基于循环神经网络序列标注的中文分词研究

刁 琦[1],古丽米拉·克孜尔别克[1],钟丽峰[2],张 健[3],张志强[1]
(1. 新疆农业大学 计算机与信息工程学院,新疆 乌鲁木齐 830052;
2. 新疆维吾尔族自治区图书馆,新疆 乌鲁木齐 830052;
3. 新疆虹联软件有限公司,新疆 乌鲁木齐 830052)

摘 要:分词是中文自然语言处理中的关键技术。在自然语言处理中,序列标注在中文分词中有着极其重要的应用。当前主流的中文分词方法是基于监督学习,从中文文本中提取特征信息。这些方法未能充分地利用上下文信息对中文进行分割,缺乏长距离信息约束能力。针对上述问题进行研究,提出在序列标注的前提下利用双向循环神经网络模型进行中文分词,避免了窗口对上下文大小的限制,可以获得一个词的前面和后面的上下文信息,通过增加上下文能够有效地解决梯度爆炸和爆的问题,然后再在输入层加入训练好的上下文词向量,取得相对较好的分词效果。实验结果表明,该算法的使用可以达到 97.3% 的中文分词准确率,与传统机器学习分词算法相比,效果较为显著。

关键词:自然语言处理;循环神经网络;序列标注;中文分词;监督学习
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)10-0065-04
doi:10.3969/j.issn.1673-629X.2017.10.014

Research on Chinese Word Segmentation Method of Sequence Labeling Based on Recurrent Neural Networks

DIAO Qi[1],Gulimila·KEZIERBIEKE[1],Zhong Li-feng[2],ZHANG Jian[3],ZHANG Zhi-qiang[1]
(1. College of Computer & Information Engineering,Xinjiang Agricultural University,Urumqi 830052,China;
2. Library of Xinjiang Uygur Autonomous Region,Urumqi 830052,China;
3. Xinjiang Honglian Software Co.,Ltd.,Urumqi 830052,China)

Abstract:Word segmentation is a key technology in Chinese natural language processing. In natural language processing,sequence labeling plays an important role in Chinese word segmentation. The current mainstream Chinese word segmentation method is based on supervised learning,extraction of feature information from the Chinese text. However,they cannot make full use of context information to segment Chinese,and lack of long-distance information constraint. In order to solve it,Chinese word segmentation is carried on based on bi-directional recurrent neural network model on the premise of sequence labeling,avoiding the limitation of window size on context,obtaining the context information of the front and back of a word. It can effectively solve the problem of gradient explosion and explosion by adding context information,and then add a good context vector in the input layer to obtain a relatively good word segmentation effect. The experimental results show that it can achieve 97.3% accuracy of Chinese word segmentation and is superior to the traditional machine learning segmentation algorithm in the effect.

Key words:natural language processing;recurrent neural network;sequence annotation;Chinese word segmentation;supervised learning

0 引 言

分词是中文处理的一项根本任务。词是“最小的能独立运用的语言单位”[1]。中文与英文有所不同,英文中词与词之间用空格天然分割,而中文具有大字符连续书写的特点,需要对其进行有效分割。分词更重要的一个功能是帮助计算机理解文字。因此,在自然语言处理中,中文分词[2]是一项重要的基础技术。

近年来,中文分词技术有了长足进步。陈硕等[3]提出一种使用误差反传神经网络与一种改进的匹配算法相结合的中文分词技术,该方法不需要标注语义信

收稿日期:2016-11-18 **修回日期**:2017-03-09 **网络出版时间**:2017-07-19
基金项目:新疆维吾尔自治区科技计划项目(2015X0106)
作者简介:刁 琦(1989-),男,硕士研究生,研究方向为智能计算及应用;古丽米拉·克孜尔别克,通信作者,副教授,研究方向为现代通信技术及嵌入式技术。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170719.1112.072.html

息,适应性、鲁棒性好,且训练结果占用空间小,有一定冗余性,对比单纯的神经网络分词方法有较大提高。巫黄旭[4]提出一种基于统计学习的分词方法,以期在最小人工干预的条件下达到尽可能高的分词性能,扩展二元语法模型至三元语法模型,提出性能优化的三元语法获取和使用方法,但语法模型结构较为简单。何嘉[5]在分析进化神经网络及神经网络分词法优势的基础上,将改进的免疫遗传算法应用到基于神经网络的中文分词模型中,对歧义进行处理。尽管这些方法效果较好,但是标记特征工作量大,训练的模型过度拟合训练语料库。

循环神经网络(RNNs)[6]广泛应用于机器翻译[7]、语音识别[8]、图像描述生成等领域。相比于传统前馈神经网络,其特点是可以存在有向环,将上一次的输出作为本次的输入。而与前馈神经网络[9]的最大区别是:前馈神经网络要求输入的上下文是固定长度的,也就是说 n-gram 中的 n 要求是个固定值,而在 LSTM 基础上扩展的循环神经网络不限制上下文的长度,可以充分利用所有上文提供的信息来预测下一个词,本次预测的中间隐层信息可以在下一次预测里循环使用。为此,文中试图将循环神经网络模型应用在中文分词方面。

1 理论研究

1.1 词向量特征

在自然语言处理中,需要将自然语言理解问题转化为机器学习问题,即自然语言的符号数学化。目前最常用的词表示方法是 One-hot Representation[10],把文本中每一个词表示为多维向量,向量的维度是词表大小,其中绝大部分数元素表示为 0,只有一个维度的值为 1,这个维度就代表了当前的词。例如,“学生”表示为[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0];“班级”表示为[0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]。

常将 One-hot 采用稀疏的方式对词进行存储,即为每个词分配对应的数字 ID。该方法简单易用,广泛应用于各种自然语言处理任务中,如 N-gram 模型[11]中就采用该方法。但这种表述方法也存在一定问题,即表示的任意两词之间是孤立的,无法表示这两个词之间的依赖关系,从词向量上看不出两个词是否存在相关关系;采用稀疏表示法[12],在处理某些任务,如构建 N-gram 模型时,会引起维数灾难问题。

而在深度学习[13]中,一般采用分布式表示(Distributed Representation)的方法表示词向量,该方法最早由 Hinton[14]提出,通常称为 Word Representation。该方法将词用一种低维实数向量表示,优点在于相似的词在距离上更接近,体现出不同词之间的相关性,从而反映词之间的依赖关系。同时,较低的维度也使特征向量在应用时有一个可接受的复杂度。因此,新近提出的许多语言模型,如潜在语义分析(Latent Semantic Analysis,LSA)模型和潜在狄利克雷分布(Latent Dirichlet Allocation,LDA)[15]模型,及目前流行的神经网络模型等,都采用这种方法表示词向量。

1.2 循环神经网络

中文分词通常被看作为基于字符的序列标签[16]。每个字符贴上{ B,M,E,S }来表示分割。一个多字符分割用{ B,M,E }表示开始、中间、结束,S 表示单个字符分割。序列标注[17]就是针对一个线性输入序列:$x = x_1, x_2, \cdots, x_n$,给线性序列中的每个元素打上标签集中的某个标签,即 $y = y_1, y_2, \cdots, y_n$。

中文分词的序列标注过程如图 1 所示。

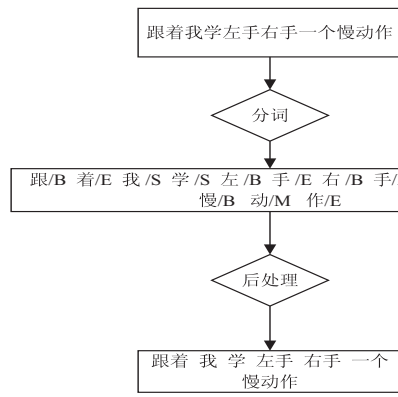

图 1 中文分词序列标注过程

神经网络的中文分词通用模块主要由三部分组成[18]:词向量化;一系列典型的神经网络层;标签推理层。通用框架如图 2 所示。

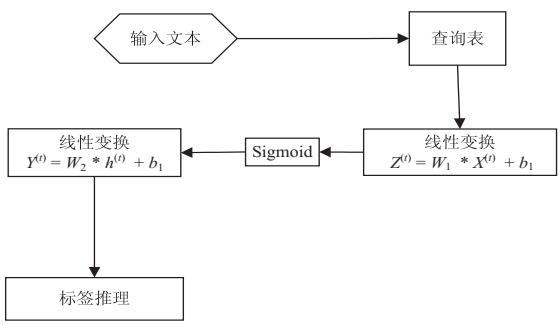

图 2 基于神经网络的中文分词通用框架

基于字标注的分词方法是基于一个局部滑动窗口,假设一个字的标签极大地依赖于其相邻位置的字。给定长度为 n 的文本序列 $c(1:n)$,大小为 k 的窗口从文本序列的第一个字 $c(1)$ 滑动至最后一个字 $c(n)$。对于序列中每个字 $c(1)$,窗口大小为 5 时,上下文信息($c(t-2)$,$c(t-1)$,$c(t)$,$c(t+1)$,$c(t+2)$)被送入查询表中,当字的范围超过了序列边界时,将以诸如“start”和“end”等特殊标记来补充。然后,将查询表中提取的字向量连接成一个向量 $\boldsymbol{X}(t)$;接着,在神

经网络下一层中, $\boldsymbol{X}(t)$ 经过先行变换后经由 sigmoid 函数 $\sigma(x) = (1 + e^{-x})^{-1}$ 或 tanh 函数激活。

$$h(t) = \sigma(w_1 x(t) + b_1) \tag{1}$$

接下来根据给定的标注集,将经过一个相似的线性变换,不同之处在于没有线性函数,得到的 $\boldsymbol{y}(t)$ 是每个可能标签的得分向量。文中选定的是更能充分表达词信息的四位标注集 $\{B, M, E, S\}$。

$$\boldsymbol{y}(t) = w_2 h(t) + b_2 \tag{2}$$

为了建模标签间依赖,引入转移得分向量 $\boldsymbol{A}_{ij}$,用于衡量从标签 i 跳转到标签 j 的概率。过往的研究表明,引入转移得分向量非常适用于中文分词等序列标注的任务,但它仅利用了长度有限的窗口信息。

1.3 LSTM 网络

Recurrent Neural Networks(RNNs)具有循环的网络结构,具备保持信息的能力,其网络结构如图 3 所示。

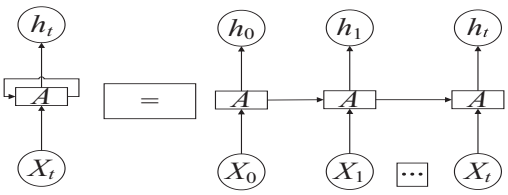

图 3　RNNs 网络结构

RNNs 中的循环网络模块将信息从网络上一层传输到下一层,网络模块的隐含层每个时刻的输出都依赖于以往时刻的信息。RNNs 的链式属性表明其与序列标注问题存在密切联系,目前已被应用到文本分类器和机器翻译等 NLP 任务中。在 RNNs 的训练中,存在梯度爆炸和消失的问题;且传统的 RNNs 难以保较长时间的记忆。

LSTM[19](Long Short-Term Memory)网络是 RNNs 的扩展,用来避免长期依赖问题。LSTM 的重复神经网络模块具有不同的结构,这与朴素 RNNs 网络不同,存在 4 个以特殊方式影响的神经网络层,网络结构如图 4 所示。

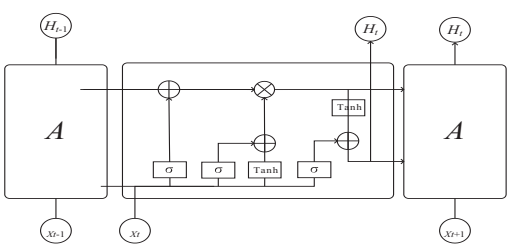

图 4　LSTM 神经网络结构

LSTM 网络的关键在于细胞状态,有点类似于传送带。在 LSTM 中,通过门结构对细胞状态增加或删除信息,而门结构采用选择性让信息通过的方式,通常由一个 sigmoid 神经网络层和逐点乘积操作组成(sigmoid 层的输出在 0 到 1 之间,定义了信息通过的程度,0 表示什么都不让过,1 表示所有都让过)。

LSTM 网络具有输入门(input gates)、忘记门(forget gates)和输出门(output gates)三种门结构,用以保持和更新细胞状态。

(1)从细胞状态中忘记信息,由忘记门的 sigmoid 层决定,以当前的输入 x_t 和上一层的输出 h_{t-1} 作为输入,在 $t-1$ 时刻的细胞状态输出 f_t 为;

$$f_t = \sigma(w_f \cdot (h_{t-1}, x_i) + b_f) \tag{3}$$

(2)在细胞状态中存储信息,主要由两部分组成:输入门的 sigmoid 层的结果 i_t,作为将更新的信息;由 tanh 层新创建向量 $\tilde{c}$,添加进细胞状态中。将旧的细胞状态 C_{t-1},乘以 f_t,用以遗忘信息,与新的候选信息 $i_t \cdot \tilde{c}$ 的和,生成细胞状态的更新。

$$i_t = \sigma(w_i \cdot (h_{t-1}, x_t) + b_i) \tag{4}$$

$$\tilde{c} = \tanh(w_c \cdot (h_{t-1}, x_t) + b_c) \tag{5}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \widetilde{c_t} \tag{6}$$

(3)输出信息由输出门决定。先使用 sigmoid 层决定要输出细胞状态的部分信息,接着用 tanh 处理细胞状态,两部分信息的乘积得到输出的值。

$$o_t = \sigma(w_o \cdot (h \cdot,) + b_o) \tag{7}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{8}$$

LSTM 网络模型已成功应用于诸如文本/情感分类、机器翻译、智能问答和看图说话[20]等自然语言处理任务中。由于 LSTM 网络记忆单元去学习从细胞状态中忘记信息、去更新细胞状态的信息,而且具有学习文本序列中远距离依赖的特性,很自然地想到可以使用 LSTM 网络模型进行中文分词的任务。

1.4 基于循环神经网络的中文分词架构

在中文分析任务中,LSTM 记忆单元的输入来自上下文窗口的汉字。对每个汉字 $C(t)$,LSTM 记忆单元的输入为 $X(t)$,由上下文字嵌入($c^{(t)}, \cdots, c^{(t+k)}$)连接而成。其中 k 代表与当前字的距离。LSTM 单元的输出在经过线性变换后用于标签推理函数,推理汉字对应的标签。

$$x(t) = v_c^{t-k} \oplus \cdots \oplus v_c^{(t+k)} \tag{9}$$

文中提出的架构如图 5 所示。为了建模标签间依赖,在以往的神经网络模型方法中引入转移得分向量 $\boldsymbol{A}_{ij}$,用于衡量从标签 i 跳转到标签 j 的概率。对于输入文本序列 $c^{(1:n)}$,其标注的标签序列为 $y^{(1:n)}$,序列级的得分是标签转移得分和网络标注得分的总和。

$$s(c^{(1:n)}, y^{(1:n)} = \sum_{t=1}^{n} A_{y^{(t-1)} y^{(t)}} + y_{y^{(t)}}^{(t)}) \tag{10}$$

2　分词评估标准

中文分词性能评估指标,采用了分词评测常用的

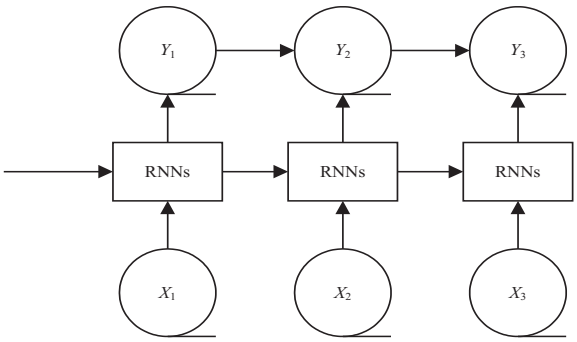

图 5 文中提出的架构

R(召回率)、P(准确率)和 F-measure(F 值),以 F 值为主要评测指标。

$$P = \text{正确词数/识别的词数} * 100\% \quad (11)$$

$$R = \text{正确词数/原有词数} * 100\% \quad (12)$$

$$F = 2PR/(P+R) \quad (13)$$

3 实 验

3.1 实验设备

实验设备如表 1 所示。

软件方面使用 python2.7,安装好了 keras,theano 及相关库。

表 1 实验设备

电脑型号	操作系统	处理器	主板	内存	主硬盘	显卡	显示器
戴尔 DELL T3400 塔式电脑	Windows 7 旗舰版 64 位 SP1 (DirectX 11)	Intel(R)Core(TM) 2Duo CPU E8400 @ 3.00 GHz 3.00 GHz	联想 31900005WIN8 STD PRC (英特尔 HM76 4 GB Express 芯片组)		ST3320418AS (320 GB)	OHY553	DELL E1909W 分辨率 1 400 * 900

3.2 实验结果

语料库由 2 亿字的中文语料训练形成,该中文语料含有 50 本电子书、2 年的人民日报,内容涵盖范围非常广泛,包含外交、政治、经济、文化、民生等众多领域。测试集为新疆维吾尔自治区科技计划项目提供的 3 组数据,共 2 000 句,内容包含政治、外交、体育、民俗、文化和日常生活等方面。实验结果如下:P 为 0.973,R 为 0.971,F 为 0.972。

4 结束语

文中实验基于循环神经网络模型,采用四词位标注,在循环神经网络层输入预先训练的词向量,对实验的中文语料库进行分词。测试结果表明,该算法较传统的中文分词效果要好。

参考文献:

[1] 汉语信息处理词汇 01 部分:基本术语(GB12200.1-90)6[S].北京:中国标准出版社,1991.

[2] 奉国和,郑 伟.国内中文自动分词技术研究综述[J].图书情报工作,2011,55(2):41-45.

[3] 李 华,陈 硕,练睿婷.神经网络和匹配融合的中文分词研究[J].心智与计算,2010(2):117-127.

[4] 巫黄旭.基于统计学习的中文分词改进及其在面向应用分词中的应用[D].杭州:浙江大学,2012.

[5] 何 嘉.基于遗传算法优化的中文分词研究[D].成都:电子科技大学,2012.

[6] Graves A. Supervised sequence labelling with recurrent neural networks[M]. Berlin:Springer,2012.

[7] 蒋锐滢,崔 磊,何 晶,等.基于主题模型和统计机器翻译方法的中文格律诗自动生成[J].计算机学报,2015,38(12):2426-2436.

[8] 王山海,景新幸,杨海燕.基于深度学习神经网络的孤立词语音识别的研究[J].计算机应用研究,2015,32(8):2289-2291.

[9] Bebis G,Georgiopoulos M. Feed-forward neural networks[J]. IEEE Potentials,1994,13(4):27-31.

[10] Landauer T K,Foltz P W,Laham D. An introduction to latent semantic analysis[J]. Discourse Processes,1998,25(2-3):259-284.

[11] 陈天莹,陈 蓉,潘璐璐,等.基于前后文 n-gram 模型的古汉语句子切分[J].计算机工程,2007,33(3):192-193.

[12] 栾悉道,王卫威,谢毓湘,等.非线性稀疏表示理论及其应用[J].计算机科学,2014,41(8):13-18.

[13] 张建明,詹智财,成科扬,等.深度学习的研究与发展[J].江苏大学学报:自然科学版,2015,36(2):191-200.

[14] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the 8th annual conference of the cognitive science society. [s. l.]:[s. n.],1986.

[15] Blei D M,Ng A Y,Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research,2003,3:993-1022.

[16] 梁喜涛,顾 磊.中文分词与词性标注研究[J].计算机技术与发展,2015,25(2):175-180.

[17] 王 昊,邓三鸿,苏新宁.基于字序列标注的中文关键词抽取研究[J].现代图书情报技术,2011(12):39-45.

[18] Zheng X,Chen H,Xu T. Deep learning for Chinese word segmentation and POS tagging[C]//Proceedings of conference on empirical methods in natural language processing. [s. l.]:[s. n.],2013:647-657.

[19] Graves A,Jaitly N,Mohamed A. Hybrid speech recognition with deep bidirectional LSTM[C]//IEEE workshop on automatic speech recognition and understanding. [s. l.]:IEEE,2013:273-278.

[20] 翟 艳,冯红梅.基于"看图说话"任务的汉语学习者口语流利性发展研究[J].华文教学与研究,2014(4):1-7.