doi:10.3772/j.issn.2095-915x.2018.01.008

基于 BLSTM 的科技文献术语抽取方法

赵东玥[1] 杜永萍[1] 石崇德[2]

1. 北京工业大学信息学部 北京 100124；
2. 中国科学技术信息研究所 北京 100038

摘 要 术语抽取是研究科技文献领域的重要技术，为进一步提高科技文献术语抽取的准确率和召回率，本文采用了基于 BLSTM（Bidirectional Long Short-Term Memory）的神经网络模型。使用预先训练的词向量字典将中文分词结果映射为向量作为 BLSTM 模型的输入，使用序列标注的方法将输出分类结果映射为术语的边界进行术语抽取。在自动化技术、计算机技术领域的数据集上，设计实验对比了使用词为特征的 BLSTM 模型和条件随机场模型的抽取结果。结果表明基于 BLSTM 模型的科技文献术语抽取得了更优的性能，在中文数据集上精确率最高 0.7821，召回率最高 0.8020，F1 值最高 0.7860，在英文数据集上分别达到 0.8525，0.8677 和 0.8543。

关键词： 术语抽取；科技文献；长短时记忆

中图分类号： TP391，G35

开放科学（资源服务）标识码（OSID）

Scientific Literature Terms Extraction Based on Bidirectional Long Short-Term Memory Model

ZHAO Dongyue[1] DU Yongping[1] SHI Chongde[2]

1. Faculty of Information Science, Beijing University of Technology, Beijing 100124, China;
2. Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract Term extraction plays an important role in the field of scientific literature. In order to improve the accuracy and recall of the term extraction, this research designed a neural network model based on BLSTM (Bidirectional Long Short-Term Memory) model. The segmentation results in Chinese were mapped into the vectors via pre-trained word vector dictionary, and the output of classification results were

项目基金： 国家自然科学基金青年基金项目“面向科技监测的实体识别与关系抽取研究”（71403257）。

作者简介： 赵东玥（1992–），硕士研究生，研究方向：自然语言处理，Email：zdy@emails.bjut.edu.cn；杜永萍（1977–），博士，副教授，研究方向：信息检索，自然语言处理；石崇德（1979–），博士，副研究员，研究方向：自然语言处理、机器翻译。

doi:10.3772/j.issn.2095-915x.2018.01.008

mapped as the term boundaries via the sequence tagging. The experiment was implemented to compare the BLSTM model with word feature and the conditional random field method in the fields of automation technology and computer technology. The results presented that the BLSTM model obtained the better performance with the highest accuracy 0.7821, the highest recall 0.8020 and the highest F1 value 0.7860 in Chinese dataset. For the English dataset, the highest accuracy, recall and F1 value is 0.8525, 0.8677 and 0.8543, respectively.

Keywords：Term extraction; scientific literature; LSTM

1 引言

根据GB/T10112-1999《术语工作原则与方法》中给出的概念，术语是专业领域中概念的语言指称，通常为名词。科技文献中的术语能够表示文献的研究领域和研究方向。科技文献术语抽取在自然语言处理领域有重要研究意义，在机器翻译和跨语言信息检索等领域中有广泛用途。

术语抽取的方法主要分为三类，基于规则的方法，基于统计的方法和基于机器学习的方法。基于规则的方法根据分析语料制定术语抽取规则[1]。基于统计的方法使用词频和词长度等信息进行术语抽取，对特定结构文本的抽取方法[2,3]，对特定类型术语的抽取方法[4]，使用互信息和位置加权的方法[5]，使用信息熵和词频的DV-entropy[6]，使用C-value[7]的方法和使用SCP[8]的方法，以及两者相结合的方法[9]，以及改进的IC-value[10]，ATValue[11]等。基于传统机器学习的方法包括支持向量机（Support Vector Machine，SVM），隐马尔可夫模型（Hidden Markov Model，HMM），条件随机场（Conditional Random Fields，CRF）[12]等使用序列标注的方法对词进行分类，以及支持向量回归（Support Vector Regression）的术语抽取方法[13]。CRF在HMM的基础上解决了标签偏置问题，提出之后被应用到了不同专业领域的术语抽取研究[14,15]中，并取得了较好的效果。基于神经网络的方法有结合词向量模型和术语部件统计的抽取方法[16]，深度神经网络的方法[17]。随着神经网络的发展，长短时记忆（Long Short-Term Memory，LSTM）[18]的提出解决了循环神经网络（Recurrent Neural Network，RNN）不能保存长期状态的问题，在此基础上产生了双向长短时记忆（Bidirectional Long Short-Term Memory，BLSTM）[19]，使用长短时记忆模型解决序列标注问题成为了一种更有效的解决方法[20-22]。

本文主要就科技文献领域的术语抽取任务展开研究，利用术语词典在无标注的原始语料上构建训练语料，使用基于双向长短时记忆的神经网络BLSTM模型进行术语抽取并比较BLSTM，LSTM，CRF方法下使用不同特征进行科技术语抽取的效果。本文第2节分别阐述了基于CRF，LSTM，BLSTM的术语抽取模型；第3节阐述了基于BLSTM模型的术语抽取方法；第4节通过实验比较了传统方法与神经网络

方法的抽取效果；第5节总结全文并提出之后的研究方向。

2 术语抽取模型

2.1 条件随机场CRF

条件随机场是Lafferty在2001年[12]提出的一种判别式图模型，可以通过观察状态预测隐含变量，在本文中使用线性链条件随机场（Linear-Chain CRF），模型描述如下：

X，Y是随机向量，$\Lambda=\{\lambda_k\}\in R^k$是参数向量，$\{f(y,y',\boldsymbol{x}_t)\}_{k=1}^{K}$是一个实数值的特征函数集合，线性链条件随机场的概率分布符合如下形式：

$$p(\boldsymbol{y}\mid\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x})}\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},\boldsymbol{x}_t)\right\}$$

（公式1）

其中X是观察序列，Y是状态序列，$\boldsymbol{x}$和$\boldsymbol{y}$分别是X和Y的取值，t是状态时刻，第k个特征函数f_k通常为二值函数，$Z(\boldsymbol{x})$是归一化函数。

$$Z(\boldsymbol{x})=\sum_{y}\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},\boldsymbol{x}_t)\right\}$$

（公式2）

通过对输入序列和对应的状态序列进行训练，CRF模型预测新的输入序列对应的状态序列。

2.2 长短时记忆模型LSTM

长短时记忆（Long Short-Term Memory，LSTM）是1997年由Sepp Hochreiter[18]提出的一种神经网络，在循环神经网络的基础上增加了单元状态（cell status）来保存长期状态，LSTM允许神经网络可以选择保存更早期的历史信息并选择何时遗忘这些信息。LSTM的结构示意图如图1。

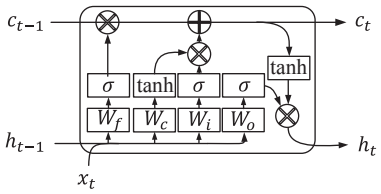

图1 LSTM单元结构

LSTM单元使用三种门限结构有选择的控制信息的流动，保护和控制单元的状态。门限结构分别是输入门，遗忘门和输出门，表示如下：

$$\boldsymbol{f}_t=\sigma(W_f[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f)\quad（公式3）$$

$$\boldsymbol{i}_t=\sigma(W_i[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i)\quad（公式4）$$

$$\tilde{\boldsymbol{c}}_t=\tanh(W_c[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c)\quad（公式5）$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t\quad（公式6）$$

$$\boldsymbol{o}_t=\sigma(W_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o)\quad（公式7）$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t)\quad（公式8）$$

公式中⊙表示向量按对应元素(elementwise)相乘，结果为一个向量。$[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]$表示将向量$\boldsymbol{h}_{t-1}$和向量$\boldsymbol{x}_t$进行拼接。$\sigma$是sigmoid激活函数$\sigma(z)=1/(1+\exp(-z))$。其中$\boldsymbol{f}_t$，$\boldsymbol{i}_t$，$\boldsymbol{o}_t$，分别为遗忘门，输入门，输出门在t时刻的输出，$\tilde{\boldsymbol{c}}_t$来源是t–1时刻单元状态输出和t时刻序列输入。$\boldsymbol{c}_t$为t时刻单元状态的向量。$W_f$，$W_i$，$W_c$，$W_o$分别是遗忘门，输入门，单元状态，输出门的权重，随迭代进行更新。$\boldsymbol{h}_t$,$\boldsymbol{h}_{t-1}$分别为t和t–1时刻LSTM隐藏层的输出向量。

2.3 双向长短时记忆模型BLSTM

双向长短时记忆（Bidirectional Long Short-Term Memory，BLSTM）在LSTM的基础上，结合了输入序列在前向和后向两个方向上的信息。对于t时刻的输出，前向LSTM层具有输

doi:10.3772/j.issn.2095-915x.2018.01.008

入序列中t时刻以及之前时刻的信息，而后向LSTM层中具有输入序列中t时刻以及之后时刻的信息。前向LSTM层t时刻的输出记作$\vec{h}_t$，后向LSTM层t时刻的输出结果记作$\overleftarrow{h}_t$，两个LSTM层输出的向量可以使用相加、平均值或连接等方式进行处理，本文选用连接的方法。BLSTM的示意图如图2。

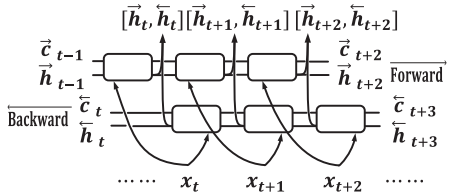

图 2 BLSTM结构

3 基于BLSTM模型的术语抽取

本文使用的模型结构如图3，输入序列的分词结果经过预先训练好的词典转化为词向量传入输入层，隐藏层使用BLSTM，输出层使用Softmax[23]。为了尽可能少的使用外部知识，神经网络模型选择使用词本身作为输入的特征。

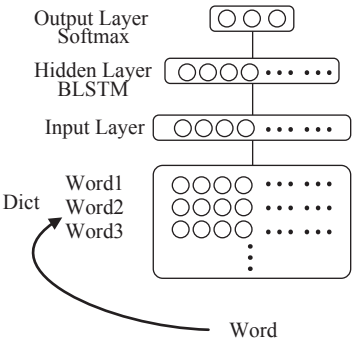

图 3 基于BLSTM的神经网络结构

神经网络的训练过程利用反向传播，使用损失函数更新门的权重，隐藏层的输出结果通过输出层Softmax函数输出分类向量，每个分量的值就是术语标注不同分类结果的概率，概率和为1。

对于大规模无监督数据，使用Word Embeding[24]将词本身转化为词向量表示，本文选择使用Word2Vec语言模型[25]。Word2Vec有两种计算模型，分别为Skip-Gram和CBOW，模型结构如图4。

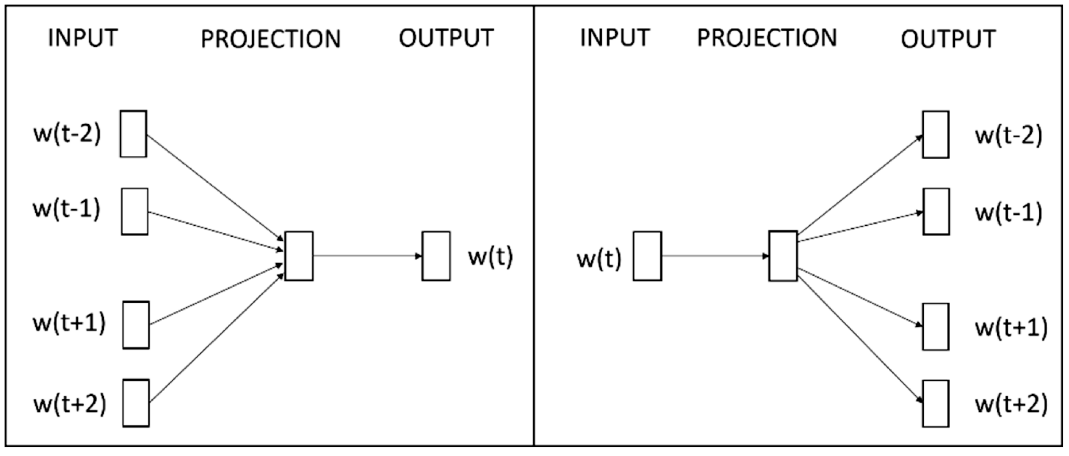

图4 CBOW模型（左）和Skip-Gram模型（右）

CBOW模型使用围绕目标的词来预测目标词。Skip-Gram模型使用目标词来预测周围的词。

CBOW模型表示如下：

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T}\log p(w_t \mid w_{t-n}, L, w_{t-1}, w_{t+1}, L, w_{t+n}) \quad (公式9)$$

Skip-Gram模型表示如下：

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T}\sum_{(-n\le j\le n),(j\ne 0)} \log p(w_{t+j} \mid w_t) \quad \text{（公式 10）}$$

其中W_i为对应位置的词。本文选择使用CBOW模型进行词向量的训练。

算法流程如表1所示：

表1 基于BLSTM的术语抽取算法

输入：

- 术语句子的分词结果数组input(词1，词2，词3，……，词n)
- 术语句子Word2Vec词向量训练的查找字典Dict{词：向量，词：向量，……，词：向量}
- 训练完毕的神经网络模型

输出：术语句子中的术语词

1. 使用查找字典Dict将input数组映射成向量数组，对未达到神经网络词数量上限m的向量数组填充-1向量，得到输入向量input_vector（向量1，向量2，向量3……向量n，向量n+1（-1，-1，-1……），……，向量m（-1，-1，-1……））
2. 将Input_vector输入神经网络，使用遮罩层处理不定词数量构成的输入向量
3. 将向量输入BLSTM层，拼接前后向的LSTM向量
4. 将隐藏层BLSTM的向量通过输出层映射为标签类别对应的3维向量，使用Softmax获得标签B（术语的首词），I（组成术语的词），O（非术语词）的概率
5. 生成单个标签B的术语词或以标签B开头，存在连续标签I的术语词

本文在TP领域数据上验证模型的有效性，对句子进行分词，在分词结果上使用术语词典正向最大匹配为语料添加标注信息。标注信息使用BIO标注方法，其中B代表术语的首词，I代表组成术语的词，O代表非术语词。表2为中文示例：

表 2 中文语料示例

编号	句子
1	JXTA技术主要用于提供P2P应用系统所需的基础服务
2	在总结传统B/S架构的远程学习系统许多弊端的基础上. 分析JXTA技术的体系结构
3	核心协议及开发P2P应用系统的优势

中文语料使用HanLp开源工具①内置的标准分词器对句子进行分词和词性标注，使用词典进行序列标注，以序号为1的句子为例，标注结果如下：

JXTA/nx/B 技术/n/I 主要/b/O 用于/v/O 提供/v/O P/nx/B 2/m/I P/nx/I 应用/vn/I系统/n/I 所需/nz/O 的/ude1/O 基础/n/B 服务/vn/I ./w/B

标注结果中第一项是词本身，第二项是词性，第三项是术语边界标记，标注项之间使用“/”号分隔，词之间使用空格分隔，如：JXTA/nx/B中，JXTA（词），nx（词性），B（术语边界）。样例中可以观察到的术语词为：JXTA技术、P2P应用系统、基础服务。

英文语料使用Stanford NLP[20]自然语言处理工具对句子进行词性标注，使用词典进行序列标注。

① http://hanlp.linrunsoft.com

doi:10.3772/j.issn.2095-915x.2018.01.008

4 实验

4.1 实验设置

文本实验语料使用自动化技术、计算机技术领域下（中图分类号TP）的中英文句子对齐语料②。每个语料有各自对应领域的术语词典。语料的数据分布如表3所示。

表3 语料数据分布

类别	句子	总字符数	句子平均字符数	词典包含术语词数量
TP领域中文	30000	1486909	49.5	664273
TP领域英文	30000	774773	25.8	465020

为验证本文术语抽取方法的效果，将BLSTM方法、LSTM方法与CRF方法进行对比。将30000句数据按照编号顺序以6000句为一组共5组，设置训练集和测试集为4：1做交叉检验。其中第1组实验取1-6000句作为测试集，第2组实验取6001-12000句作为测试集，依此类推。

实验对于输入的不同特征进行组合和比较，CRF的方法选择使用词、词+词性特征进行术语抽取，窗口大小为5。LSTM和BLSTM方法设置输入层为200维的词向量，隐藏层150维，输出层3维，batchsize设为64。抽取结果使用准确率（Precision，P），召回率（Recall，R）以及F1值作进行评价。

$$P=\frac{\text{模型正确抽取的术语词数量}}{\text{模型抽取的术语词数量}}\times 100\% \quad \text{（公式11）}$$

$$R=\frac{\text{模型正确抽取的术语词数量}}{\text{测试集中的术语词数量}}\times 100\% \quad \text{（公式 12）}$$

$$F1=\frac{2\cdot R\cdot P}{R+P} \quad \text{（公式 13）}$$

4.2 评价结果

实验采用不同模型进行术语抽取，CRF模型使用词和词性作为特征，LSTM和BLSTM使用相同的词向量作为特征，分别在中文和英文语料上进行实验。在5组实验数据上中文数据实验结果如表4所示。

表 4 中文语料术语抽取评价结果

		CRF（词）	CRF（词 + 词性）	LSTM	BLSTM
1	P	0.7642	**0.7673**	0.7664	0.7538
	R	0.7698	0.7718	0.7801	**0.8020**
	F1	0.7670	0.7695	0.7732	**0.7771**
2	P	0.7630	0.7673	**0.7682**	0.7550
	R	0.7657	0.7698	0.7796	**0.7932**
	F1	0.7644	0.7686	**0.7738**	0.7736
3	P	0.7679	0.7715	0.7693	**0.7821**
	R	0.7770	0.7775	**0.7953**	0.7900
	F1	0.7724	0.7745	0.7820	**0.7860**
4	P	0.7671	0.7718	**0.7738**	0.7728
	R	0.7755	0.7776	0.7818	**0.7944**
	F1	0.7713	0.7747	0.7778	**0.7835**
5	P	0.7680	0.7709	0.7583	**0.7724**
	R	0.7737	0.7735	**0.7896**	0.7887
	F1	0.7708	0.7722	0.7741	**0.7805**

对于中文 TP 领域的数据，基于神经网络模型（LSTM，BLSTM）性能优于 CRF，CRF 模型（词 + 词性）特征的术语抽取性能整体高于 CRF（词特征），BLSTM 模型结果的 F1 值大多优于 LSTM，仅在第 2 组测试结果上稍弱于 LSTM，BLSTM 获得较高性能主要原因是结构获得了前向和后向两部分的上下文信息，捕获了更多特征。

英文数据上术语抽取的实验结果如图 5。

② 语料来源：中国科学技术信息研究所重点工作项目“日汉机器翻译双语资源建设与翻译引擎研发ZD2017-4”

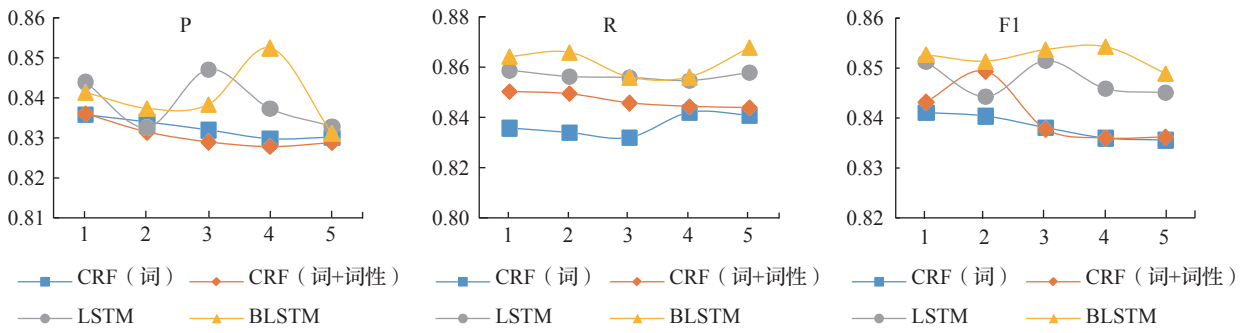

图 5 英文语料术语抽取评价结果

英文领域的术语抽取效果整体高于中文领域语料，其中P值最高为0.8525，R值最高为0.8677，F1值最高为0.8543。对比CRF采用的（词）和（词+词性）的方法，英文领域数据中加入词性特征后召回率提升，准确率下降，总体结果稍高于仅采用词特征的方法。对比LSTM和BLSTM的方法，在5组数据上BLSTM模型的召回率和F1值都优于LSTM，在第2和第4组数据上准确率高于LSTM。总体上看，BLSTM优于传统的CRF方法。LSTM，BLSTM在使用较少特征的情况下也可以获得较好的效果，进一步证明了神经网络模型的优势。

5 总结

本文使用基于双向长短时记忆BLSTM的模型对计算机、自动化领域下的中英文科技文献进行术语抽取，比较了该模型与传统的机器学习方法的性能。LSTM模型解决了长距离依赖问题，在此基础上BLSTM可以有效获取语料中更多信息。作为序列标注问题的一种有效的方法，BLSTM模型具有领域的通用性，可以在不依赖外部知识的情况下获得与传统方法相近甚至更好的抽取结果。

本文目前的研究为单语语料的术语抽取，如何拓展到使用双语对齐语料进行术语抽取，解决抽取术语和术语对齐问题是后续工作研究的目标。

参考文献

[1] 周浪, 史树敏, 冯冲, 等. 基于多策略融合的中文术语抽取方法[J]. 情报学报, 2010, 29(3):460-467.
[2] 屈鹏, 王惠临. 面向信息分析的专利术语抽取研究[J]. 图书情报工作, 2013, 57(1):130-135.
[3] 何远标, 乐小虬, 张帆. 学术论文大纲中关键术语抽取方法研究[J]. 现代图书情报技术, 2014, 30(3):73-79.
[4] 化柏林. 针对中文学术文献的情报方法术语抽取[J]. 现代图书情报技术, 2013(6):68-75.
[5] 曾文, 徐硕, 张运良,等. 科技文献术语的自动抽取技术研究与分析[J]. 现代图书情报技术, 2014, 30(1):51-55.
[6] 李丽双, 王意文, 黄德根. 基于信息熵和词频分布变化的术语抽取研究[J]. 中文信息学报, 2015, 29(1):82-87.
[7] Frantzi K T, Ananiadou S, Tsujii J. The C-value/NC-value Method of Automatic Recognition for Multi-word Terms[J]. Lecture Notes in Computer Science, 1998, 1513:585-604.
[8] Silva J F, Lopes G P. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units[J]. 1999.
[9] 韩红旗, 安小米. C-value值和unithood指标结合

doi:10.3772/j.issn.2095-915x.2018.01.008

的中文科技术语抽取[J]. 图书情报工作, 2012, 56(19):85-89.

[10] 胡阿沛, 张静, 刘俊丽. 基于改进C-value方法的中文术语抽取[J]. 现代图书情报技术, 2013, 29(2):24-29.

[11] 杨雅娜, 刘胜奇. 基于TValue融合领域度的术语抽取法[J]. 情报工程, 2015, 1(5):25-31.

[12] 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.

[13] 蒋婷, 孙建军. 基于SVR模型的中文领域术语自动抽取研究——面向图书情报领域[J]. 情报理论与实践, 2016, 39(1):24-31.

[14] 何宇, 吕学强, 徐丽萍. 新能源汽车领域中文术语抽取方法[J]. 现代图书情报技术, 2015, 31(10):88-94.

[15] 王昊, 王密平, 苏新宁. 面向本体学习的中文专利术语抽取研究[J]. 情报学报, 2016, 35(6):573-585.

[16] 姜霖, 王东波. 采用连续词袋模型(CBOW)的领域术语自动抽取研究[J]. 现代图书情报技术, 2016, 32(2):9-15.

[17] 张海楠, 伍大勇, 刘悦,等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4).

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

[19] Graves A, Schmidhuber J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks the Official Journal of the International Neural Network Society, 2005, 18(5-6):602.

[20] 朱丹浩, 杨蕾, 王东波. 基于深度学习的中文机构名识别研究——一种汉字级别的循环神经网络方法[J]. 现代图书情报技术, 2016, 32(12):36-43.

[21] 任智慧, 徐浩煜, 封松林,等. 基于LSTM网络的序列标注中文分词法[J]. 计算机应用研究, 2017, 34(5):1321-1324.

[22] 万圣贤, 兰艳艳, 郭嘉丰,等. 用于文本分类的局部化双向长短时记忆[J]. 中文信息学报, 2017, 31(3):62-68.

[23] Bishop C M. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]. Springer-Verlag New York, Inc. 2006.

[24] Bengio Y, Vincent P, Janvin C. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2006, 3(6):1137-1155.

[25] MIKOLOV T, CORRADO G, CHEN K, et al. Efficient Estimation of Word Representations in Vector Space[C]. International Conference on Learning Representations. Scottsdale, 2013:1-12.

[26] Manning C D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]. Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.