基于注意力机制的包装命名实体识别

冀相冰[1,2]，朱艳辉[1,2]，徐啸[1,2]，梁文桐[1,2]，詹飞[1,2]
（1.湖南工业大学 计算机学院，湖南 株洲 412008；
2.湖南省智能信息感知及处理技术重点实验室，湖南 株洲 412008）

摘要：目的 为了解决包装行业相关文本命名实体识别困难问题，提出在 BiLSTM（Bidirectional Long Short-Term Memory）神经网络中加入注意力机制（Attention）和字词联合特征，构建一种基于注意力机制的 BiLSTM 深度学习模型（简称 Attention-BiLSTM），以识别包装命名实体。**方法** 首先构建包装领域词典匹配包装语料中词语的类别特征，同时将包装语料转换为字特征和词特征联合的向量特征，并且在过程中加入 POS（词性）信息。然后将以上特征联合馈送到 BiLSTM 网络，以获取文本的全局特征，并利用注意力机制获取局部特征。最后根据文本的全局特征和局部特征使用 CRF（Conditional Random Field）解码整个句子的最优标注序列。**结果** 通过对《中国包装网》新闻数据集的实验，获得了 85.6%的 *F* 值。**结论** 所提方法在包装命名实体识别中优于传统方法。

关键词：命名实体识别；包装；注意力机制；BiLSTM；字词联合特征
中图分类号：TP391.1　　**文献标识码：**A　　**文章编号：**1001-3563(2019)15-0024-06
DOI：10.19554/j.cnki.1001-3563.2019.15.004

Packaging Named Entity Recognition Based on Attention Mechanism

JI Xiang-bing[1,2], *ZHU Yan-hui*[1,2], *XU Xiao*[1,2], *LIANG Wen-tong*[1,2], *ZHAN Fei*[1,2]
(1.School of Computer, Hunan University of Technology, Zhuzhou 412008, China;
2.Hunan Key Laboratory of Intelligent Information Perception and Processing Technology, Zhuzhou 412008, China)

ABSTRACT: The work aims to add attention mechanism (Attention) and Joint Characteristics of Words in BiLSTM (Bidirectional Long Short-Term Memory) neural network to construct a BiLSTM deep learning model (Attention-BiLSTM) based on attention mechanism, so as to solve the problem of difficult identification of text-related entities in the packaging industry and recognize the packaging named entity. Firstly, the packaging domain dictionary was built to match with the category features of the words in the packaging corpus, and the packaging corpus was converted into the vector features of the word feature and the character feature, and then POS (part of speech) information was added in the process. The above features were then fed jointly to the BiLSTM network to obtain the global features of the text, and the attention mechanism was used to acquire the local features. Finally, the CRF (Conditional Random Field) was used to decode the optimal label sequence of the entire sentence according to the global features and local features of the text. The final *F* score was 85.6% on the "China Packaging Network" news dataset. The proposed method is superior to the traditional method in packaging named entity recognition.

收稿日期：2019-04-27
基金项目：国家自然科学基金（61402165）；湖南省自然科学基金（2018JJ2098）；湖南工业大学重点项目（17ZBLWT001KT006）
作者简介：冀相冰（1991—），男，湖南工业大学硕士生，主攻自然语言处理与知识工程。
通信作者：朱艳辉（1968—），女，湖南工业大学教授，主要研究方向为自然语言处理与知识工程。

KEY WORDS: named entity recognition; packaging; attention mechanisms; BiLSTM; joint characteristics of word

在大数据、“互联网+”和“工业 4.0”迅速发展的浪潮中，如何从浩如烟海的互联网数据中提取有价值的信息成为目前的研究热点，为此知识图谱技术应运而生。知识图谱也被称为科学知识图谱或知识域可视化，是用可视化技术描述知识资源及其载体，挖掘、分析、构建和显示知识与它们之间的相互联系，并以图的方式展示知识。中国包装产业大数据知识图谱的构建和应用是服务包装行业、包装企业决策的需要，将填补包装行业大数据领域的空白，为包装企业提供重要的智库支持，可以实时了解和分析我国包装行业的现状、规模和动态发展，最终促进企业的创新力、竞争和特色发展。中国包装产业大数据知识图谱的构建离不开大规模的数据支撑，包装行业数据分散的特性和非结构文本的内容丰富性决定了对包装命名实体识别的研究成为构建包装产业大数据知识图谱数据来源的主要途径。包装命名实体是指能体现出包装领域特色的特定名词与短语，覆盖包装工程、印刷工程、艺术设计学等多个学科，如：包装机械、包装制品和包装材料等，因其种类繁多，覆盖范围较广，因此信息抽取比较困难。

近年来，NLP（Natural Language Processing）学者们把基于规则、统计学习和深度学习等命名实体识别方法应用在各个领域，并进行深入研究。Lample[1]采用双向 LSTM 和条件随机场对多种语言进行命名实体识别。冯艳红[2]把领域知识应用到神经网络中提升了识别性能。Wang[3]利用片段级神经网络捕获片段信息进行中文命名实体识别，避免了字符序列模型的缺陷。Borthwick[4]把最大熵统计模型应用到命名实体识别中，对信息提取任务具有特殊意义。Feng[5]利用词嵌入+CRF 的领域术语识别方法解决了传统方法忽视语义的问题。Bahdanau[6]等在神经网络中引入注意力机制，解决了机器翻译问题。Wang[7]利用一种混合深度神经网络（DNN）挖掘嵌入在无标签语料库中的隐式信息。Chiu[8]采用在神经网络编码部分匹配词典的方法，取得了良好性能。张海楠[9]把字特征与词特征联合统一起来，提升了 NER 系统性能。Sundermeyer 等[10]通过 LSTM 神经网络进行语言建模，以评估模型，验证 2 个量的强相关性。侯伟涛[11]应用双向 LSTM 对医疗文本事件进行抽取，得到的研究效果优于以往的研究效果。

包装语料具有字词边界比较模糊且有一词多义的特点，传统仅利用字特征或词特征进行命名实体识别的效果欠佳。传统方法在提取特征过程中，因过于重视文本全局特征，从而忽视了局部特征对命名实体识别效果的重要影响。在包装领域进行命名实体识别时，直接应用通用方法较困难，因为垂直领域内的命名实体划分细致，且非专业人士一般无法辨别命名实体类别等信息。

文中提出基于注意力机制的包装命名实体识别方法，采用注意力机制对文本进行重要度计算，以获得文本局部特征，解决以往方法不能充分提取文本特征的问题；构建包装领域词典，解决通用方法应用于包装领域识别率低的问题；利用字词联合特征，解决因字特征窗口限制和词特征分词错误导致上下文信息缺失的问题。

1　包装领域词典的构建

1.1　词典构建方法

为了弥补知识领域的不足，特邀请包装领域专业学者协助进行包装相关的词语分类，词典构建详细步骤如下所述。

1）通过网络爬虫抓取包装语料，通过分词、去重等文本预处理之后获得仅包含包装实体的文件。

2）请包装领域专家协助对包装实体进行归纳分析，将领域内实体由大到小划分为一、二、三级标签，其中一级标签包含 10 个种类；二级标签为一级标签的子类别；三级标签为包装命名实体。

3）为每个一级标签自定义专属标记，采用标签的英文缩写方式，例如企业标记为 CO，人标记为 PER 等。

1.2　词典构建结果

在包装领域词典中，一级标签包括企业、人、专利、机构、技术说明及其他出版物、包装知识点、论文、事件、产品和地点等 10 个类别。二、三级标签分类见表 1，因篇幅限制，仅列出部分示例。对于人、专利、技术说明及其他出版物、包装知识点、论文等 5 种标签，考虑其特点，未进行细致划分，一级标签与二级标签相同。

完成包装词典构建之后，可根据词典中的语义信息匹配出包装语料中命名实体的类别特征，然后把类别特征与其他特征一起传入神经网络学习。

2　包装命名实体识别模型

2.1　Attention-BiLSTM 模型

传统命名实体识别过程中通常用循环神经网络（RNN）捕获长距离依赖关系[12]，但是这样容易出现梯度爆炸或消失等问题。后来出现的 LSTM[13]可以很好地解决这类问题，但是单向 LSTM 只能获取过去的

信息，对未来的状态信息不可知，GRAVES[14]提出的解决方案是利用 BiLSTM 网络。

文中提出在 BiLSTM 神经网络中加入注意力机制，基于上下文的语义信息可以弥补深度网络获取局部特征方面的不足。文本的局部特征表示文本中部分内容之间的关联特征，例如在句子“引入自动化生产设备实现智能制造”中，“自动化生产设备”是一个命名实体，各个字之间的关联密切，而实体前一个字“入”和后一个字“实”与它的关联较弱。

文中提出的模型结构见图 1。它的基本思路是首

表 1　包装领域词典示例

Tab.1 Dictionary for packaging field

一级标签	二级标签	三级标签	标记
企业	国有企业	湖南省包装总公司	CO
企业	民营企业	甘肃华丽包装制品有限公司	CO
人	人	邹明龙	PER
产品	设计	药品包装设计	PRO
产品	材料	胶印耗材	PRO
事件	偶然事件	加多宝包装纠纷案	EVT
事件	常规事件	亚洲纸展	EVT
机构	学校	湖南工业大学	ORG
机构	政府机关	安徽省质监局	ORG
专利	专利	牡丹盆花寄递专用包装	ZL
地点	省	湖南省	LOC
地点	市	东莞市	LOC

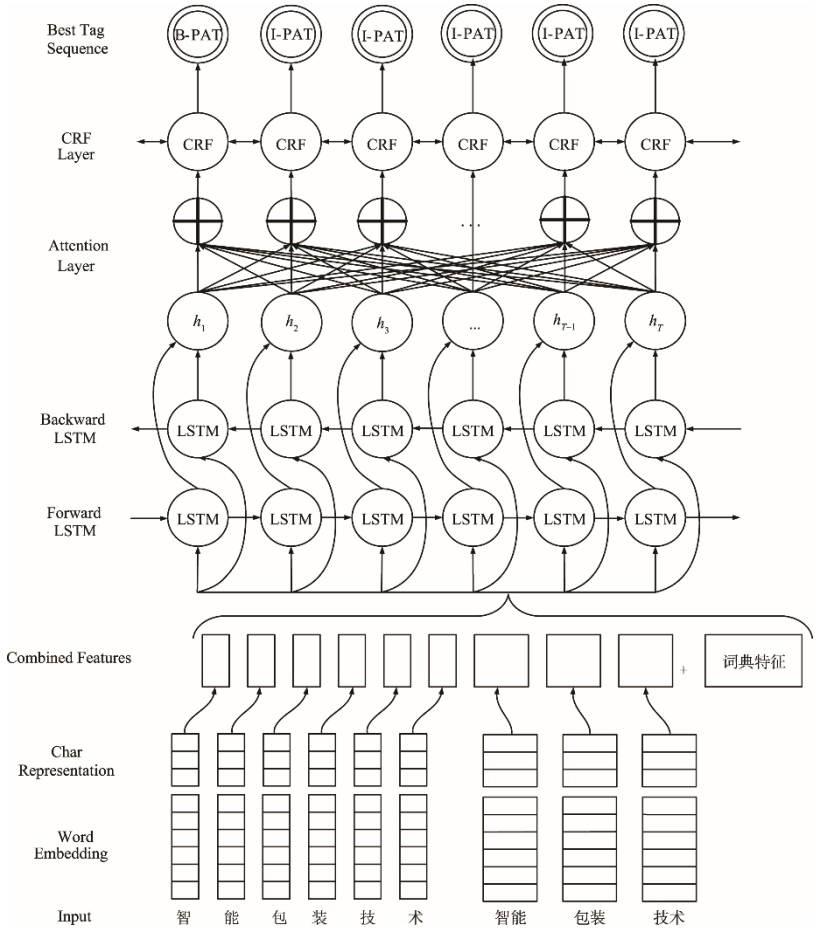

图 1　Attention-BiLSTM 模型结构

Fig.1 Structure of Attention-BiLSTM model

先把字词联合向量序列和包装词典类别特征输入 BiLSTM 网络，其中 LSTM 将所有序列展开为 2 个单独的隐藏状态，其中一个向前捕获历史的信息，另外一个向后捕获未来的信息，形成全局特征输出。然后将 BiLSTM 的输出向量输入 Attention 机制给全局特征中不同的特征向量赋予不同的权重，以提取局部特征[15]，最后将包括全局特征和局部特征的字词联合特征向量序列馈送到 CRF 层架构神经网络模型。

定义 $x_1, x_2, x_3 \cdots x_{T-1}, x_T$ 为 BiLSTM 神经网络输入的字词联合向量序列；a_{tj} 为 Attention 机制给所有特征向量赋予的权重，公式如下：

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})} \tag{1}$$

其中：

$$e_{tj} = a(c_{t-1}, h_j) = v_a^T \tan(w_a c_{t-1} + u_a h_j) \tag{2}$$

式中：c_{t-1} 为注意力模型上一时刻状态；v_a 为全局的权值；h_j 为 BiLSTM 神经网络输出的特征向量序列；u_a 为上一时刻的特征向量的权值；w_a 为上一时刻注意力机制的状态的权值。

注意力机制模型最后的输出状态 c_t 计算方法见式（3）。

$$c_t = \sum_{j=1}^{T} a_{tj} h_j \tag{3}$$

2.2　字词联合特征

在包装命名实体识别中，仅利用字特征或者词特征进行 NER 的效果不佳，采用字词联合方法构建特征向量（见图 2）能有效提升系统性能。

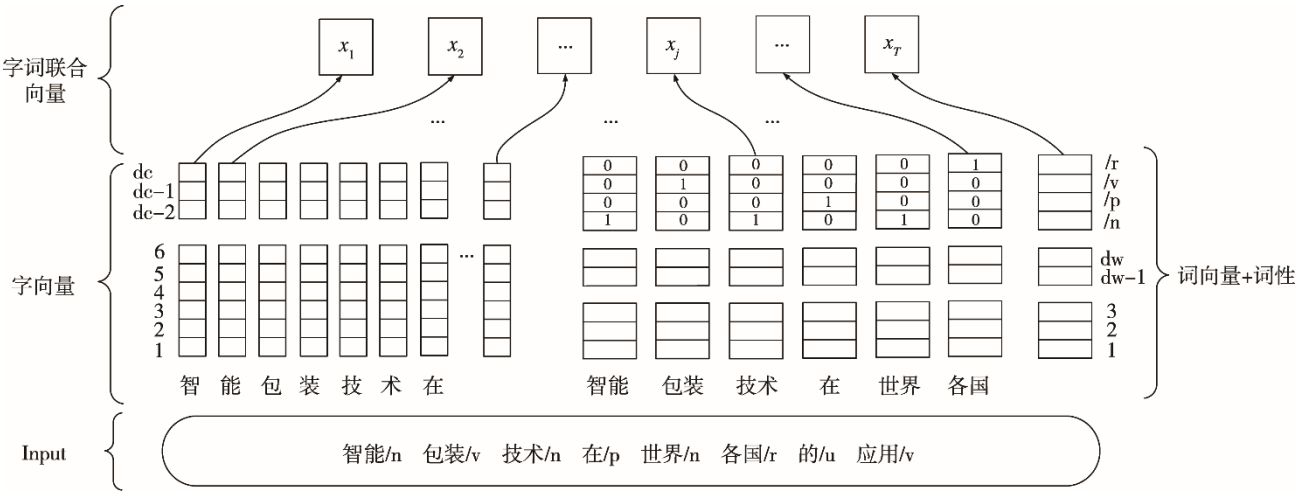

图 2　字词联合特征

Fig.2 Joint characteristics of words

为了充分挖掘包装语料中的特征，预定义字典 D_C 和词典 D_W[9]，利用字向量矩阵 $\boldsymbol{M}^c \in R^{d^c \times |D_C|}$ 存储字向量特征，利用词向量矩阵 $\boldsymbol{M}^w \in R^{d^w \times |D_w|}$ 存储词向量特征。

在字词联合向量中，预定义字词之间的映射：

$$c_i \in w_j \tag{4}$$

设定包装命名实体识别的标签种类为 T，BiLSTM 神经网络最后输出一个 $|T|$ 维的向量，使用“B，I，O”等 3 种标签对包装命名实体进行标注。它表示滑动窗口输入字 c_i 的标签概率，定义输出向量的预测函数如下：

$$\text{Score}(\theta, c_i) = f_3\left(g(f_2(f_1(c_i)))\right) = W_{\text{out}} \times g(W_{\text{hid}} \times f_1(c_i) + b_{\text{hid}}) + b_{\text{out}} \tag{5}$$

在字词联合特征向量中加入 POS（词性）信息，图 2 中左边部分为字向量特征，右边部分为带 POS 的词向量特征。以“智能包装技术在世界各国的应用”为例，通过中文分词和词性标注处理，然后使用 word2vec 工具训练文本字向量和词向量，最后把字向量和词向量的联合特征作为这句话的整体特征输入下一层神经网络。

3　实验与分析

3.1　数据集及标注模式

实验数据集采用中国包装网 2017 年的新闻语料，共计 500 篇，并按照 6∶2∶2 的比例进行划分，见表 2。包装语料中命名实体类别见表 3。

表 2　实验数据集划分

Tab.2 Division of experimental data set

类别	数据量/MB
训练集	6.18
验证集	1.53
测试集	1.55

表 3　包装实体类别
Tab.3 Packaging entity categories

实体类别	标注编码
企业	CO
人	PER
专利	ZL
机构	ORG
技术说明及其他出版物	PAT
包装知识点	KNP
论文	PAP
事件	EVT
产品	PRO
地点	LOC

实验采用的标注模式为 BIO 模式。例如：包装企业相关实体标注为 B-CO/I-CO。

3.2　实验设计

为了验证文中模型方法在包装语料集上的命名实体识别效果，分别与其他 NER 方法进行实验对比。在模型中加入包装领域词典和 POS 信息（见表 4）。

使用 NLPIR 中文分词系统对原始文本进行加工和处理，采用 word2vec 工具训练文本词向量，然后利用 Tensorflow 进行建模和数据处理。

文中实验的软硬件环境配置情况见表 5。

3.3　BiLSTM 参数设置

采用 L2 正则化算法和 Dropout 技术避免试验数据过拟合问题。列举了包装 NER 实验所需的参数，见表 6。

表 4　词性信息
Tab.4 Part-of-Speech information

词性标记
Ag,Bg,Dg,Mg,Ng,Rg,Tg,Vg,Yg,a,ad,an,b,c,d,e,f,h,I,j,k,l,m,n,nx,nz,o,p,q,r,s,t,u,v,vd,vn,w,y,z

表 5　软硬件环境
Tab.5 Software and hardware configuration

项目	环境
系统	Ubuntu16.04 LTS
GPU	NVIDIA QuadroK1200
硬盘	1T
内存	16 GB
Python版本	Python 3
TensorFlow版本	TensorFlow 1.2.1

表 6　神经网络参数设置
Tab.6 Parameter settings of neural network

参数	值
学习率	0.002
L2正则化	0.0001
词向量维	200
Batch size	10
隐藏节点个数	150
词向量窗口大小	10
Dropout	0.5
迭代次数	120

3.4　实验结果及分析

为了验证文中方法的识别效果，分别采用 CRF，LSTM 和 BiLSTM 方法在同一包装语料集上进行实验。实验对比结果见表 7。

表 7　不同命名实体识别方法对比
Tab.7 Comparison of different named entity identification methods

方法	评价标准		
	准确率/%	召回率/%	*F*值/%
CRF	68.4	62.7	65.4
LSTM	79.8	75.1	77.4
BiLSTM	80.2	79.5	79.8
Attention-BiLSTM	83.2	81.4	82.3
Attention-BiLSTM-词典	86.7	84.6	85.6

通过不同命名实体识别方法在同一包装语料上的实验可以观察到，与传统 CRF，LSTM，BiLSTM 等方法比较，文中模型 *F* 值得到明显提升。这是因为汉语的语言特点，基于字向量的 NER 容易受到窗口大小的限制，词向量容易受到因为词典稀疏导致分词错误的影响，而利用字词向量特征的结合很好地解决了这些问题。另外，文中方法加入 Attention 机制进一步加强了模型的标记预测能力。因为不同的字词对上下文的贡献程度不一样，在特征提取过程中加入文本的局部特征，有效弥补了传统仅注重全局特征提取的缺陷，且可以抽取更多的文本上下文特征。从以上实验数据还可以观察到，加入包装领域词典特征之后，准确率、召回率和 *F* 值比未加入词典亦得到较好的提升，说明在包装命名实体识别中加入包装词典信息对模型性能提供的帮助较大，验证了文中方法在包装命

名实体识别系统中的有效性。

4　结语

采取基于注意力机制的包装命名实体识别方法构造包装领域词典，能够更好地学习包装语料中的复杂语义关系。注意力机制能够获取文本的局部特征，抑制文本冗余信息，很好地解决了传统模型抽取特征不全面的弊端。应用字特征和词特征的联合向量进行包装领域的实体识别，解决了因字特征或词特征本身缺陷导致系统识别性能较差的问题。实验证明，该方法可以很好地完成对包装命名实体的识别任务。文中方法仅针对包装领域进行了命名实体识别，对于其他语言和领域的序列标注任务还未进行细致研究。接下来将考虑丰富领域词典特征或者加入其他更好方法应用到其他领域的研究中。

参考文献：

[1] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural Architectures for Named Entity Recognition[C]// Proceedings of NAACL-HLT 2016, San Diego, 2016: 260—270.

[2] 冯艳红, 于红, 孙庚, 等. 基于 BLSTM 的命名实体识别方法[J]. 计算机科学, 2018, 45(2): 261—268.
FENG Yan-hong, YU Hong, SUN Geng, et al. Named Entity Recognition Method Based on BLSTM[J]. Computer Science, 2018, 45(2): 261—268.

[3] 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.
WANG Lei, XIE Yun, ZHOU Jun-sheng, et al. Segment-level Chinese Named Entity Recognition Based on Neural Network[J]. Chinese Journal of Information Science, 2018, 32(3): 84—90.

[4] BORTHWICK A. A Maximum Entropy Approach to Named Entity Recognition[D]. New York : New York University, 1999.

[5] FENG Y H, YU H, SUN G, et al. Domain-specific Terminology Recognition Method Based on Word Embedding and CRF[J]. Journal of Computer Applications, 2016, 36(11): 3146—3151.

[6] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]// International Conference on Learning Representations 2015, San Diego, 2015.

[7] WANG G Y, CAI Y Q, GE F J. Using Hybrid Neural Network to Address Chinese Named Entity Recognition[C]// IEEE,International Conference on Cloud Computing and Intelligence Systems, 2014: 433—438.

[8] CHIU J P C, NICHOLS E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357—370.

[9] 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): 28—35.
ZHANG Hai-nan, WU Da-yong, LIU Yue, et al. Chinese Named Entity Recognition Based on Deep Neural Network[J]. Journal of Chinese Information Processing, 2017, 31(4): 28—35.

[10] SUNDERMEYER M, NEY H, SCHLUTER R. From Feed Forward to Recurrent LSTM Neural Networks for Language Modeling[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2015, 23(3): 517—529.

[11] 侯伟涛, 姬东鸿. 基于 Bi-LSTM 的医疗事件识别研究[J]. 计算机应用研究, 2018, 35(7): 1974—1977.
HOU Wei-tao, JI Dong-hong. Research on Medical Event Recognition Based on Bi-LSTM[J]. Application Research of Computers, 2018, 35(7): 1974—1977.

[12] MAX X, 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.

[13] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735—1780.

[14] GRAVES A, SCHMIDHUBER J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks, 2005, 18(5): 602—610.

[15] GUL K S Q, 尹继泽, 潘丽敏, 等. 基于深度神经网络的命名实体识别方法研究[J]. 信息网络安全, 2017(10): 29—35.
GUL K S Q, YIN Ji-ze, PAN Li-min, et al. Research on Named Entity Recognition Method Based on Deep Neural Network[J]. Information Network Security, 2017(10): 29—35.