

基于 BLSTM-CNN-CRF 的中文命名实体识别方法

刘宇鹏, 栗冬冬

(哈尔滨理工大学 软件与微电子学院, 哈尔滨 150080)

摘 要: 传统的命名实体识别方法依赖大量的人工选择的特征和专业领域的外部知识, 针对这一问题, 提出了一种新颖的神经网络结构, 该算法结合了双向 LSTM, CNN 和 CRF 可以同时自动获取到基于字符级别和词语级别的表示, 是一种真正意义上的端到端的结构, 不再需要人工选择特征和数据的预处理, 可以应用到各个领域的命名实体识别任务中去。最后, 通过实验证明该算法在医疗领域和新闻领域的 F1 值分别达到了 90.97% 和 92.19%。

关键词: 命名实体识别; 长短期记忆网络; 卷积神经网络; 条件随机场

DOI: 10.15938/j.jhust.2020.01.017

中图分类号: TP391.1 **文献标志码:** A **文章编号:** 1007-2683(2020)01-0115-06

Chinese Named Entity Recognition Method Based on Bi-directional LSTM-CNN-CRF

LIU Yu-peng, LI Dong-dong

(School of Software and Microelectronics, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: Traditional algorithms of named entity recognition rely on a large number of artificial features and external knowledge. To solve this problem, we propose a novel neural network structure which combines the bidirectional LSTM (Long Short-Term Memory), CNN (Convolution Neural Network) and CRF (Conditional Random Field). This algorithm can also automatically get both character-level and word-level representations. It is a truly end-to-end system, requiring no feature engineering and data processing. Thus, it can be easily applied to various fields of named entity recognition task. Finally, the experimental results show that the F1 measure of the algorithm in the medical and news field is 90.97% and 92.19%, respectively.

Keywords: named entity recognition; long short-term memory; convolution neural network; conditional random field

0 引 言

命名实体识别(named entity recognition, NER)作为自然语言处理的一项基础任务,在自然语言处

理领域有着很重要的作用^[1-2],分为在不同语言上的命名实体识别^[3-4],微博上的命名实体识别^[5]和 Twitter 上的命名实体识别^[6-8]等。在机器对话、问答系统、机器翻译等任务中都有着重要的应用。传统方法中表现最优的就是基于统计的模型,例如隐

收稿日期: 2018-03-08

基金项目: 国家自然科学基金(61300115)。

作者简介: 栗冬冬(1991—),男,硕士研究生。

通信作者: 刘宇鹏(1978—),男,博士,副教授, E-mail: flyeagle99@126.com。

马尔可夫模型(hidden markov models HMM)、条件随机场模型(conditional random fields, CRF)^[8],这些方法需要大量的人工选择的特征和外部知识信息,例如字典信息,很难直接解决一个全新领域的命名实体识别问题。近年来,利用神经网络获取词语表示作为模型的输入(例如词向量)在自然语言处理领域取得了巨大的成功^[10-12]。循环神经网络及其变形长短期记忆网络(long-short term memory, LSTM)在解决序列标注任务中表现优异^[13-15]。

本文提出了一种结合双向 LSTM、卷积神经网络(convolutional neural network, CNN)、CRF 的方法,首先利用 CNN 获取词语基于字符的表示,然后结合词向量加入到双向 LSTM 中,最后利用线性 CRF 联合解码标签。该模型不需要外部知识和手工选择特征,在不同领域的数据上进行实验验证了其可行性和有效性。

1 问题描述

NER 是指从一段文本中提取出具有特定含义的词语,即实体^[14]。实体的类型有很多种,常见的有人名、地名、机构名等,在一些特定领域会有其独特的实体,例如医疗领域的病症名称、药物名称等。这些实体信息对进一步分析句子含义有很大的帮助。

一般会把 NER 问题看成序列标注问题,标注问题一般是监督学习问题,也可以看成是一种复杂的结构预测问题的简单形式^[17]。NER 任务分为学习和标注两个过程,首先有一个带标注信息的数据集作为训练数据。

$$T = \{ (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) \} \quad (1)$$

式中 $x_i = (x_i^1, x_i^2, \dots, x_i^n)^T$, $i = 1, 2, 3, \dots, N$ 是输入的观测序列,相当于一个句子,其中的每一项相当组成句子的字或者词。

$y_i = (y_i^1, y_i^2, \dots, y_i^n)^T$ 是输入所对应的输出的预测标注序列,即对输入句子的每个字或词的标注结果。根据训练数据系统需要学习出一个模型,用条件概率分布表示为

$$P(Y^1, Y^2, \dots, Y^n | X^1, X^2, \dots, X^n) \quad (2)$$

式中每个 x^i 表示所有可能的观测,对应的每个 y^i 表示其标注的所有可能。系统学习得到模型后就可以对新的输入序列计算得出其对应的标注序列。

2 神经网络结构

2.1 CNN 提取字符级别的表示

之前的研究表明卷积神经网络(convolutional neural network, CNN)在提取局部特征方面十分有效^[18],因此在本系统中先利用 CNN 提取出词语基于字符的特征将其编码到神经网络中。其结构表示图如图 1 所示, CNN 的输入为字符的 embedding,图 1 中虚线部分为 dropout 层。

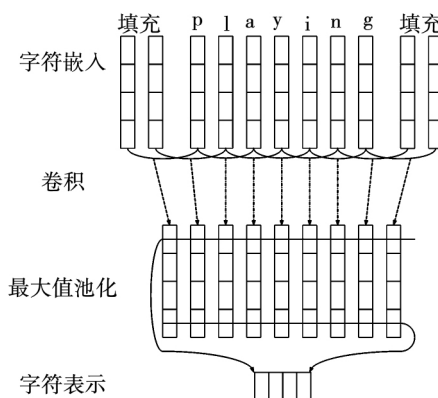


图1 CNN 提取词语字符级别表示图

Fig.1 Char-level Word Representation Using CNN

2.2 双向 LSTM

循环神经网络(recurrent neural networks, RNNs)是一种可以捕捉到时序信息的网络结构,但在获取长距离依赖信息时 RNNs 会出现梯度消失问题。LSTM 是 RNNs 的一种变形,它可以解决 RNN 中的梯度消失问题。一般地, LSTM 单元由 3 个不同功能的‘门’组成,遗忘门(forget gate)、输入门(input gate)、输出门(output gate),这种‘门’的结构用来控制信息在传递过程中的取舍。图 2 展示了 LSTM 单元的基本结构。在 LSTM 具体实现过程中,在时刻 t 时 LSTM 单元通过以下公式来实现更新:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \quad (3)$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \quad (4)$$

$$\tilde{c}_t = \tanh(W_c h_{t-1} + U_c x_t + b_c) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (6)$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \quad (7)$$

$$h_t = o_{t-1} \odot \tanh(c_t) \quad (8)$$

其中: σ 是 sigmoid 函数; $\odot$ 是元素对应乘积; x_t 是 t 时刻的输入向量(例如词向量 wordembedding); h_t 是隐含状态向量(也称为输出向量), h_t 中储存着 t 时刻和 t 之前时刻的所有信息。 U_i, U_f, U_c, U_o 是不

同‘门’对输入向量 x_t 的权重; W_i, W_f, W_c, W_o 是不同‘门’对隐含状态向量 h_t 的权重; b_i, b_f, b_c, b_o 是偏置向量。

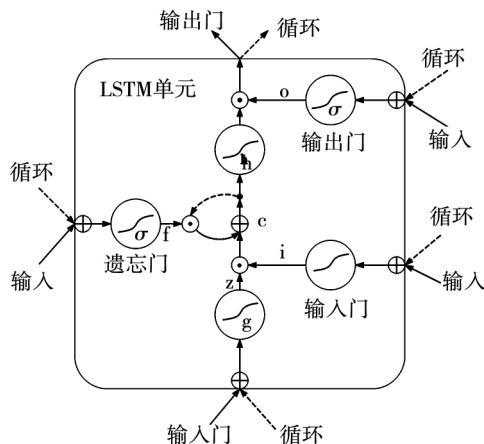


图2 LSTM单元结构图

Fig. 2 Cell Structure of LSTM

实体识别问题本质是一种序列标注问题,对许多序列标注问题来说,实体标注不仅和其当前时刻之前的信息相关,也和其之后的信息有关,并且有特殊情况即输入数据的最前端是没有上文信息的,要想判断其含义,利用其下文信息也是可以。所以判断一个词语的上下文信息对判断其是否为实体有着重要的作用。然而 LSTM 的隐藏状态 仅仅捕获到了当前时刻 t 之前的信息,并不能获取之后的信息。双向长短期记忆网络(bi-directional long and short term memory, BLSTM) 能够强有力的解决这一难题。

BLSTM 的基本思想其实很简单,在 LSTM 的基础上增加一个向后的循环神经网络,然后共同输出到同一个输出层上。这样就可以利用到数据的过去和未来的信息。这种结构的网络可以给输入层序列中的每一个单元某时刻其过去的信息和将来的信息。因此在这里选取了双向的 LSTM 结构。

2.3 条件随机场 CRF

由于 BLSTM 很好的利用了上下文信息,但不能充分利用输出信息的预测标签信息,就是预测结果中可能出现{ B-Body, I-Symptoms, ... } 这种情况。因此在模型最上层加入了 CRF 层,把 BLSTM 的输出信息作为输入信息加到 CRF 中。在实体识别任务中,输入是一个句子文本,如果能利用上相邻标签的相关性信息对解码出最好的预测结果有着很大的帮助。例如在词性标注问题中一个形容词之后的词的词性是名词词性的概率要大于动词的词性,在 NER 中,按照 BIOES 标准标注^[9],标签 I-PER 后面的标

签不可能是 I-ORG。因此,本文中模型采用条件随机场 CRF 来联合解码,而不是单独的解码每一个标签。

一般地,用 $z = \{z_1, \dots, z_n\}$ 来表示一个输入序列,其中 z_i 表示第 i 个输入词的向量, $y = \{y_1, y_2, \dots, y_n\}$ 表示输入 z 的标签序列。 $Y(z)$ 表示 z 的可能标签序列。序列 CRF 概率模型定义了一系列的条件概率 $p(y|z; W, b)$, 给定 z 所有可能的序列 y 的概率由以下公式计算得出:

$$p(y|z; W, b) = \frac{\prod_{i=1}^n \varphi(y_{i-1}, y_i, z)}{\sum_{y' \in Y(z)} \prod_{i=1}^n \varphi(y'_{i-1}, y'_i, z)} \quad (9)$$

式(9)中 $\varphi(y'_i, y, z) = \exp(W_{y'_i, y}^T + b_{y'_i, y})$ 是势函数, $W_{y'_i, y}^T$ 和 $b_{y'_i, y}$ 分别是对于标签对 (y', y) 的权重向量和偏置。

对于 CRF 模型训练,使用最大似然估计。对于训练集 $\{(z_i, y_i)\}$, 似然估计算法由公式(10)给出:

$$L(W, b) = \sum \log p(y|z; W, b) \quad (10)$$

最大似然训练过程就是选择参数使得 $L(W, b)$ 最大。解码过程就是寻找条件概率最大的标注序列 y^* :

$$y^* = \operatorname{argmax}_{y \in Y(z)} p(y|z; W, b) \quad (11)$$

对于序列 CRF 模型,只考虑两个连续标签之间的相互作用,训练和解码过程采用维特比算法(Viterbi algorithm) 即可取得较好的效果。

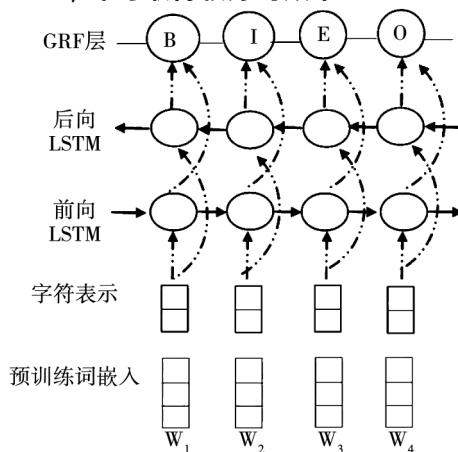


图3 BLSTM-CNN-CRF 模型图

Fig. 3 BLSTM-CNN-CRF Model

2.4 BLSTM-CNN-CRF 模型

最后,CRF 输出最终的标注结果,由此得到混合模型,此模型结合了深度学习的方法和传统的机器学习算法 CRF,取得了较好的预测效果。图3 详细地介绍了这种模型的结构。

在图3 中最下方的预训练的 wordembedding 是利用未标注数据预训练的一个词向量,对于每个输

入序列,通过 CNN 计算出字符级别的表示,然后把字符级别的表示和预先训练好的词向量结合到一起输入到 BLSTM 网络中^[24]。最后 BLSTM 的输出向量作为输入加入到 CRF 中,然后联合解码出最优标注序列。如图 3 中所示,在 BLSTM 网络的输入和输出向量中都加入了 dropout 层。实验表明使用 dropout 层可以显著的提高模型的性能。

3 模型参数配置及实验评价标准

3.1 模型参数配置

本文中所有的实验是在 Theano 0.9.0 平台上做的,每个模型的计算都是在 GPU 上完成的。

权重矩阵和偏置向量: 矩阵参数从 $[-\sqrt{\frac{6}{r+c}}, +\sqrt{\frac{6}{r+c}}]$ 随机初始化,其中 r 和 c 是矩阵和行数和列数。偏置向量除了 LSTM 的遗忘门中的偏置 b_f 初始化为 1.0,其他均为 0。

模型的参数采用随机梯度下降法来进行优化,模型训练的批处理大小 (minibatch) 设置为 100,动量 (momentum) 设置为 0.9。初始学习率 η_0 (learning rate) 设置为 0.015,学习率每轮按照公式 $\eta_t = \eta_0 / (1 + \rho t)$ 更新,其中衰减率 (decay rate) $\rho = 0.05$, t 是训练的轮数,表格 1 总结了实验中的参数配置情况。

表 1 模型参数
Tab. 1 Model Parameter

	超参数	数值
CNN	窗口大小	3
	过滤器个数	30
LSTM	状态大小	200
	初始状态	0.0
	孔洞	无
	Dropout 率	0.5
其它	批量大小	100
	初始学习率	0.015
	衰减率	0.05
	梯度剪裁	5.0

通过随机搜索调整开发集上的参数。由于时间的限制,在整个参数空间上进行随机搜索是不可行的,因此,对于实体识别任务,试着尽可能多地共享参数。在模型中设置 LSTM 的 state size 为 200,调整此参数并没有显著影响模型的性能。对于 CNN,则采用了 30 个滤波器,窗口大小设置为 3。

在训练过程中,对每个 embedding 会微调 (fine-tune) 其初始 embeddings,并在神经网络反向传播梯度更新过程中进行修正。这种方法在序列和结构化标注预测问题中十分有效。

在一般模型训练过程中出现过拟合 (overfitting) 问题,为了避免这一现象,在模型中使用了 dropout 方法来正则化模型,在图 4 中可以看到,在字符 embedding 加入 CNN 之间添加了 dropout 层并且在 BLSTM 的输入和输出都加了 dropout 层,在实验中的所有 dropout rate 都设置为 0.5。

3.2 实验评价标准

本实验的任务为输入一组原始的语料数据,输出给定句子或文档中出现的实体名字 (字符串边界),并把每个实体归到其预先规定好的类别中去。例如:

输入: 右眼睑及右侧面部肿胀、淤血

输出: 右眼睑 0-2 身体部位; 右侧面部 5-7 身体部位; 肿胀 8-9 症状和体征; 淤血 11-12 症状和体征

本任务采用实体识别中常用的 3 个指标作为评价指标,分别是精确率 (Precision)、召回率 (Recall) 以及 F1-Measure。输出结果集合记为 $S = \{s_1, s_2, \dots, s_m\}$,人工标注的结果 (Gold Standard) 集合记为 $G = \{g_1, g_2, \dots, g_n\}$ 。定义 $s_i \in S$ 与 $g_j \in G$ 严格等价,当且仅当:

1. $s_i.mention = g_j.mention$
2. $s_i.pos_b = g_j.pos_b$
3. $s_i.pos_e = g_j.pos_e$
4. $s_i.category = g_j.category$

基于以上等价关系,定义集合 S 和 G 的交集为 $\cap$,由此得到评价指标:

$$P = \frac{|S \cap G|}{|S|}, R = \frac{|S \cap G|}{|G|}, F_1 = \frac{2PR}{P+R}.$$

4 实验设计及分析

为了验证本文中模型的通用性良好,分别在新闻领域和医疗领域的数据上进行了实验。新闻数据采用的是 1998 年的人民日报数据,实体类别包括人名、地名和机构名。医疗数据采用的是 2017 年全国知识图谱与语义大会 (CCKS2017) 发布的数据,实体类别包括症状和体征、检查和检验、疾病和诊断、治疗、身体部位,数据均采用 BIOES 方式对字进行标注,所有实验只需要预先训练一个词向量 (因为

是对字进行标注,所有此时的词向量用字向量替换),除此之外均不需要对数据进行预处理和外部的知识信息。

4.1 医疗领域的命名实体识别实验

CCKS2017 共发布了 400 组已标注的中文电子医疗文本,其中 300 组数据作为训练数据,100 组作为测试数据。除此之外 CCKS 还发布了 2605 组未标注的数据集。电子病历每组有 4 部分组成,包括一般项目,病史特征,诊疗过程和出院情况。其中训练集总共包含 29866 个句子,测试集包含 9493 个句子。预训练的字向量是利用开源工具 Glove 在 2605

组未标注数据上训练获得。

实验部分分别采用了两种方法,在单一模型和混合模型上做了对比实验。一是利用 300 组已标注数据训练出一个模型,然后利用该模型对测试集进行标注,实验结果即表 2 中 supervised learning 的结果。二是先利用第一种 300 组标注训练数据训练出的模型对 2605 组未标注数据进行标注,然后将其和原先的 300 组数据混合到一起,利用所有的 2905 组数据重新训练出一个模型,最后利用新的模型对测试集进行标注,结果如表 2 中 semi-supervised learning 结果所示。

表 2 医疗领域数据不同模型下的实验结果
Tab. 2 Different Model Result in Medical Domain

	模型	精度 / %	召回率 / %	F1 / %
有监督学习	CRF	87.77	86.83	87.29
	BLSTM	89.53	88.67	89.09
	CNN - BLSTM - CRF	90.68	89.67	90.14
半监督学习	CRF	88.92	88.10	88.51
	BLSTM	90.21	89.53	89.86
	CNN - BLSTM - CRF	91.25	90.69	90.97

所有实验均分别在单一的 CRF 模型、BLSTM 模型和 CNN-BLSTM-CRF 模型上进行。

由实验结果可以看到,本文中的混合模型的性能要明显优于单一模型,两种实验方法中,混合模型相对于 CRF、BLSTM 分别提高了 2.85%、1.05% 和 2.46%、1.11%。并且采用半监督方法后测试集的 F1 值提高了 0.83%,因此增加标注数据的数量有助于提高模型的性能。

4.2 新闻领域的命名实体识别

新闻领域采用的是 1998 年的人民日报数据,总共包含 107976 个已标注的句子。由于没有测试数据集,在此部分采用了 4 折实验的方法,预训练的字向量是用开源工具 Glove 在中文维基百科数据上训练获得。所有实验同样是在单一模型和混合模型上进行。实验结果如表 3、表 4、表 5 所示。

表 3 CRF 模型上的 4 折实验
Tab. 3 4-Fold Experiment of CRF Model

折数	精度 / %	召回率 / %	F1 / %
f1	91.22	89.20	90.20
f2	90.92	89.14	90.02
f3	90.89	89.32	90.10
f4	91.03	89.21	90.11
平均	91.01	89.21	90.11

表 4 BLSTM 模型上的 4 折实验
Tab. 4 4-Fold Experiment of BLSTM Model

折数	精度 / %	召回率 / %	F1 / %
f1	91.37	90.23	90.79
f2	91.18	90.13	90.65
f3	91.02	90.07	90.54
f4	91.41	90.38	90.89
平均	91.25	90.20	90.71

表 5 CNN-BLSTM-CRF 模型上的 4 折实验
Tab. 5 4-Fold Experiment of CNN-BLSTM-CRF Model

折数	精度 / %	召回率 / %	F1 / %
f1	92.49	92.14	92.31
f2	92.43	92.13	92.28
f3	92.15	92.20	92.05
f4	92.31	91.95	92.13
平均	92.34	92.04	92.19

实验结果表明混合模型在新闻领域的实体识别任务中也高于单一的模型,相对 CRF、BLSTM 模型分别提高了 2.08%、1.48%。

4.3 实验结果分析

通过在两组不同领域的数据集上进行实验,本

文中混合模型在各组实验中均取得了高于单一模型的预测效果。且在新闻领域和医疗领域中,实体类别区别非常大,使用本文中的模型,不需要任何的数据预处理和引入外部知识信息。验证了该模型可以很好的应用到不同领域的命名实体识别任务中,模型的通用性较好。

5 结 论

针对命名实体识别在不同领域中的类型千差万别,传统的命名实体识别方法中需要大量人工选择的特征和引入外部的知识信息,本文提出了一种结合 CNN、BLSTM、CRF 的模型,可以直接应用到不同类型的实体标注中去。通过在新闻领域和医疗领域数据集上的多组对比实验,该系统均取得了高于传统单一模型的结果。在接下来的工作中,可以对模型进一步改进,探索多任务学习方法,针对不同的领域结合更多有用的相关信息。

参 考 文 献:

- [1] DAVID N, SATOSHI S. A Survey of Named Entity Recognition and Classification [J]. Logistical Investigations, 2007, 30(1): 3.
- [2] 刘浏, 王东波. 命名实体识别研究综述 [J]. 情报学报, 2018, 37(3): 329.
LIU Liu, WANG Dongbo. A Survey of Named Entity Recognition [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 329.
- [3] 王路路, 艾山, 吾买尔. 基于 CRF 和半监督学习的维吾尔文命名实体识别 [J]. 中文信息学报, 2018, 32(11): 16.
WANG Lulu, AI Shan, WU Maier. A Semi-supervised Approach to Uyghur Named Entity Recognition Based on CRF [J]. Journal of Chinese Information Processing, 2018, 32(11): 16.
- [4] 徐广义, 严馨, 余正涛, 等. 融合跨语言特征的柬埔寨语命名实体识别方法 [J]. 云南大学学报 (自然科学版), 2018, 40(5): 865.
XU Guangyi, YAN Xin, YU Zhengtao, et al. A Khmer Named Entity Recognition Method Incorporating Cross-lingual Features [J]. Journal of Yunnan University (Natural Sciences Edition), 2018, 40(5): 865.
- [5] 朱颖东, 杨立志, 丁温雪, 等. 基于主题标签和 CRF 的中文微博命名实体识别 [J]. 华中师范大学学报 (自然科学版), 2018, 52(3): 316.
ZHU Haodong, YANG Lizhi, DING Wenxue, et al. Named Entity Recognition of Chinese Microblog based on Theme tag and CRF [J]. Journal of Central China Normal University (Natural Sciences), 2018, 52(3): 316.
- [6] RITTER A, CLARK S, ETZIONI O. Named Entity Recognition in Tweets: an Experimental Study [C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 1524.
- [7] LIU X, ZHANG S, WEI F, et al. Recognizing Named Entities in Tweets [C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2011: 359.
- [8] LI C, WENG J, HE Q, et al. Twiner: Named Entity Recognition in Targeted Twitter Stream [C] // Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, 2012: 721.
- [9] 张祝玉, 任飞亮, 朱靖波. 基于条件随机场的中文命名实体识别特征比较研究 [C] // 第四届全国信息检索与内容安全学术会议, 2008: 8.
ZHANG Zhuayu, REN Feiliang, ZHU Jingbo. A Comparative Study of Chinese Named Entity Recognition Features based on Conditional Random Fields [C] // NCIRC, 2008: 8.
- [10] WU Y, JIANG M, LEI J, et al. Named Entity Recognition in Chinese Text Using Deep Neural Network [J]. Studies in Health Technology and Information, 2015, 15(1): 624.
- [11] 刘树杰, 董力, 张家俊. 深度学习在自然语言处理中的应用 [J]. 中国计算机学会通讯, 2015, 18(2): 11.
LIU Shujie, DONG Li, ZHANG Jiajun. The Application of Deep Learning in Natural Language Processing [J]. Communications of the CCF, 2015, 18(2): 11.
- [12] WU Y, JIANG M, LEI J, et al. Named Entity Recognition in Chinese Clinical Text using Deep Neural Network [J]. Studies in Health Technology and Informatics, 2015, 216: 624.
- [13] CICERO D, VICTOR G, RJ N, et al. Boosting Named Entity Recognition with Neural Character Embeddings [C] // In Proceedings of NEWS 2015 The Fifth Named Entities Workshop, 2015: 25.
- [14] WU Y, JIANG M, LEI J, et al. Named Entity Recognition in Chinese Text Using Deep Neural Network [J]. Studies in Health Technology and Information, 2015, 18(1): 462.
- [15] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging [J]. Computer Science, 2015, 4(1): 1508.
- [16] ANDO R, ZHANG T. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data [J]. The Journal of Machine Learning Research, 2005(6): 1817.
- [17] 陈锋, 翟羽佳, 王芳. 基于条件随机场的学术期刊中理论的自动识别方法 [J]. 图书情报工作, 2016, 60(2): 122.
CHEN Feng, ZHAI Yujia, WANG Fang. Automatic Theory Recognition in Academic Journals Based on CRF [J]. Library and Information Service, 2016, 60(2): 122.
- [18] CHIU J, NICHOLS E. Named Entity Recognition with Bidirectional LSTM-CNNs [J]. Transactions of the Association for Computational Linguistics, 2016(4): 357.
- [19] 曲春燕, 关毅, 杨锦锋, 等. 中文电子病历命名实体标注语料库构建 [J]. 高技术通讯, 2015(2): 143.
QU Chunyan, GUAN Yi, YANG Jinfeng, et al. The Construction of Annotated Corpora of Named Entities for Chinese Electronic Medical Records [J]. Chinese High Technology Letters, 2015(2): 143.
- [20] PENG N, DREDZE M. Named Entity Recognition for Chinese social Media with Jointly Trained Embeddings [C] // Proceedings of EMNLP, 2015: 548.

(编辑: 温泽宇)