying the words, it is more suitable to recognize the words of Chinese traditional medicines and Chinese medicine formulae directly by training neural network model than to segment words before recognition by the off-the-shelf programming models.

Keywords: natural language processing; Named Entity Recognition; BLSTM neural network; Chinese word segmentation

当今，信息数字化的飞速进展使信息获得的方式变得更加容易。但是，随着电子化文本数量的日益增加，通过人工从海量的文本数据中提取关键信息的工作也逐渐变得繁重。信息抽取技术可以减轻人力筛选数据的工作量，该技术是通过一系列自然语言处理方法从自然语言文本中抽取关键信息，并将文本中非结构化的数据转换为结构化的数据。命名实体识别通常作为信息抽取中的第一步需要执行的处理，用来识别并分类在文本中所出现的指定名词[1]。如今，虽然命名实体识别技术已经比较成熟，但是该技术通常被用来识别人名、组织、地点等较为常见的专有名词，如果将该技术应用于识别中药名词和方剂名词，那么这将推动自然语言处理技术在处理中医药领域文本的发展，推动中医药领域内的研究。

电子数据数量的迅猛增长也同时带动了神经网络的复兴[2]，神经网络已经在自然语言处理中得到了深入的研究和广泛的应用。其中，双向长短期记忆（Bidirectional Long Short-Term Memory，BLSTM）神经网络是长短期记忆（Long Short Term Memory，LSTM）神经网络的一种演变[3]，在中文分词和命名实体识别等方面的表现尤为突出[4-6]。

本次研究的目的为通过两种方法，实现中药名词和方剂名词识别的目的。这两种方法为1）使用分词工具对文本分词后，进行中药名词和方剂名词的匹配；2）训练能够直接识别中药名词和方剂名词的BLSTM神经网络模型。之后，通过实验对两种方法的表现进行比较，以确定其中较好的方法。最终，本次研究的结果为通过训练BLSTM神经网络模型而实现的中药名词和方剂名词命名实体识别表现较好。

1 研究方法

中文与拉丁语系的语言不同，中文不显式地使用空白作为词之间分界符[7]。在应用自然语言处理中文时，中文分词往往是必不可少的预处理步骤[8]。进行中文命名实体识别时，既可以对文本进行分词预处理后再进行识别，也可以不预先进行分词直接在原始文本中进行识别。本次研究比较了这两种方法，具体如下所述：

1）先分词，后识别。先对原始文本进行分词处理，再在分词结果中搜索中药和方剂名词。

2）不进行分词，直接识别。不对原始文本做分词处理，而直接搜索并识别原始文本中出现的中药名词和方剂名词。

而在应用命名实体识别处理中文时，是否应当进行分词预处理已经被深入研究。较为一致的观点为不进行分词预处理而直接进行命名实体识别的效果较好[9]。在本次研究中，这两种方法将被专门用于识别中药名词和方剂名词。

目前，有较多可以直接使用的开源第三方分词器。与之相比，虽然目前也有可以进行命名实体识别的代码库，例如Stanford NLP[10]、Tencent AI中的自然语言处理SDK和Baidu AI中的自然语言处理SDK。但是，它们都无法专门识别中药名词和方剂名词。所以，上面列举的两种方法中，第一种方法较为简单。通过直接调用可以使用分词器，减少编写代码的工作量。如果分词结果准确，那么之后中药名词和方剂名词的识别将比直接在原始文本中进行识别简单。如果分词结果不准确，那么错误的分词将导致中药名词和方剂名词的匹配也出现错误。此时，采用第二种方法“不分词直接识别”并通过实验检测其表现。具体而言，通过搭建并训练用于命名实体识别的BLSTM神经网络模型直接识别出文本中中药名词和方剂名词，而不预先进行分词处理。

2 研究实验

为比较在研究方法中所介绍的两种方法的表现，本节详细描述本次研究中用于比较这两种方法而进行的实验及其实验结果，以此验证这两种方法在应用于中药名词和方剂名词标注时的实际表现。

实验中所使用的编程语言为Python（版本 3.6.8），使用3.6.x版本是为了兼容建立神经网络模型时所需的TensorFlow。所使用的计算机的操作系统为Microsoft Window 10（x64）家庭版。计算机的处理器为Intel Core i7-8750H，内存大小为16 GB，显卡为NVIDIA GeForce GTX 1060。

2.1 先分词，后识别

在本次实验中，被用来比较的第三方中文分词器为“结巴”中文分词工具（版本 0.39）[11]、清华大学THU Lexical Analyzer for Chinese（THULAC）中文分词工具（版本 0.1.1）[12]、pkuseg中文分词工具（版本 0.0.15）[13]。这三款中文分词器在分词领域被广泛使用，本次研究将比较它们在切分中药名词和方剂名词的表现。通常，用来比较分词表现的指标为精确率（Precision）、召回率（Recall）和F_1值[14]。因为本次研究仅测试分词工具在切分中药名词和方剂名词的表现，精确率、召回率和F_1值的计算方式与传统比较分词工具的计算方式不同，具体的计算公式如下所示：

$$\text{精确率}=\frac{\text{正确切分的中药名词和方剂名词数量}}{\text{正确切分的中药名词和方剂名词数量+错误切分的中药名词和方剂名词数量}}$$

$$\text{召回率}=\frac{\text{正确切分的中药名词和方剂名词数量}}{\text{正确切分的中药名词和方剂名词数量+没有被识别的中药名词和方剂名词数量}}$$

$$F_1=\frac{2\times\text{准确率}\times\text{召回率}}{\text{准确率=召回率}}$$

其中，F_1值是一种特殊的F度量（F-measure），而F度量是用于融合精确率和召回率的度量指标。在精确率计算公式中，“错误切分的中药名词和方剂名词”被定义为1）分词器在中药名词和方剂名词中添加分词标记。例如，中药“炙甘草”被错误分词为“炙”和“甘草”。2）本不应当作中药名词或方剂名词的词被切分为中药名词或方剂名词。比如，“柽柳注射液”被错误地分词为“柽柳”和“注射液”。另外，当包含若干中药名词的方剂名词被错误切分为若干中药名词时，只认为该方剂名词被错误切分，所以错误切分的数量只记为1个。例如，当“麻黄附子细辛汤”被切分为“麻黄”“附子”“细辛”和“汤”时，仅计算为1次错误切分。在召回率计算公式中，“没有被识别的中药名词和方剂名词”被定义为分词器没有在中药名词或方剂名词的周围添加正确的分词标志。例如，“取川椒”本应该被分词为“取”和“川椒”。

用于测试的文本节选自第七版《中药学》教材，其中包含411个中药名词，161种中药名词；201个方剂名词，79种方剂名词。所测试的三个分词器表现如表1、表2所示。

表1 切分中药的精确率、召回率和F_1值

分词工具	精确率	召回率	F_1值
“结巴”	79.62%	92.14%	0.8542
THULAC	84.91%	80.78%	0.8279
pkuseg	80.82%	82.00%	0.8140

表 2 切分方剂的精确率、召回率和 F_1 值

分词工具	精确率	召回率	F_1 值
“结巴”	100.00%	79.55%	0.8542
THULAC	100.00%	12.32%	0.2193
pkuseg	100.00%	17.16%	0.2929

在表1中，虽然“结巴”分词工具切分中药名词的精确率不是最高的，但是其F_1值较高。因此，切分的精确率和召回率同样重要时，“结巴”分词切分的切分表现相对较好。表2中的数据显示，“结巴”中文分词切分方剂名词的表现同样优于另外两个分词工具。因此可以认为：“结巴”中文

自动分词工具切分中药名词和方剂名词的表现最好。但是，其错误分词的数量仍然不能忽略不计。另外，“结巴”中自动分词工具允许添加自定义词典，以矫正部分分词错误。但是，自定义词典中用于决定分词优先级的词频参数较难被合理地设置。比如，“加大黄”的合理分词结果应该为“加”和“大黄”。在不使用自定义词典或自定义词典中的“大黄”词频较小时，“加大黄”会被切分为“加大”和“黄”，因为“加大”也是中文中的常见词。但是，如果为“大黄”设置较大的词频参数，“加大黄油的用量”又会被不合理地切分为“加/大黄/油的用量”（斜线代表分词标志）。而手动为自定义词典中所有中药词条和方剂词条设置合适的词频是无法实现的。所以通过本次实验可以得出结论，由于现有第三方分词器对中药名词和方剂名词的分词表现并不好。所以先利用分词器进行分词，并在之后识别中药名词和方剂名词的方法较难实现本次研究的目的。虽然通过重新训练能够专门处理包含中药名词和方剂名词的分词工具，有可能使“先分词后识别”的方法成功。但是，训练完整的分词模型与本次研究的目的相偏离。而且，采用“先分词后识别”的方法是因为其实现的简易性。所以可以得出结论，利用“先分词，后识别”的方法并不适合于实现本次研究的目的。在下一小节中，方法二，“不分词，直接识别”的可行性将被验证。

2.2 不分词，直接识别

本节将验证使用BLSTM神经网络模型识别中药名词和方剂名词的可行性。通常，训练神经网络模型包含准备训练数据和设置神经网络的结构和超参数两个主要阶段。本节将先描述准备训练数据和字嵌入字典的方法。之后，描述所搭建BLSTM神经网络模型及其超参数的设置。最后通过实验验证此方法的可行性。

2.2.1 准备语料库和字嵌入字典

语料库是指大量的机器可以使用的文本或语音信息。本次研究中所收集语料的来源为第七版《中药学》教材。在处理作为语料素材的原始文本时，以所有非文字符号（包括任意标点符号、特殊符号和任意长度的空格）作为标记，将语料素材切分为只包含文字的字符串，并仅保留切分后长度小于40个字符的字符串。因为语料素材的分句结果中，长度超过40个字符的字符串极少。之后，筛除不包含中药名词和方剂名词的字符串。最后，使用基础的“IOB”标注[15]的方式标注字符串中的每个字：“B”代表被标注字为中药名词或方剂名词的起始字，“I”代表被标注为中药名词或方剂名词中的中间字，“O”代表不是中药名词或方剂名词的字。例如，“在此方剂中麻黄是中药的组成药”的“IOB”序列为“OOOOOBIOOOOOOO”。最后所收集的语料库中共包含2206条包含中药名词或方剂名词的中文字符串，及其对应的“IOB”序列。另外，文字在被输入到神经网络模型之前需要通过字嵌入转化为字向量。本次研究中，使用one-hot嵌入的方式将汉字转换为字向量。使用one-hot嵌入是因为不需要额外训练字嵌入模型。使用one-hot字嵌入需要预先收集字典以将字典中的汉字编号。本次研究所使用的字典中收录了通用汉字规范表（2013）中所有的汉字，其中包含了现代汉语中较为常见的汉字，所收集的字嵌入字典中共包含8103个汉字。为了处理可能出现的未识别字，在字典末尾额外添加“<UNK>”代表未识别字。

2.2.2 搭建BLSTM神经网络及超参数设置

本研究中所使用的神经网络框架为建立在GPU版本Tensorflow（版本1.12.0）之上的Keras（版本2.2.4）。最终，本次研究中构建的BLSTM模型的结构如图1所示。该BLSTM神经网络由两层隐藏层组成：BLSTM层和SoftMax层。其中BLSTM层由两个不同方向的LSTM层构成，图中红色的LSTM层和绿色的LSTM层分别代表从左到右和从右到左两个方向的LSTM层，它们共同组成BLSTM层。被输入的字符串经one-hot嵌入转换为一串字向量后被输入进该神经网络的BLSTM层。之后，BLSTM层的输出被输入到激活函数为SoftMax的隐藏层（同时也是该神经网络的输出层）。最终，SoftMax层输出该神经网络所预测的“IOB”标签序列。

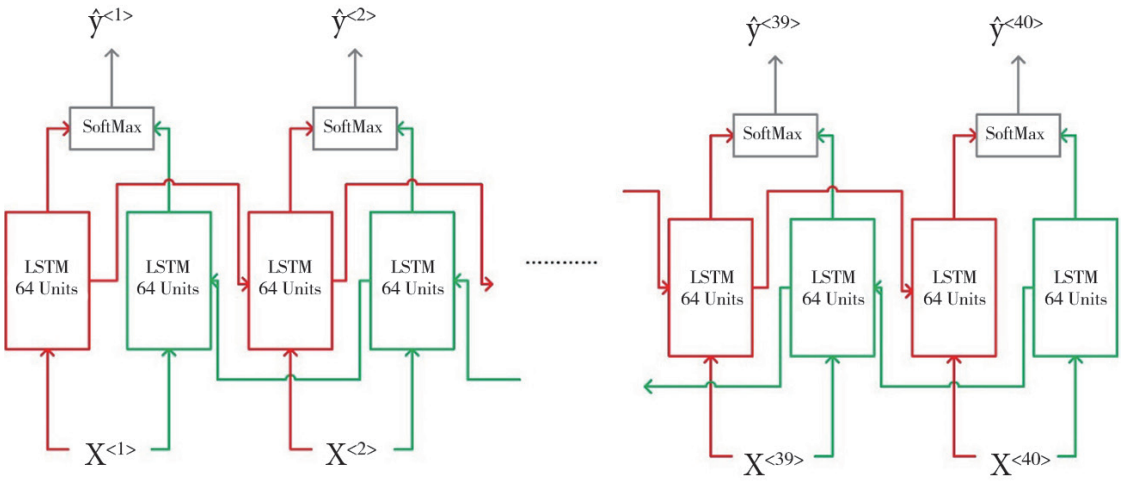

图1 BLSTM神经网络的结构

图1中$X^{<t>}$（{t|1≤t≤40，t∈Z}）代表用于输入到神经网络的字向量。由于语料库中最长的汉字字符串包含40个汉字，所以BLSTM神经网络的长度也设置为40。因为字嵌入字典中共包含8104个条目，所以每一个输入到BLSTM神经网络的字是大小为8104×1的字向量。因此，被输入进BLSTM神经网络的每个字符串可以表示为一个大小为8104×40的矩阵，其中每一列代表该字符串中的一个汉字。当字符串包含的汉字数量少于40个时，不足的部分用空白字填充。每一个空白字用所有元素都为0的字向量代表。之后，$X^{<t>}$被输入到神经网络的BLSTM层。BLSTM层在Keras中建立方式为使用双向的CuDNNLSTM层。在Keras中，CuDNNLSTM层与LSTM层的功能相同。不过CuDNNLSTM层能够利用GPU更快地完成模型的训练。此外，组成BLSTM的两个LSTM层均采用tanh激活函数（Activation Function）。之后，BLSTM层的输出被传入激活函数为SoftMax的隐藏层中。由于神经网络对每一个输入的汉字标签的预测属于多分类预测，所以该隐藏层的激活函数选择SoftMax较为合理，并且损失函数（Loss Function）也应该选择“categorical_crossentropy”。与输入进神经网络的字向量相似，因为每个汉字可以被标注为“I”“O”或“B”，所以每个汉字的标注用大小为3×1向量代表。例如，“I”标签可以用向量$[1,0,0]^T$表示。所以在输出层中，隐藏单元（Hidden Units）的数量应为3。所以图1中$\hat{y}^{<t>}$是大小为3×1的“IOB”向量代表BLSTM神经网络所预测输入字的字向量$X^{<t>}$的标签。本质上，用于命名实体识别的BLSTM神经网络是使用大小为8104×40汉字字符串矩阵，生成该字符串对应的大小为3×40“IOB”标签矩阵。另外，因为语料库中共包含2206条字符串，所以整个语料库中的数据可以被一个大小为2206×8104×40的张量（Tensor）代表。与之对应，语料库中所有的“IOB”标签序列可以被一个大小为2206×3×40的矩阵代表。另外，为了在实验中调整和评估BLSTM神经网络，语料库中2206条字符串中，15%的语料作为开发集，15%的语料作为测试集，剩余的语料作为训练集。为了防止过拟合，两个隐藏层中采用了L2正则化。为了加快模型学习速度，该神经网络模型采用了Adam优化器。该模型的超参数如表3所示。

表3 训练BLSTM神经网络的超参数设置

超参数	数值
Learning rate	0.01
Beta_1	0.9
Beta_2	0.99
Epochs	20
Batch Size	64
Learning rate decay	0.01
L2 regularization（λ）	0.001
Number of hidden units in BLSTM	64

在表3中没有列出的超参数，保留为所使用的Keras方法的默认参数值。进行20次迭代训练后，训练结果：训练集的准确率为95.41%，损失函数的值为0.085；测试集的准确率为92.25%，损失函数的值为0.137。训练集的准确率和测试集的准确率的差别较小，所以可以认为所训练模型不存在过拟合的现象，因此L2正则化参数不需要调整。

在实验最后阶段，使用所训练的BLSTM神经网络模型标注了部分包含中药名词或方剂名词的句子，标注结果如表4所示。通过观察表4中的例子可以发现，所训练的BLSTM神经网络模型在处理歧义时表现较好。但是仍然可能出标注错误，其原因可能为语料库中的训练数据过少。

表4 BLSTM对5个中文字符串中的中药名词和方剂名词的标注结果

包含中药名词的中文字符串	BLSTM神经网络标注结果
当归是中药	BIOOO
大黄是中药	BIOOO
当归国华侨回到故乡时	OOOOOOOOOO
对门家的狗叫大黄	OOOOOOOO
麻黄附子细辛汤的药理作用	BBIIIIOOOOOO

3 讨论

以上两个实验证明了“结巴”中文分词工具、THULAC中文分词包和pkuseg中文分词包在切分中药名词和方剂名词时表现较差。因此，使用它们难以实现以“先分词，后识别”的方式识别中药名词和方剂名词的目的，而重新训练专门能够切分中药名词和方剂名词的分词器的工作量较大。所以，本次研究中的“先分词，后识别”的方法不可行。相对而言，通过训练BLSTM神经网络以采用“不分词，直接识别”的方法在本次研究中表现较好。在实验中，尽管用于训练模型的语料较少且模型结构较为简单，但是所训练的BLSTM神经网络模型已具有一定的中药名词和方剂名词标注能力。尽管如此，本次研究所训练模型仍有如下不足。

（1）语料库中训练样本不平衡：实验中所收集语料全部来自中医教材，因此，其中所出现的中药名词和方剂名词往往仅作为中医领域的专业名词使用。这导致语料库中仅包含大量的正样本，而缺少负样本。

（2）语料库规模过小：本次研究中所收集的语料仅有2206条，有很多中药词条和方剂词条没有在语料库中出现过。这导致所训练模型无法正确识别没有在语料库中出现的中药名词和方剂名词。

（3）无法区分中药名词和方剂名词：由于所收集的语料库中的方剂名词数量较少，所以语料库中的中药名词和方剂名词采用了同样的标注方式。如果方剂名词的数量足够，方剂名词可以采取区别于中药名词标注的标签。例如，用“B_herb”和“I_herb”标注中药名词，而用“B_pres”和“I_pres”标注方剂名词。

4 结论和结语

本次研究发现利用现有的分词工具进行“先分词，后识别”的方法识别文中的中药名词和方剂名词准确率不高。相比而言，通过训练BLSTM神经网络模型，以“不分词，直接识别”的方法，识别中药名词和方剂名词具有更高的可行性。

本次研究实现并比较了两种可行的命名实体识别方法，并应用在中药名词和方剂名词的识别中。实验结果说明了下一步研究较为合理的方向应为建立并训练能够直接识别中药名词和方剂名词的神经网络。

参考文献：

[1] Daniel Jurafsky. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition[M]. 2nd

Edition. New Jersey：Pearson Education，2009：759.

[2] Lecun Y，Bengio Y，Hinton G. Deep Learning[J]. Nature，2015，521（7553）：436.

[3] Hochreiter S，Schmidhuber J. Long Short-term Memory[J]. Neural Computation，1997，9（8）：1735-1780.

[4] Yao Y，Huang Z. Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation[C]//International Conference on Neural Information Processing. Cham，Switzerland：Springer International Publishing，2016：345-353.

[5] Ma J，Ganchev K，Weiss D. State-of-the-art Chinese Word Segmentation with Bi-LSTMs[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels，Belgium：Association for Computational Linguistics，2018：4902-4908.

[6] Ouyang L，Tian Y，Tang H，et al. Chinese Named Entity Recognition Based on B-LSTM Neural Network with Additional Features[C]//International Conference on Security，Privacy and Anonymity in Computation，Communication and Storage. Cham，Switzerland：Springer International Publishing，2017：269-279.

[7] Xue N，Shen L. Chinese Word Segmentation as LMR Tagging[C]//Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume 17. Sapporo，Japan：Association for Computational Linguistics，2003：176-179.

[8] Chen X，Qiu X，Zhu C，et al. Long Short-term Memory Neural Networks for Chinese Word Segmentation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon，Portugal：Association for Computational Linguistics，2015：1197-1206.

[9] Zhang Y，Yang J. Chinese NER Using Lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers）. Melbourne，Australia：Association for Computational Linguistics，2018：1554-1564.

[10] Manning C，Surdeanu M，Bauer J，et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics：System Demonstrations. Baltimore，Maryland USA：Association for Computational Linguistics，2014：55-60.

[11] 结巴中文分词[EB/OL].（2017-08-31）[2019-04-02]. https://github.com/fxsjy/jieba.

[12] Sun M，Chen X，Zhang K，et al. Thulac：An Efficient Lexical Analyzer for Chinese [EB/OL].（2017-01-17）[2019-04-02]. https://github.com/thunlp/THULAC-Python.

[13] Sun X，Wang H，Li W. Fast Online Training with Frequency-adaptive Learning Rates for Chinese Word Segmentation and New Word Detection[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics：Long Papers-Volume 1. Jeju Island，Korea：Association for Computational Linguistics，2012：253-262.

[14] Sproat R，Emerson T. The First International Chinese Word Segmentation Bakeoff[C]// Association for Computational Linguistics. Sighan Workshop on Chinese Language Processing. Stroudsburg，PA，USA：Association for Computational Linguistics，2003.

[15] Ramshaw L A，Marcus M P. Text Chunking Using Transformation-based Learning[M]//Armstrong S，Church K，Isabelle P，et al. Natural Language Processing Using Very Large Corpora. Dordrecht：Springer Netherlands，1999：157-176.

（收稿日期 2019年3月4日 编辑 王雅雯）