

基于 BLSTM 的命名实体识别方法

冯艳红^{1,2} 于红^{1,2} 孙庚^{1,2} 孙娟娟¹

(大连海洋大学信息工程学院 大连 116023)¹

(大连海洋大学辽宁省海洋信息技术重点实验室 大连 116023)²

摘要 传统的命名实体识别方法直接依靠大量的人工特征和专门的领域知识,解决了监督学习语料不足的问题,但设计人工特征和获取领域知识的代价昂贵。针对该问题,提出一种基于 BLSTM(Bidirectional Long Short-Term Memory)的神经网络结构的命名实体识别方法。该方法不再直接依赖于人工特征和领域知识,而是利用基于上下文的词向量和基于字的词向量,前者表达命名实体的上下文信息,后者表达构成命名实体的前缀、后缀和领域信息;同时,利用标注序列中标签之间的相关性对 BLSTM 的代价函数进行约束,并将领域知识嵌入模型的代价函数中,进一步增强模型的识别能力。实验表明,所提方法的识别效果优于传统方法。

关键词 BLSTM,命名实体,词向量,代价函数

中图分类号 TP183 文献标识码 A DOI 10.11896/j.issn.1002-137X.2018.02.045

Named Entity Recognition Method Based on BLSTM

FENG Yan-hong^{1,2} YU Hong^{1,2} SUN Geng^{1,2} SUN Juan-juan¹

(College of Information Engineering, Dalian Ocean University, Dalian 116023, China)¹

(Key Laboratory of Marine Information Technology of Liaoning Province, Dalian Ocean University, Dalian 116023, China)²

Abstract Traditional named entity recognition methods directly rely on plenty of hand-crafted features and special domain knowledge, and have resolved the problem that there are few supervised learning corpora which are available. But the costs of developing hand-crafted features and obtaining domain knowledge are expensive. To solve this problem, a neural network model based on BLSTM(Bidirectional Long Short-Term Memory) was proposed. This method does not directly use hand-crafted features and domain knowledge any more, but utilizes the word embedding based on context and word embedding based on characters. The former expresses the information about context of named entities, and the latter expresses the information about prefix, postfix and domain knowledge which make up the named entities. Simultaneously, it constrains the cost function of BLSTM by using the dependency between the labels in tagged sequence, and integrates the domain knowledge into the cost function, furtherly improving the recognition ability of the model. The experiments show that the recognition effect of the method in this paper is superior to traditional methods.

Keywords BLSTM, Named entity, Word embedding, Cost function

1 引言

命名实体识别是自然语言处理领域的关键任务,对信息检索、自动问答、机器翻译和知识库构建等研究和应用有重要意义。目前,命名实体识别主要利用有监督的机器学习方法,然而,用于监督学习的训练数据通常由人工标注,耗时耗力,导致大规模标注数据稀缺;同时,样本数据的约束不强,导致很难在小样本数据上产生泛化的命名实体识别结果。学者们针对这类问题,提出了利用人工特征的机器学习方法识别文本中的命名实体。孙丽萍等人利用条件随机场(Conditional Random Field, CRF)对企业简称进行识别^[1]; Duan 等人利用 CRF 对人民日报语料中的人名、地名和机构名进行识别^[2];

黄德根等人利用 CRF 和支持向量机(Support Vector Machine, SVM)构建了双层模型对机构名称进行识别^[3]。在这类方法中,学者们针对不同的识别问题提取了大量不同的人工特征,如词语的前后缀特征、语法特征、构成命名实体的字符特征等,这些特征针对具体问题设计,能精确地获取对问题有价值的信息,弥补了标注语料稀缺的问题,取得了很好的识别效果。除了利用人工特征外,也有学者利用了领域知识,如领域词典。冯艳红等人利用 CRF 对渔业领域的命名实体进行识别^[4],用到了渔业领域的词典;冯蕴天等人利用 CRF 对军事领域的命名实体进行识别^[5],用到了军事领域的词典。这种利用人工特征和领域知识的方法取得了很好的识别效果,但人工特征的代价昂贵,并且不同领域选取的特征不尽相

到稿日期:2016-11-02 返修日期:2017-03-22 本文受大连市科技计划项目海洋渔业大数据管理与集成关键技术研究(2015A11GX022)资助。
冯艳红(1980-),女,硕士,讲师,CCF 会员,主要研究方向为自然语言处理、信息检索;于红(1968-),女,博士,教授,CCF 会员,主要研究方向为数据挖掘、信息检索, E-mail: yuhong@dloou.edu.cn(通信作者);孙庚(1979-),男,硕士,副教授,主要研究方向为嵌入式系统;孙娟娟(1993-),女,硕士生,主要研究方向为数据挖掘、信息检索。

同,不同领域的词典不通用,导致识别方法不通用,泛化能力差,识别工作量大。近几年,利用基于神经网络的深度学习获取特征的方法在图像、语音和自然语言处理领域取得了很好的效果。在自然语言处理领域,用词向量表达文本中词语是将深度学习算法引入自然语言处理领域的一个核心技术^[6]。词向量是通过训练神经网络语言模型得到的一种分布表示特征^[7-8],即用一个连续的实数向量表达文本中的一个词语的特征。有学者提出将词向量作为词语的特征,替代机器学习方法中的人工特征,在文献[4]和文献[9]中,利用词向量作为词语的特征,分别对渔业领域和生物领域的命名实体进行识别。这类方法解决了人工提取特征所需代价较高的问题,同时也提高了识别效果。为了进一步减小对人或其他系统的依赖,有学者提出用神经网络模型解决命名实体识别问题,替代传统的CRF和SVM等机器学习模型。文献[10]和文献[11]利用多层前馈神经网络对人民日报语料中的命名实体进行识别,相比于CRF方法,其地名和机构名的F值分别提升了3.76%和2.35%。文献[12]利用卷积神经网络对Reuters语料中的英文命名实体进行识别,该方法在不加入人工特征时,F值略低于利用人工特征的方法;在加入人工特征和词典后,F值高于传统的利用人工特征的方法。这类方法通常不再依赖于人工特征或领域知识,实现了端到端的命名实体识别模式,减少了人工特征提取所需的代价,所以本文利用神经网络的方法研究命名实体识别问题。

命名实体识别是典型的序列标注问题,而循环神经网络(Recurrent Neural Network,RNN)是一种很有效地解决序列标注问题的神经网络模型,能够有效地利用数据的序列信息,具有一定的记忆功能。但RNN无法很好地处理长距离依赖问题,并且训练算法存在梯度消失或爆炸问题。文献[13]提出了一种利用门限机制对历史信息进行过滤的长短时记忆(Long Short-Term Memory,LSTM)模型,有效地解决了RNN中存在的问题。针对本文要解决的问题和LSTM模型的优点,提出一种中文命名实体识别方法,该方法基于3点考虑:1)文本是否被识别为命名实体与其上下文有关,也与构成命名实体的每个字及字序有关。所以本文提出的方法利用了无标注语料中的上文信息和下文信息,解决了大规模标注语料稀缺的问题;利用基于字的分布表示来表达词语的字序特征,解决了传统识别方法过于依赖人工特征的问题。2)受文献[12]启发,考虑标注序列中标签间的相关性,对本文提出的模型的代价函数进行约束,在小的训练数据上尽可能挖掘有价值的信息,以提高命名实体识别的效果。3)学者们已经证明,传统识别方法中的人工特征和领域知识对命名实体的识别效果的提升有重要影响,但设计人工特征和获取领域知识的代价昂贵,所以本文利用了少量易获取的领域知识,将其融入模型的代价函数和基于字的词向量中,以进一步提高命名实体识别的效果。

本文第2节阐述基于BLSTM的命名实体识别模型;第3节阐述基于上下文和字序的词向量训练模型;第4节通过实验验证了所提方法的识别效果;最后总结全文。

2 基于BLSTM的命名实体识别模型

2.1 BLSTM结构

给定输入序列 (x_1, x_2, \dots, x_n) ,RNN模型返回一个关于

输入序列的表示序列 (h_1, h_2, \dots, h_n) 。该模型能动态地捕获序列数据的信息,并且对信息具有记忆能力,但在算法实现上,容易产生梯度消失或爆炸问题。文献[13]中提出的LSTM模型引入了记忆单元和门限机制,实现了对长距离信息的有效利用,并解决了梯度消失问题。有学者对LSTM模型的记忆单元和门限机制做了改进以提高效率^[14],本文采用文献[14]提出的改进的门限机制, t 时刻,给定输入 x_t ,LSTM的隐藏层的输出表示 h_t 的具体计算过程如式(1)~式(5)所示:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad (5)$$

其中, W 表示连接两层的权重矩阵(如 W_{xi} 表示输入层到隐藏层的输入门的权重矩阵), b 表示偏置向量(如 b_i 表示隐藏层的输入门的偏置向量), c 表示记忆单元的状态, σ 和 \tanh 表示两种不同的神经元激活函数, i 、 f 和 o 分别表示输入门、遗忘门和输出门。这种门限机制能够对记忆单元的信息进行有效过滤和记忆,从而解决RNN存在的问题。LSTM只捕获了文本的上文信息,对于命名实体识别任务,下文信息也具有重要的参考价值。文献[15]提出的双向RNN(Bidirectional RNN,BRNN)模型同时考虑了历史信息(即上文信息)和未来自信息(即下文信息)。文献[16]用LSTM记忆单元替代BRNN中的隐藏层神经元,构建了BLSTM模型,该模型有效地利用了序列数据的上文信息和下文信息,其结构如图1所示。

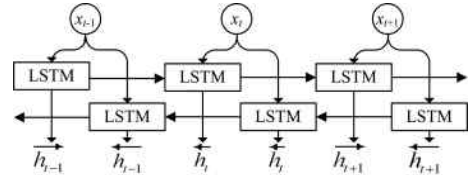


图1 BLSTM模型结构

Fig. 1 Structure of BLSTM model

图1中,圆圈中的 x_t 表示模型 t 时刻的输入;方框中的LSTM为式(1)~式(5)描述的模型; \vec{h}_t 为前向LSTM在 t 时刻的输出表示, \overleftarrow{h}_t 为反向LSTM在 t 时刻的输出表示,则BLSTM在 t 时刻的输出表示定义为 $h_t = [\vec{h}_t; \overleftarrow{h}_t]$,即直接拼接 \vec{h}_t 与 \overleftarrow{h}_t 。这种表示同时包含了上文信息和下文信息,适用于标签种类较多的命名实体识别任务。

2.2 命名实体识别模型

利用BLSTM在 t 时刻的输出表示 h_t 对 t 时刻的标注结果直接预测,对于标签之间独立的任务(如词性标注任务)有很好的识别效果,但命名实体识别任务的标签之间并不独立,而是具有较强的依赖关系(如B-PER标签后面可能是I-PER标签和E-PER标签,但不可能是I-LOC标签,详见本文4.1节),所以本文的标注决策同时考虑两个条件:输出表示 h_t 对标注的预测和标签间的依赖关系对标注的预测。建立命名实体识别模型,模型的结构如图2所示,该模型记为BLSTM_NER(BLSTM Named Entity Recognition)。

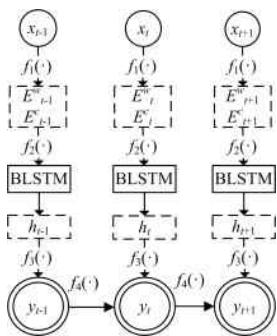


图 2 命名实体识别模型结构

Fig. 2 Structure of named entity recognition model

图 2 所示模型的输入为词语序列,记为 X 。圆圈中的 x_t 表示模型在 t 时刻的输入,模型的输入记为 $X = [x]_T^T$ 。 $f(\cdot)$ 为变换函数,表示词语到词向量的变换函数,本文定义为查找词向量表(LT)变换,记为 $f_1(x_t) = LT(x_t)$ 。关于词向量的内容详见本文第 3 节。虚线方框中的内容表示输入数据的分布表示,包括 E^w , E^c 和 h ,则 $LT(x_t) = [E_t^w; E_t^c]$ 。本文定义 f_2 变换为拼接操作(Concat),记为 $f_2([E_t^w; E_t^c]) = Concat([E_t^w; E_t^c]) = E_t$ 。实线方框为图 1 所示的 BLSTM 模型,则 $h_t = BLSTM(E_t)$,得到输入 x_t 的分布表示 h_t 。 f_3 变换为分布表示 h_t 到标注结果的决策概率,本文将定义为具有表达概率功能的 *softmax* 函数,记为 $f_3(h_t) = softmax(h_t)$ 。双圆圈中的 y_t 表示模型的标注结果,则将输入 x_t 时预测结果为 y_t 的概率记为 $P_{x_t, y_t} = softmax(h_t)$ 。 f_4 变换为标签 y 之间的变换, y_t 到 y_{t+1} 之间的依赖关系以标签转移概率形式表达,记为 $A_{y_t, y_{t+1}} = f_4(y_t, y_{t+1})$ 。

由该模型可知,当给定词语序列 $X = [x]_T^T$,由模型得到预测序列 $y = [y]_T^T$,则预测结果正确的可能性量化定义如式(6)所示:

$$S(X, y) = \sum_{t=0}^T A_{y_t, y_{t+1}} + \sum_{t=1}^T P_{x_t, y_t} \quad (6)$$

由式(6)可以看出,本文在评价标注的好坏时,除了使用 BLSTM 的输出表示得到的概率 P 外,还考虑了标签之间的约束关系 A 。对标注决策进行约束,当概率 P 和 A 的值较大时,正确的可能性得分 $S(X, y)$ 也会变大。式(6)中的未知参数包括词向量 E 、BLSTM 中的权重矩阵 W 和偏置向量 b 、标注转移概率 A ,记 $\theta = \{E, W, b, A\}$ 。根据本文引言部分提出的 3 点考虑中的第 3 点,将领域知识融入到代价函数中,具体方法为:将一个小规模的命名实体词典作为领域知识,当输入词语为词典中的命名实体时,正确的可能性得分 $S(X, y)$ 应该变大,根据文献[17]中阐述的加入正则项的做法,本文在式(6)的基础上加入正则项 $R(\theta)$,如式(7)所示:

$$S(X, y) = \sum_{t=0}^T A_{y_t, y_{t+1}} + \sum_{t=1}^T P_{x_t, y_t} + R(\theta) \quad (7)$$

本文取 $R(\theta)$ 为参数 θ 的二范数,即 $R(\theta) = \|\theta\|_2$ 。

所有可能的标注序列记为 Y ,则标注为 y 的概率如式(8)所示:

$$p(y/X) = \frac{e^{S(X, y)}}{\sum_{y' \in Y} e^{S(X, y')}} \quad (8)$$

在所有可能的标注序列上执行 *softmax* 函数,即可得到标注为序列 y 的概率 $p(y/X)$ 。

模型的目标是使得正确标注的序列的概率最大,所以采

用对数最大似然估计得到该模型的代价函数,如式(9)所示:

$$\begin{aligned} \log(p(y/X)) &= S(X, y) - \log\left(\sum_{y' \in Y} e^{S(X, y')}\right) \\ &= S(X, y) - \underset{y' \in Y}{\text{logadd}} S(X, y') \end{aligned} \quad (9)$$

利用随机梯度下降学习算法训练参数 θ ,求得参数 θ 后,用 Viterbi 算法求得所有序列上打分最高的序列,并将其作为最终的命名实体识别的标注结果,如式(10)所示:

$$y^* = \arg \max_{y' \in Y} S(X, y') \quad (10)$$

3 词向量

命名实体识别模型中的词向量 E 通常具有较高的维度,导致模型具有大量的参数,而实际中,有标注的监督训练语料通常不足,所以难以学习到准确的参数。本文受文献[8, 12, 18-19]启发,采取预训练词向量的方法,解决训练数据不足的问题。主要从两方面对词向量进行预训练:1)词语是否为命名实体与其上下文有关,所以利用语言模型[8, 18]在大规模无标注语料上训练词向量;2)词语是否为命名实体与构成命名实体的字及字序有关,即与其前缀和后缀有关,同时,领域知识对命名实体识别的效果有较大的提升,所以利用 BLSTM 模型在小规模的标注语料(不使用数据的标签)和小规模命名实体词典上分别训练两种词向量。3.1 节阐述利用上下文词语得到的词向量,3.2 节阐述利用字序和领域词典得到的字符级词向量。

3.1 基于上下文词语的词向量

虽然有监督的训练语料不足,但随着互联网的发展,大规模的未标注的无监督数据容易获取,利用语言模型在大规模的无监督的训练语料上学习词向量是一个解决问题的可行方案。文献[12]在解决序列标注问题中已经证明,先用大规模的无标注语料训练得到词向量,再将其作为标注模型的词向量的初始值,比直接使用随机值作为标注模型的初始值的性能有提升。Word2Vec 是语言模型的一种简单且有效的实现[8, 18],有两种实现模型:Skip-gram 和 CBOW。Skip-gram 是由当前词语预测上下文词语,而 CBOW 是由上下文词语预测当前词语。文献[20]对两种模型进行了比较,当语料规模在百兆级别时,CBOW 模型体现出较大的优势和较好的性能。本文选用的搜狗新闻语料约有 250MB 词语,所以利用 CBOW 模型在该语料上训练基于上下文词语的词向量,得到的词向量记为 E^w 。

3.2 字符级词向量

在中文语言形态学中,词语具有一定的内部结构和构词规则,为了让词向量能更准确地表达词语,本文根据文献[19]提出了一种利用 BLSTM 模型训练基于字符的词向量的方法,充分利用语料的语言学特征,进一步解决监督训练语料不足的问题。字符级的词向量训练模型结构如图 3 所示,该模型记为 BLSTM_E,得到的词向量记为 E^c 。

图 3 中 f_5 变换是将输入 x 变换为字符序列 c_i , c_i 是构成 x 的第 i 个字符,记为 $f_5(x) = c_0 \cdots c_{i-1} c_i c_{i+1} \cdots$ 。可以看出,该模型利用了词语的前缀、后缀及字序信息。对于有特定前缀(姓氏)的人名、复合地名(例如,安徽省 合肥市 长江路)和复合机构名(例如,上海 交通大学;国防部 第五 研究院 火箭 总体 设计部),这类命名实体通常有特定的前缀或后缀,对词语是否被识别为命名实体有重要的影响,所以该模型学习到

的这种形态学特征有助于提高命名实体识别的效果。同样,若将领域词典这种先验知识作为该模型的输入,得到的词向量会更准确地表达命名实体的语义和结构信息,进而提升命名实体的识别效果。所以若图3所示的模型的输入词语 x 来自小规模标注语料(不使用数据的标签),则得到的词向量 E^c 为词语的前缀信息和后缀信息的分布表示;若图3所示的模型的输入词语 x 来自小规模命名实体词典,则得到的词向量 E^c 为词语的前、后缀信息和领域知识的分布表示。

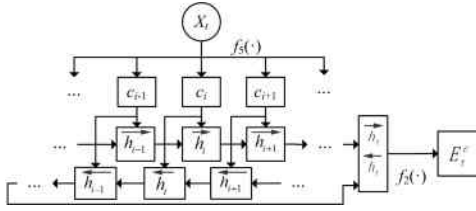


图3 训练字符级词向量模型结构

Fig. 3 Structure of training character level word embedding model

4 实验结果及分析

为验证本文所提命名实体识别方法的识别效果,将该方法与文献[2]中的CRF方法和文献[11]中的SVM方法、DNN(Deep Neural Network)方法进行比较。为体现文献中识别方法的实验条件与本文中识别方法的实验条件的异同,4.1节阐述了语料及标注模式,给出本文、文献[2]和文献[11]中的数据集、标注模式和预处理步骤等内容;由于本文提出的识别方法需要大量的参数,因此在4.2节阐明了本文识别方法的实验设置,同时给出了文献[2]和文献[11]中识别方法的实验设置,供实验对比使用;为综合评价本文所提识别方法的效果,在4.3节阐明了评价指标体系;最后在4.4节给出实验结果和分析。

4.1 语料及标注模式

本实验所用数据包括4个中文文本语料集。第一个语料集为大规模无标注语料,本文选择搜狗全网新闻语料集(SogouCA)^[21],记为DataSet1,用于训练基于上下文的词向量;第二个语料集为人民日报1998年1月份的标注语料,记为DataSet2;第三个语料集为哈尔滨工业大学命名实体识别标注语料,记为DataSet3;第四个语料集为命名实体词典,本文选择搜狗输入法词库中的部分数据^[22],包括常见的中国人名、中国地名、国家机关组织机构名,拆分为字符序列数据,记为DataSet5。第二个和第三个语料集用于训练和测试BLSTM_NER模型。为训练BLSTM_E^c模型中基于字符的词向量,本文将DataSet2和DataSet3中的标签进行删除,并拆分为字符序列数据,得到的数据集记为DataSet4,作为BLSTM_E^c模型的训练语料。DataSet5用于训练BLSTM_E^c模型中融入领域知识的词向量。对于文献[2]和文献[11]中实验所用数据,仅包括本文的DataSet2数据集,所以4.4节的实验对比仅限于在该数据集上的对比。

本文识别中文文本中的人名、地名和机构名3种命名实体,5个数据集中的词语总数、3种命名实体在DataSet2,

DataSet3,DataSet4和DataSet5中的数量,以及DataSet4和DataSet5中的字符总数的统计结果如表1所列,其中DataSet2中去除了每行数据前面的日期。

表1 数据集统计信息

Table 1 Statistical information about data sets

数据集	词语总数 /MB	人名数量 /kB	地名数量 /kB	机构名数量 /kB	字符总数 /MB
DataSet1	250	—	—	—	—
DataSet2	1.120	19.98	22.43	11.15	—
DataSet3	1.620	25.36	9.22	8.49	—
DataSet4	2.740	45.34	31.65	19.64	4.36
DataSet5	0.014	5.80	8.10	0.45	0.23

为验证本文所提方法的效果,本文在DataSet2和DataSet3上采用五折交叉验证实验,即DataSet2的训练集有0.9MB词语,其余0.22MB为测试集;DataSet3的训练集有1.3MB词语,其余0.32MB为测试集。文献[2]中DataSet2上的训练集和测试集的划分为7:3,文献[11]中DataSet2上训练集和测试集的划分方法与本文相同。

监督训练数据常用的标注模式有BIO,BIEO和BIESO等,B表示实体开始,I表示实体中间,E表示实体结束,S表示单个实体,O表示不是实体。由于BIESO标注模式的标签种类最多,理论上,其表达能力最强,因此本文的监督训练数据的标注模式选择BIESO模式。人名记为PER,所以人名的开始记为B-PER,地名记为LOC,地名的开始记为B-LOC,机构名记为ORG,机构名的开始记为B-ORG。易知,标签一共有13种,即本文采用13标签标注模式。由于DataSet2和DataSet3的原始语料的标注模式不一致,因此实验中应首先统一标注模式。标注时,如果有嵌套现象,取嵌套的最外层的命名实体的标注标签为最终标注结果;每个词只标注一种标签,不重复标注。文献[2]采用了3种标注模式:3标签标注模式、4标签标注模式和6标签标注模式;文献[11]中采用的是7标签标注模式。

由于数据集DataSet2是公开数据集,并且已经给出分词结果和词性标注,在文献[11]和本文中,除了处理标注模式外,均未对该数据集做其他预处理,而是直接用于实验;但由于文献[2]中CRF方法需要大量的人工特征(包括字符、词性、前缀、后缀及前后缀组合),因此做了用于提取特征的预处理。由于本文实验使用了DataSet1至DataSet5共计5个数据集,并对不同的数据集做了不同的预处理,为让读者了解本文实验中的预处理过程,给出预处理的源代码¹⁾。

4.2 实验设置

基于上下文词语的词向量 E^w 由CBOW模型在DataSet1上训练得到。设置CBOW模型的参数:隐藏层的神经元数量为100,即词向量的维度为100;上下文窗口大小为5;初始学习率为0.015,该值随训练误差的降低而减小;经过统计,DataSet1的词表大小约为752kB,CBOW模型的输出层神经元数量是词表的大小,过大的输出导致训练效率低下,本文采用负采样的方法提高训练效率,采样频率设置为0.1。

字符级的词向量 E^c 由BLSTM_E^c模型分别在DataSet4和DataSet5上训练得到。DataSet4的规模与PTB(Penn Treebank)数据集^[23]的规模在数量级上相当,参照文献[23]

¹⁾ https://github.com/fengpp/blstm_ner

在 PTB 数据集上训练语言模型的参数设置,设置 BLSTM_E^o 模型在 DataSet4 上训练的参数:隐藏层为 1 层,前向和反向 LSTM 的隐藏层的神经元数量都为 25,所以模型的输出表示 E^o 的大小为 50;模型的权重参数随机初始化,在 $[-0.1, 0.1]$ 范围内均匀分布;初始学习率设置为 1.0,每训练 1 代学习率降低 0.5 倍;梯度范数设置为 5;batch 大小设置为 200;总共训练 10 代。数据集 DataSet5 的规模较小,与 DataSet4 的规模相差一个数量级,所以本文的 BLSTM_E^o 模型在 DataSet5 上训练的参数设置:隐藏层仍为 1 层,前向和反向 LSTM 的隐藏层的神经元数量仍然都为 25,所以模型的输出表示 E^o 大小仍然为 50,所以模型的参数个数与 DataSet4 上的模型的参数个数相同;模型的权重参数随机初始化,在 $[-0.1, 0.1]$ 范围内均匀分布;由于模型的参数个数相同,而训练数据规模变小,因此降低初始学习率,设置其为 0.1,每训练 1 代学习率降低 0.5 倍;梯度范数设置为 5;同时降低 batch 大小,设置其为 50;为避免过拟合,减少训练的代数,设置其为 5。

本文提出的 BLSTM_NER 模型分别在 DataSet2 和 DataSet3 上实验,后者的人名数量多于前者的人名数量,后者的地名和机构名的数量略少于前者的地名和机构名的数量,且在命名实体的总数上后者也略少于前者。综合来看,这两个数据集的训练数据规模和命名实体的数量都很相近,所以采用相同的参数设置,具体为:词表大小设置为 60kB,未登录词的词向量的分量均取值为 0.5;神经网络的输入层的大小为 150,隐藏层为 1 层,隐藏层神经元数量为 300,输出层大小为 13;学习率设置为 0.001;梯度范数设置为 5;batch 大小设置为 400;总共训练 10 代。为提升模型的泛化性能,对 BLSTM_NER 模型的输入到隐藏层的权重采用了 Dropout 技术。

文献[2]中的 CRF 方法由开源项目 CRF++5.4 实现,由于该方法采用的是传统的机器学习方法识别命名实体,因此参数设置与本文的方法不具备可比性,此处略去其参数设置。文献[11]中的 SVM 方法由开源工具 LIBLINEAR 实现,输入特征为由文献[11]的 DNN 模型得到的分布式特征。文献[11]中的 DNN 方法由自动编码器(AutoEncoder, AE)实现,网络顶层的分类器分别由 Logistic 和 Softmax 实现,达到最优识别结果时的参数设置为:顶层分类器由 Logistic 实现,上下文窗口为 5,隐藏层神经元数量为 100,学习率为 0.2,总共训练 50 代后收敛。相较于本文的方法,DNN 方法所采用的神经网络规模略小,但训练代数多于本文的识别方法。由于 DNN 方法所采用的网络结构和本文提出的基于 BLSTM 的方法所采用的网络结构不同,因此网络规模和训练代数的不同并不能完全决定二者的优劣,主要以识别效果作为优劣评价指标。

4.3 评价指标

本文将命名实体识别问题建模为多分类问题,共计 13 类,在评价时,将其转换为二分类问题,具体转换方法阐述如下。

根据 4.1 节阐述的标注模式,将人名、地名和机构名分为 3 个独立的问题,每个问题为一个五分类问题。以人名为例,分为 B-PER, I-PER, E-PER, S-PER 和 O 5 类,将识别出的人名的开始、中间和结束合并为一个实体,完整地识别一个命名

实体为正确识别,记为二分类的正类;而只识别对实体的开始、中间或结束不算识别正确,记为二分类的负类,识别结果为 O 也记为负类。地名和机构名的转换方法与此相同。

采用二分类问题常用的评价指标体系对本文研究的方法的识别效果进行评价,指标体系中的具体指标项阐述如下。

将样本分为正类(Positive)和负类(Negative)。样本是正类并且被模型预测为正类的数目(True Positive)记为 TP;样本是负类并且被模型预测为正类的数目(False Positive)记为 FP;样本是负类并且被模型预测为负类的数目(True Negative)记为 TN;样本是正类并且被模型预测为负类的数目(False Negative)记为 FN。

精确率(Precision, P)、召回率(Recall, R)、F 测度值(F-score, F)、敏感度(Sensitivity, Sent)、特异性(Specificity, Spec)、1-特异性(1-Spec)、准确率(Accuracy, Acc)指标定义如式(11)~式(17)所示:

$$P = TP / (TP + FP) \quad (11)$$

$$R = TP / (TP + FN) \quad (12)$$

$$F = 2 * P * R / (P + R) = 2 * TP / (2 * TP + FP + FN) \quad (13)$$

$$Sent = TP / (TP + FN) = R \quad (14)$$

$$Spec = TN / (TN + FP) \quad (15)$$

$$1-Spec = FP / (TN + FP) \quad (16)$$

$$Acc = (TP + TN) / (TP + TN + FP + FN) \quad (17)$$

其中,式(11)的 P 表达识别出正确的正类占识别出的正类的总数的比例;式(12)的 R 表达识别出正确的正类占样本中正类的总数的比例;式(13)的 F 为 P 和 R 的调和平均数,表达了二者的综合效果,以上 3 个指标主要评价正类的识别效果。式(14)的 Sent 与式(12)的 R 等价;式(15)的 Spec 表达识别出正确的负类占样本中负类的总数的比例。由 Sent 和式(16)的 1-Spec 构成的 ROC(Receiver Operating Characteristic)曲线是一种有效地评价模型的方法,同时评价了正类和负类的识别效果。对于样本的正负类不均衡的问题,该评价方法具有较高的参考价值,本文的数据集 DataSet2 和 DataSet3 属于正负样本不均衡的情况。为进一步验证模型的效果,给出 AUC(Area Under the Curve)指标,该指标为 ROC 曲线与坐标轴构成的区域的面积,范围为 $[0, 1]$,该值越大说明模型的效果越好。

综上,本文给出不同数据集上不同模型的最优 P, R, F, Acc 值、同一数据集上不同模型的 ROC 曲线和相应的 AUC 的值,作为本文所提方法的识别效果的评价。

4.4 实验结果及分析

为验证本文提出的命名实体识别方法的识别效果,实验分两组进行。在第一组实验中,侧重对比基于神经网络的模型和传统的机器学习模型的识别效果,令 BLSTM_NER 模型的输入词向量随机产生,不加入标签之间的约束,该模型记为 BLSTM_NER_1;在第二组实验中,侧重对比本文提出的识别方法在依次加入本文提出的 E^o 和 E^w、及标签之间的约束关系和融入领域知识之后对识别效果的影响。E^o 由 DataSet4 训练得到,BLSTM_NER 中的打分方法见式(6),该模型记为 BLSTM_NER_2。在 BLSTM_NER_2 的基础上,BLSTM_

NER 中的打分方法替换为式(7)的模型,记为 BLSTM_NER_3。在 BLSTM_NER_3 的基础上,将 E^* 替换为由 DataSet5 训练得到的模型,该模型记为 BLSTM_NER_4。该组实验侧重对比本文提出的 BLSTM_NER 模型的 4 种形式的识别效果。两组实验分别在 DataSet2 和 DataSet3 上进行。

表 2 模型 BLSTM_NER_1 与 CRF 和 SVM 的识别效果的对比实验结果/%

Table 2 Comparison of recognition effect about BLSTM_NER_1,CRF,SVM model/%

数据集	模型名称	人名			地名			机构名		
		P	R	F	P	R	F	P	R	F
DataSet2	CRF	95.40	85.70	90.30	95.70	89.30	92.40	93.70	87.60	90.50
	SVM	93.56	82.41	87.63	84.93	73.15	78.60	87.73	63.78	73.86
	DNN	97.55	87.10	92.03	93.58	77.11	84.55	87.83	71.94	79.09
DataSet3	BLSTM_NER_1	97.72	88.21	92.72	94.31	80.83	87.05	88.97	80.65	84.61
	BLSTM_NER_1	97.67	88.19	92.69	95.42	81.62	87.98	89.86	81.76	85.66

从表 2 所列的实验数据可得出,在 DataSet2 上:1)对于人名这类命名实体,基于深度神经网络的 DNN 模型的识别效果好于传统的机器学习方法 CRF 和 SVM,BLSTM_NER_1 模型的识别效果略好于 DNN 模型的识别效果,但差别不大,原因是由于本文所选的 DataSet2 数据集中的人名实体绝大部分是中国人名,因此一般只由姓氏和名字构成,这类实体构成简单,需要的长距离记忆信息较少,所以选择 BLSTM 模型还是 DNN 模型对识别效果的影响较小。2)对于地名和机构名这两类实体,CRF 模型的识别效果最优,分析原因是由于在 CRF 模型中加入了字符、词性和前后缀等重要的人工设计特征,这些特征由人工针对特定问题精心设计,精确地获得了词语中有价值的信息,对实体识别结果有很重要的正面影响,而 BLSTM_NER_1 模型仅利用到了前后缀特征,因此结果不如基于人工特征的 CRF 方法好,但基于人工特征的方法的代价昂贵;另外,在利用卷积神经网络(Convolutional Neural Network,CNN)模型^[12]识别命名实体的方法中,不加入人工特征和领域词典时,F 值也略低于利用人工特征的方法,这说明本文的方法与利用 CNN 模型的方法在不加入人工特征时具有相似的识别能力。对于地名和机构名这两类实体,BLSTM_NER_1 模型的识别效果好于 DNN 模型,分析其原因是地名和机构名这两类实体通常较长,BLSTM 捕获的长距离信息起到了积极的作用。

因为 DataSet3 上传统方法没有实验数据,所以对比模型

表 3 模型 BLSTM_NER_1 至 BLSTM_NER_4 的识别效果的对比实验结果/%

Table 3 Comparison of recognition effect about BLSTM_NER_1 to BLSTM_NER_4 model/%

数据集	模型名称	人名				地名				机构名			
		P	R	F	Acc	P	R	F	Acc	P	R	F	Acc
DataSet2	BLSTM_NER_1	97.72	88.21	92.72	99.75	94.31	80.83	87.05	99.52	88.97	80.65	84.61	99.71
	BLSTM_NER_2	97.74	88.28	92.77	99.75	96.74	82.78	89.22	99.6	90.56	83.72	87.01	99.75
	BLSTM_NER_3	97.98	88.89	93.21	99.77	96.93	84.45	90.26	99.63	91.13	84.55	87.72	99.76
	BLSTM_NER_4	98.23	89.49	93.66	99.78	97.52	89.34	93.25	99.74	93.74	87.98	90.77	99.82
DataSet3	BLSTM_NER_1	97.67	88.19	92.69	99.78	95.42	81.62	87.98	99.87	89.86	81.76	85.66	99.86
	BLSTM_NER_2	97.69	89.21	93.26	99.8	96.91	82.97	89.4	99.88	90.89	83.35	86.96	99.87
	BLSTM_NER_3	97.81	89.17	93.29	99.8	97.15	84.87	90.6	99.89	91.36	85.28	88.22	99.88
	BLSTM_NER_4	97.92	89.34	93.43	99.81	97.62	89.41	93.33	99.93	94.17	88.21	91.09	99.91

首先,根据表 3 从总体上分析识别的最优结果:1)数据集 DataSet1 和 DataSet2 上所有模型的 Acc 值都很高,而且差异不大,所以不具备评价价值;2)依次加入词向量 E、标签约束和领域知识后,两个数据集上模型 BLSTM_NER_1 至

4.4.1 与传统的机器学习方法比较的实验结果

该组实验结果如表 2 所列,CRF 模型的实验结果为文献[2]中的最优实验结果,SVM 和 DNN 模型的实验结果为文献[11]中的最优实验结果。由于文献[2]和文献[11]中的识别方法的评价指标仅列出 P,R 和 F,因此表 2 中仅对比这 3 个指标。

BLSTM_NER_1 在数据集 DataSet2 和 DataSet3 上的实验结果:1)DataSet3 上 BLSTM_NER_1 模型对人名这类命名实体的识别效果略差于该模型在 DataSet2 上的识别效果,其原因是由于 DataSet3 数据集中的人名数量多于 DataSet2 数据集中的人名数量,且有部分外国人名在分割时按照中国人名的分割方式进行了分割;2)地名和机构名这两类实体的识别效果在 DataSet3 上略好于 DataSet2,原因是由于 DataSet3 中的地名和机构名的平均长度略长于 DataSet2 数据集上的地名和机构名的长度,有效地利用了 BLSTM 捕获的长距离信息。

4.4.2 加入词向量 E、标签约束和领域知识的对比实验结果

由第一组实验得出,文献[2]中的 CRF 方法的综合效果最优,在“输入词向量随机产生,不加入标签之间的约束”时,本文的方法并未体现出优势。所以进行了第二组实验,由表 2 和表 3 的实验数据可以看出,在 BLSTM_NER_4 上的实验结果全面超越了 CRF 方法的识别效果。

在第二组实验中,加入了预训练的词向量、标签之间的约束关系和少量领域知识,DataSet2 和 DataSet3 上的人名、地名和机构名的 P,R,F 和 Acc 指标的对比实验结果如表 3 所列。表 3 所列实验数据为综合考虑模型的 P,R 和 1-Spec 3 个指标、在特定的模型输出阈值下给出的最优结果。读者可根据应用命名实体识别问题的场合,调整模型的输出的阈值,选择具有最优的 R 值、或 P 值、或 1-Spec 值、或 Acc 值的模型进行使用。

BLSTM_NER_4 的人名、地名和机构名 3 种实体的 P,R 和 F 的识别效果逐渐提升,BLSTM_NER_4 在 DataSet2 上的识别效果超过了传统的命名实体识别方法中的最优结果。

其次,根据表 3 从模型间的对比角度分析识别的最优结

果:1)在 DataSet2 和 DataSet3 上,人名这类命名实体的识别效果在增加前后缀和标签约束后的 BLSTM_NER_2 模型上较 BLSTM_NER_1 略有提升,分析原因是人名这种实体的前后缀信息和标签约束对识别结果有一定的影响,但影响不大;在加入领域知识的 BLSTM_NER_3 和 BLSTM_NER_4 模型上的提升较之前稍微明显,这说明领域知识无论以词向量的形式还是正则项的形式融入模型对模型的效果都有较大的提升;2)而地名和机构名这两类命名实体在增加前缀、后缀和标签约束后,BLSTM_NER_2 模型的识别效果明显好于 BLSTM_NER_1 模型的识别效果,分析其原因是由于地名和机构名这两类命名实体的长度通常较人名长,而且构成复杂,由基于上下文的词向量和 BLSTM_E 模型训练得到的词向量对识别效果有积极的影响,也就是词语的上下文和前后缀信息对命名实体识别有重要影响,这一点在 CRF 模型中也得到了验证。因此建议将命名实体识别的有效特征向量化,作为神经网络模型的特征输入,对命名实体识别的效果会有较显著的提升。在加入领域知识后,地名和机构名的识别效果在 BLSTM_NER_3 和 BLSTM_NER_4 上有更明显的提升,说明即使使用少量的地名和机构名词典也会起到较大的积极作用,建议读者在获取领域知识代价不高的情况下尽量使用领域知识。

为检验本文提出的 BLSTM_NER_1 至 BLSTM_NER_4 模型的有效性,给出 ROC 曲线,如图 4—图 6 所示,AUC 值如表 4 所列。图 4—图 6 和表 4 所列实验数据为验证模型在不同的输出阈值时模型的识别效果。由于所有模型的识别效果提升幅度不大,且 1-Spec 的值在不超过 0.02 时 Sent 的值就达到 0.8,为放大不同模型的识别效果间的差异,图 4—图 6 中的纵轴 Sent 范围从 0.7 开始、略去 0.0 到 0.7 之间的 ROC 曲线走势。

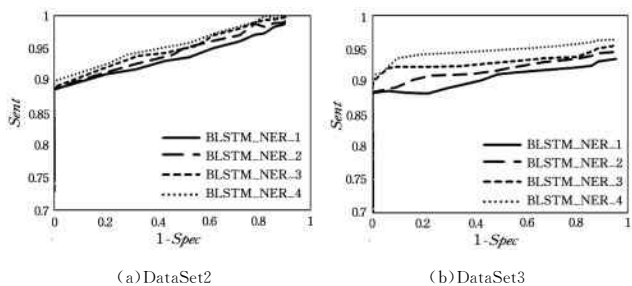


图 4 人名实体的 ROC 曲线图

Fig. 4 ROC curve of person named entities

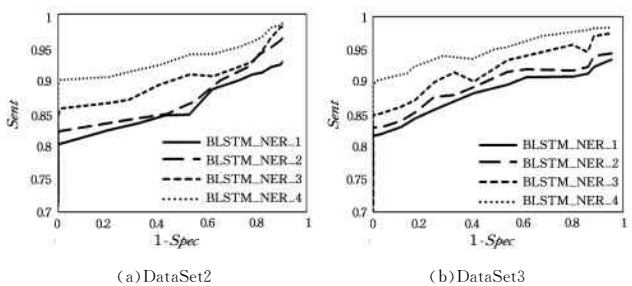


图 5 地名实体的 ROC 曲线图

Fig. 5 ROC curve of location named entities

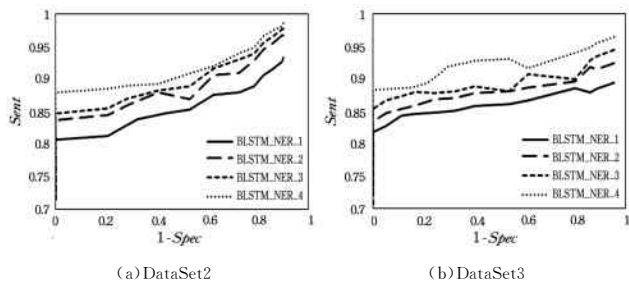


图 6 机构名实体的 ROC 曲线图

Fig. 6 ROC curve of organization named entities

表 4 在 DataSet2 和 DataSet3 上不同模型、不同实体的 AUC 值
Table 4 AUC values for different entities in different models on DataSet2 and DataSet3

数据集	模型	实体类型		
		人名	地名	机构名
DataSet2	BLSTM_NER_1	0.9387	0.8668	0.8627
	BLSTM_NER_2	0.9462	0.8803	0.8919
	BLSTM_NER_3	0.9520	0.9057	0.9023
	BLSTM_NER_4	0.9571	0.9359	0.9183
DataSet3	BLSTM_NER_1	0.9072	0.8850	0.8640
	BLSTM_NER_2	0.9196	0.8973	0.8832
	BLSTM_NER_3	0.9316	0.9203	0.8953
	BLSTM_NER_4	0.9471	0.9498	0.9222

首先,根据模型随 1-Spec 值和 Sent 值变化的情况,从总体上分析模型的识别效果。1)根据图 4—图 6 的 ROC 曲线走势和表 4 所列的 AUC 值上可以看出,无论在哪个数据集上,随着 1-Spec 的提高,Sent 稳步增加,这说明模型的识别效果比较稳定。2)在依次加入前后缀信息、标签约束信息和领域知识后,模型的识别效果也依次提升;但在人名实体上的提升效果不如在地名和机构名实体上明显,其原因是由于人名词语长度较短、人名的姓氏和名字没有较强的约束关系、人名词典中的人名与待识别的文本中的人名实体没有很强的相关性,因此前后缀信息、标签约束信息和领域知识对人名这类实体有一定的影响,但影响不大。

其次,根据模型随 1-Spec 值和 Sent 值变化的情况,从模型间的对比角度分析模型的识别效果。1)根据图 4(a)和图 4(b)可以看出,人名实体在数据集 DataSet3 上的提升高于 DataSet2,分析原因是 DataSet3 中的人名实体包含国外的音译名字(松/B-PER 德/I-PER 拉/I-PER 尔/E-PER),一个名字由超过两部分构成的情况较 DataSet2 上的要多,导致人名词语的平均长度大于 DataSet2 上的人名词语的平均长度,所以前缀、后缀、标签间约束和领域知识起到了更积极的作用;但 BLSTM_NER_4 较 BLSTM_NER_3 的效果提升较其他模型间的提升较小,这说明人名实体在加入正则项的提升上不明显,尤其在数据集 DataSet2 上。2)根据图 5(a)和图 5(b)可以看出,地名实体的 BLSTM_NER_3 和 BLSTM_NER_4 模型的提升效果较 BLSTM_NER_2 更明显,这说明加入领域知识对地名实体识别的效果提升较加入标签约束更明显,建议读者在获取领域知识代价不大的前提下,可在地名实体识别模型中融入少量领域知识。3)根据图 6(a)和图 6(b)可以看出,机构名实体的 BLSTM_NER_2 模型较 BLSTM_NER_1 模型的效果提升明显,说明前缀信息、后缀信息和标签间约

束作用明显;BLSTM_NER_4模型较BLSTM_NER_3模型的效果提升明显,也就是机构名的前、后缀信息特征和领域知识作用明显,分析原因是中文的机构名有较明显的前缀信息和后缀信息(大学,公司);但BLSTM_NER_3模型较BLSTM_NER_2模型的效果提升不够明显,说明正则项对效果提升略差,建议读者在识别机构名时不使用正则项。

结束语 本文提出了一种利用神经网络模型解决命名实体识别问题的有效方法,该方法不直接依赖人工特征和外部资源,只是利用了少量的监督数据、领域知识和大量的无标注数据,解决了目前的机器学习方法中过度依赖人工特征和领域知识及语料不足的问题。本文提出的命名实体识别方法中融入了词语的上下文信息、词语的前后缀信息和领域词典,将这种信息特征化为词的分布表示特征;考虑了词语的标签之间的约束关系,进一步提高了识别的效果。但本文只是顺序地读取数据对命名实体进行识别,每个词语对命名实体的影响同等重要,并没有考虑不同的词语对命名实体的不同影响,如何将深度学习的注意力机制引入本文的模型中、重点关注对命名实体识别有重要影响的词语,是进一步需要解决的问题。

参考文献

- [1] SUN L P, GUO G, TANG W W, et al. Enterprise Abbreviation Prediction Based on Constitution Pattern and Conditional Random Field[J]. Journal of Computer Applications, 2016, 36(2): 449-454. (in Chinese)
孙丽萍, 过弋, 唐文武, 等. 基于构成模式和条件随机场的企业简称预测[J]. 计算机应用, 2016, 36(2): 449-454.
- [2] DUAN H Z, ZHENG Y. A Study on Features of the CRFs-based Chinese Named Entity Recognition[J]. International Journal of Advanced Intelligence Paradigms, 2011, 3(2): 287-294.
- [3] HUANG D G, LI Z Z, WAN R. Chinese Organization Name Recognition Using Cascaded Model Based on SVM and CRF[J]. Journal of Dalian University of Technology, 2010, 50(5): 782-787. (in Chinese)
黄德根, 李泽中, 万如. 基于SVM和CRF的双层模型中文机构名识别[J]. 大连理工大学学报, 2010, 50(5): 782-787.
- [4] 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. (in Chinese)
冯艳红, 于红, 孙庚, 等. 基于词向量和CRF的领域术语识别方法[J]. 计算机应用, 2016, 36(11): 3146-3151.
- [5] FENG Y T, ZHANG H J, HAO W N. Named Entity Recognition for Military Text[J]. Computer Science, 2015, 42(7): 15-18, 47. (in Chinese)
冯蕴天, 张宏军, 郝文宁. 面向军事文本的命名实体识别[J]. 计算机科学, 2015, 42(7): 15-18, 47.
- [6] BENGIO Y, SCHWENK H, WENÉCAL J S, et al. A Neural Probabilistic Language Model[J]. Springer Berlin Heidelberg, 2006, 3(6): 1137-1155.
- [7] MNIH A, HINTON G E. A Scalable Hierarchical Distributed Language Model[C]//NIPS2008: Advances in Neural Information Processing Systems 21. Vancouver, British Columbia, Canada: Curran Associates, Inc., 2008: 1081-1088.
- [8] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Neural Computation, 2014, 4: 1771-1800.
- [9] LI L S, HE H L, LIU S S, et al. Research of Word Representations on Biomedical Named Entity Recognition[J]. Journal of Chinese Computer System, 2016(2): 302-307.
- [10] 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. IEEE, 2014: 433-438.
- [11] WANG G Y. Research of Chinese Named Entity Recognition Based on Deep Learning[D]. Beijing: Beijing University of Technology, 2015: 33-38. (in Chinese)
王国昱. 基于深度学习的中文命名实体识别研究[D]. 北京: 北京工业大学, 2015: 33-38.
- [12] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural Language Processing (almost) from Scratch[J]. The Journal of Machine Learning Research, 2011, 12(1): 2493-2537.
- [13] HOCHREITER S, SCHMIDHUBER J. Long Short-term Memory[J]. Neural computation, 1997, 9(8): 1735-1780.
- [14] GRAVES A, MOHAMED A R, HINTON G. Speech Recognition with Deep Recurrent Neural Networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, 2013. Vancouver, BC, Canada, 2013: 6645-6649.
- [15] SCHUSTER M, PALIWAL K K. Bidirectional Recurrent Neural Networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.
- [16] GRAVES A, SCHMIDHUBER J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks, 2005, 18(5): 602-610.
- [17] NIELSEN M A. Neural Networks and Deep Learning[EB/OL]. (2015) [2017-02-14]. <http://neuralnetworksanddeeplearning.com/chapter3.html>.
- [18] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and Their Compositionality[C]//NIPS2013: Advances in Neural Information Processing Systems 26. Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2013: 3111-3119.
- [19] WANG L, LU'IS T, MARUJO L, et al. 2015b. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.
- [20] LAI S W. Word and Document Embedding Based on Neural Network Approaches[D]. Beijing: University of Chinese Academy of Sciences, 2016: 27-39. (in Chinese)
来斯惟. 基于神经网络的词和文档语义向量表示方法研究[D]. 北京: 中国科学院大学, 2016: 27-39.
- [21] 搜狗全网新闻数据(SogouCA)[EB/OL]. [2017-02-14]. <http://www.sogou.com/labs/dl/ca.html>.
- [22] 搜狗输入法词库[EB/OL]. [2017-02-14]. <http://pinyin.sogou.com/dict>.
- [23] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent Neural Network Regularization[EB/OL]. (2015-2) [2016-09]. <http://arxiv.org/abs/1409.2329v5>.