计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(1):239−243 [doi: 10.15888/j.cnki.csa.006156]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度神经网络的武器名称识别①

游　飞[1], 张　激[2], 邱　定[1], 于铭华[1]

[1](华东计算技术研究所 系统平台部, 上海 201808)
[2](华东计算技术研究所 总师办, 上海 201808)
通讯作者: 游　飞, E-mail: youfefi@mail.ustc.edu.cn

摘　要: 科学技术的进步, 推进着军事武器装备的快速更新. 在高度信息化的时代, 急需智能化军事信息处理技术. 本文针对飞行器、坦克车辆、火炮弹炮、导弹武器等军事文本中的武器命名实体, 提出了基于词向量、词状态的特征, 利用深度神经网络模型的识别方法. 实验表明: 在测试语料上取得 F-1 值 0.9102 的效果.

关键词: 深度神经网络; 词向量; 命名实体识别

引用格式: 游飞,张激,邱定,于铭华.基于深度神经网络的武器名称识别.计算机系统应用,2018,27(1):239–243. http://www.c-s-a.org.cn/1003-3254/6156.html

Weapon Named Entity Recognition Based on Deep Neural Network

YOU Fei[1], ZHANG Ji[2], QIU Ding[1], YU Ming-Hua[1]

[1](Department of System Platform, The 32nd Research Institute of China Electronics Technology Group Corporation, Shanghai 201808, China)
[2](Chief Engineer Office, The 32nd Research Institute of China Electrons Technology Group Corporation, Shanghai 201808, China)

Abstract: The development of computer science makes military weapons and equipment update fast. In the highly-informed society, the intelligent information processing technology in military, is badly needed. This paper proposes an identification method with the model of deep neural network based on the character of word vector and state. It's for weaponry in electronic text, such as aircraft, tank vehicle, artillery missile and missile weapon. The experiment shows the value of F-1 which equals 0.9102 on the test corpus.

Key words: deep neutral network; word vector; named entity recognition

科学技术的进步推进着军事武器装备的快速更迭, 同时信息处理技术的快速发展, 我军的高度信息化时代正式到来. 在日常军事训练和行动中, 产生了大量的以电子文本形式存在的信息. 如何高效自动化地处理这些海量的文本成为急需解决的问题.

命名实体识别 (Named Entity Recognition, NER) 已经成为许多自然语言处理应用的重要步骤, 例如问答系统、信息提取和机器翻译[1], 是自然语言处理中的一项重要的基础工作. 然而命名实体识别的效果受限于特定的领域和语言, 这就需要为不同领域不同语言量身定制一套识别系统.

命名实体识别最初是在第六届 MUC 会议作为一个子任务提出的[2]. 命名实体识别的主要任务是识别文本中出现的专有名称和数量短语, 并对其加以归类. 早期的命名实体基于字典和规则的方法识别, 字典和规则的编写需要语言专家的参与, 且不能完全覆盖所有的实体. 之后, 人们开始提出基于将统计的方法, 统计的方法能够有效的捕捉到命名实体的位置或特征现象, 接着用维特比 (Viterbi) 算法求解最佳的状态序列. 基于统计方法的优点是不需要丰富的语言学知识、可移

① 收稿时间: 2017-04-10; 修改时间: 2017-04-26; 采用时间: 2017-05-08; csa 在线出版时间: 2017-12-22

植性较好, 缺点是需要大量的人工进行语料的标注. 基于统计方法主要的有: Bikel 等人[3]最早将隐马尔科夫(Hidden Markov Model, HMM) 方法用于命名实体识别. Ratnaparkhi 等人[4]提出最大熵 (Maximum Entropy, EM) 模型用于语言分类的问题.

中文的命名实体的研究紧跟其后, 始于上世纪 90 年代初. 由于语言的特性, 中文的命名实体识别的难度较难, 效果较差. 命名实体识别任务中涉及到分词和词法分析等任务, 英文中词的边界明显, 词性特性特征显著, 而中文中存在一词多义, 词边界模糊等现象. 国内的孙茂松等[5]较早地对中文人名进行识别. 俞鸿魁等[6]基于层叠隐马尔科夫模型进行中文命名实体识别, 达到较高的识别准确率. 姜文志等[7]基于条件随机场(Conditional Random Field, CRF) 和规则的方法对军事命名实体进行了识别.

最近, 由于深度学习能够从大量的无标记的语料中学习特征, 利用深度学习模型解决命名实体识别已经成为的趋势[8]. 深度学习属于机器学习领域, 它能够通过构造深度神经网络 (Deep Neural Networks, DNN) 模型学习高层的特征[9]. Dr. Ronan Collobert 等人[10]基于深度神经模型处理词性标注、命名实体识别等问题, 并取得了当时最好水平.

在军事信息处理领域, 军事专有名词的识别是非常重要的一项工作. 目前许多军事信息处理系统的实体基于字典、规则或统计的方法. 本文主要研究词的向量的表示和词向量模型的训练, 借鉴已有的深度神经网络模型, 在训练集上训练模型, 观察不同参数下的测试结果.

1 深度神经网络模型

深度神经网络从狭义上讲是一个具有多层感知机模型, 近些年深度神经网络模型被应用在自然语言处理的许多任务中并取得了显著的效果, 如: 词性标注、命名实体识别、语块识别等. 本文基于深度神经网络构建出武器名称识别的模型. 模型的结构如图 1 所示. 底层是神经网络的输入层, 即连续化的词向量窗口. 由于模型的输入是固定的格式, 本文将固定维度的词向量和词性向量作为输入, 通过中间隐层的非线性变换, 学习到高层的特征, 即词的上下文的特征, 本文将词对应实体识别的四种状态, 作为网络模型的输出. 最后通过再结合训练集的状态转移概率求得句子的最佳标注序列.

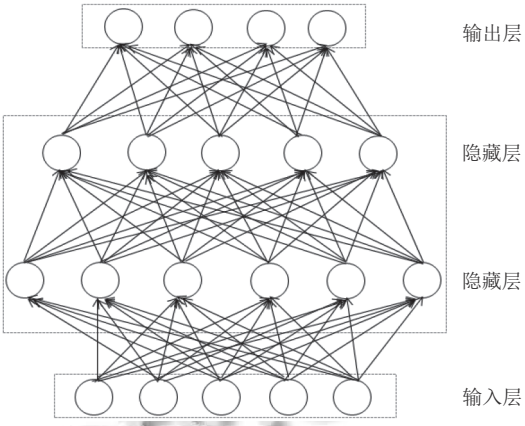

图 1 深度神经网络模型

1.1 词向量模型

将每个词语表示成一个低维的实数向量, 那么任意两个词语之间的距离可以用欧式距离表示. 这种特征表示可以解决机器学习中维数灾难和局部泛化等问题[10]. 与传统的基于统计记录上下特征的方法相比, 它可以更好的捕捉到数据之间的固有联系, 而且不需要进行人工标注.

在基于基于词向量特征的命名实体识别任务中, 常把训练集的单词 W, 表示为一个固定维度的列向量, 作为深度神经的输入. 该向量可以很好的表示句子信息和语义相似度. 理想状态下, DNN 的输入为若干词语的存储矩阵 $M^{(D\times W)}$, D 是一个词语向量的维度, 而 W 是领域词语字典的大小. 在命名实体识别任务中, 当前的句子能够很好的体现的词语之间的关联, 而句子之间的词语关联较弱. 因为 DNN 模型的输入是固定的格式, 本文大小为 W 的窗口作为输入, 窗口中间是当前词为 M_i, 则它前后的 $(k\text{-}1/2)$ 个词语代表它的上下文, 即为词序 $[M_{i-(w-1)/2}, M_{i-(w-3)/2}, \ldots, M_i, \ldots, M_{i+(w-3)/2}, M_{i+(w-1)/2}]$. 对于位于句前和句尾的当前词, 本文动窗口的前部或尾部做随机填充处理, 考虑到词性在特定语言中有普遍的规律, 本文选用参考北大计算所词性标注集简表, 选用常用的词性 14 个, 并将其他词性视为统一词性, 将窗口中每个词映射到 15 维的词性向量中, 并将词性向量与词向量拼接, 即把这 W 个词语的特征向量 $M^{((D+15)\times W)}$ 作为模型的输入.

1.2 隐藏层

两层隐藏层进行非线性变换, 变换后的窗口向量为:

$$f_{\theta 3}(M) = g(W_3(g(W_2(f_{\theta 1}(M)) + b_2)) + b_3) \quad (1)$$

上式中, $W^2 \in R^{h2\times D}$、$W^3 \in R^{h3\times h2}$ 分别是隐藏层中, 第二层和第三层的变换矩阵, $h2$、$h3$ 分别代表两层隐藏层的节点单元个数 (该参数可调节). $b^2 \in R^{h2}$、$b^3 \in R^{h3}$ 表示第二层和第三层的偏置矩阵. 采用 *ReLU* 函数作为激活函数:

$$\begin{aligned} lg(x) =& ReLU(x) \\ =& max(0, W_i \cdot f_{\theta i}(M) + b_i) \end{aligned} \tag{2}$$

1.3 输出层

对于军事武器名称识别任务, 在给定电子文本中, 利用当前词语的上下文环境, 识别该词是否为武器名称, 故设计输出层的节点个数为 4, 对应词语的四个状态标注值: $F = \{F_1, F_2, F_3, F_4\} = \{B, I, O, E\}$. F 集合中四种状态的含义为: B 代表该词语为武器名称的第一个词, I 代表武器名称的中间词, E 代表武器名称的尾部词, O 代表该词语不是武器名称.

输出层的输入来自上层隐藏层的输出, 该输入为一个 $h3$ 维的向量 z, 输出层的非线性变换为 $W_4 \cdot z + b_4$, $W_4 \in R^{h4\times h3}$ 为该层的变换矩阵, b_4 该层的偏置矩阵, $h4$ 为输出层神经单元的个数. 变换后得到一个没有归一化的 $h4$ 维向量, 本文用 *Softmax* 函数对其进行归一化处理, z_i 表示输出向量的第 i 个值:

$$soft\max(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{h4} e^{z_j}} \tag{3}$$

进行归一化处理后可得到 $P\{F_t|W_i\}$, 输出层的第 i 个值表示当前词语为状态 F_i 的概率.

1.4 标注推断

测试数据通过该模型训练后, 输出词在状态 $F_i(i = 1,2,3,4)$ 的值, 即 $p(W_i|c_{[i-2,i+2]})$, 记为词语的上下文得分. 本文考虑到词语标记和特征之间的关系在命名实体识别中很重要, 加入词状态之间的转移概率 T_{ij}, 记为状态转移得分, 表示在整个训练集中, 词从状态 F_i 到 F_j 的转移概.

命名实体识别的输出是一个状态序列标记的问题.对于句子的一种标记序列为 $w_{[1:n]}$, 在已知上下文得分和状态转移得分的情况下, 计算最高得分的标记路径 $t_{[1:n]}$ 的问题可以通过维特比 (Viterbi) 算法求解. 算法的递推关系如下.

初始化:

$$\varphi_1(j) = \mu \cdot A_{0i} + p(j|c_{[j-2,j+2]}) \tag{4}$$

递推关系:

$$\begin{aligned} \varphi_i(k) = & max_{1\le j\le 4}\{\varphi_{i-1}(j) + \mu \cdot A_{jk} \\ & + p(k|c[k-2, k+2])\}, i = 2,3,... \end{aligned} \tag{5}$$

上式中 $\varphi_i(k)$ 表示: 句子词个数为 i 且最后一个词的状态为 F_k 时, 句子词序列的最大标注概率. 考虑到实验中构造的词语状态转移的特征较粗糙, 加入常数 $\mu = 0.4$ 来减少转移得分的权重.

2 参数训练

本文模型中的参数有 $(W_2, b_2, W_3, b_3, W_4, b_4, A)$, 参数的训练是利用训练数据反复对参数进行修改的过程. 参数 θ 的更新使用常用的梯度随机下降 (stochastic gradient descent) 算法, 公式如式 (6):

$$\theta = \theta - \lambda\Delta_\theta \tag{6}$$

式中 λ 为学习率, 取其值为 0.02. Δ_θ 为下降的梯度, 参数的估计采用最大似然估计的方法, 即:

$$\Delta_\theta = \frac{\partial \lg p(F_t|F_{t-\frac{w}{2}}, ..., Ft, ..., F_{t+\frac{w}{2}})}{\partial\theta} \tag{7}$$

为了避免在训练过程中出现参数过拟合的发生, 在模型的每层激活函数加入 dropout 正则化, dropout 的参数设置为 0.2.

3 实验结果和分析

3.1 实验设置

在词的向量表示部分, 本文采用开源工具包 word2vec, 该工具实现了 Mikolov 等人提出的连续词袋 (constant bag of words) 模型[11,12]的向量表示. 该模型的训练语料来自搜狐实验室全网中文新闻数据 (SogouCA)2012 年 6 月至 7 月的语 (http://www.Sogou.com/labs/resource/ca.php), 大小共计 711MB. 使用北京大学计算语言学研究所的云分词服务对该语料进行分词后, 利用 word2vec 学习词语的向量表示, 词向量的维度为 100 维至 400 维, 步长为 60 维.

由于目前没有较权威开放的中文军事语料[13], 本文爬取环球军事网、中华网等军事网站文章共 7500 篇, 对武器名称进行标注后作为实验语料, 随机抽取其中 80%(6000 篇文章) 作为训练集, 剩下的 20% (1500 篇文章) 作为测试数据. 本实验设置 3 组实验.

实验一. 利用词向量表示模型, 对训练集进行词的向量表示, 设置词性向量维数为 15, 将其与词向量拼接作为深度神经网络模型的输入. 标注推断仅考虑词的上下文得分. 构建并训练 4 层神经网络模型, 各层神经

单元个数为 250, 150, 100, 4. 在词向量的维度训练上, 设置维度在 100 至 400 之间, 步长为 60, 观测试验结果.

实验二. 利用词向量表示模型, 对训练集进行词的向量表示. 设置词性向量维数为 15, 将其与词向量拼接作为深度神经网络模型的输入. 标注推断仅考虑词的上下文得分. 构建并训练 5 层神经网络模型, 各层神经单元个数为 250, 200, 150, 100, 4. 设置词向量的维度为 280, 观测试验结果.

实验三. 利用词向量表示模型, 对训练集进行词的向量表示. 设置词性向量维数为 15, 将其与词向量拼接作为深度神经网络模型的输入. 标注推断结合词的上下文得分和状态转移得分. 构建并训练 5 层神经网络模型, 各层神经单元个数为 250, 200, 150, 100, 4. 设置词向量的维度为 280, 观测试验结果.

3.2 实验结果

本实验以 F-1 值作为实验评判标准准确 F-1 值表示如下:

$$\mathrm{F}-1=\frac{2\times P\times R}{P+R}\times 100\% \tag{8}$$

$$P(\text{正确率})=\frac{\text{正确识别实体的个数}}{\text{总识别实体的个数}}\times 100\% \tag{9}$$

$$R(\text{召回率})=\frac{\text{正确识别实体的个数}}{\text{待识别实体的个数}}\times 100\% \tag{10}$$

对三组实验结果做如下分析.

图 2 表示词向量维数的增加, F-1 值的变化情况.在维度为 100 至 400 之间, F-1 值缓慢上升. 在维度为 280 时达到最大, 为 0.9021, 在 340 维度时, 有所下降.这说明词向量的维度不是越大越好, 它存在局部最优值, 这可能与文本长度和文本词语分布有关.

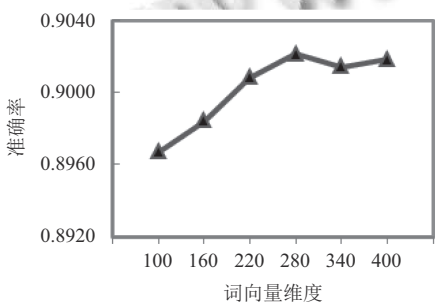

图 2 不同纬度下 F-1 值的分布

在词向量表示的最优维度 (280 维) 的情况下, 表 1 是三组不同实验情况下的 F-1 值. 试验二的 F-1 值为 0.9076, 较实验一 (280 维) 的识别效果提升了 0.609%, 说明增加一层隐层捕获了更多的特征信息. 实验三的 F-1 值为 0.9102, 较实验二的识别效果提升了 0.396%, 说明融合状态的转移得分, 可以提升命名实体的性能.

表 1 三组试验结果 F-1 值

实验序号	F-1值
实验一(280维)	0.9021
实验二	0.9076
实验三	0.9102

4 总结

我国拥有漫长的国界线和海岸线, 提升军事信息智能处理能力具有重要的战略意义. 命名实体识别作为自然语言处理的重要一环, 是军事信息化建设上的基础, 如智能问答、信息提取、舆情分析等. 本文针对军事文本中常出现的几类武器名词, 提出了基于词向量特征利用深度神经网络模型, 再融合词性和状态转移得分的特征, 在测试数据集上达到 0.9102 的识别精准度.

由于实验基于移动窗口来代表词语的前后文, 移动窗口不能捕获词语在句子中的特征. 下一步待改进的是如何捕获基于语义的特征, 以及如何减少深层网络的训练时间.

参考文献

1 McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the Seventh Conference on Natural Language Learning at Hlt-Naacl. Edmonton, Canada. 2003. 188–191.

2 Grishman R, Sundheim B. Message understanding conference-6: A brief history. Proceedings of the 16th Conference on Computational Linguistics. Copenhagen, Denmark. 1996. 466–471.

3 Bikel DM, Schwartz R, Weischedel RM. An algorithm that learns what's in a name. Machine Learning, 1999, 34(1-3): 211–231.

4 Ratnaparkhi A. A simple introduction to maximum entropy models for natural language processing. IRCS Technical Reports. Pennsylvania: University of Pennsylvania, 1997.

5 孙茂松, 黄昌宁, 高海燕, 等. 中文姓名的自动辨识. 中文信息学报, 1995, 9(2): 16–27.

6 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别. 通信学报, 2006, 27(2): 87–94.

7 姜文志, 顾佼佼, 丛林虎. CRF 与规则相结合的军事命名实体识别研究. 指挥控制与仿真, 2011, 33(4): 13–15.

8 Collobert R, Weston J, Bottou L, *et al*. Natural language processing (Almost) from scratch. The Journal of Machine Learning Research, 2011, 12(1): 2493–2537.

9 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]

10 Mansur M, Pei W, Chang B. Feature-based neural language model and chinese word segmentation. Proceedings of the 6th International Joint Conference on Natural Language Processing. Nagoya, Japan, 2013: 1271–1277.

11 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Computer Science, 2013.

12 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 2013, (26): 3111–3119.

13 冯蕴天, 张宏军, 郝文宁. 面向军事文本的命名实体识别. 计算机科学, 2015, 42(7): 15–18. [doi: 10.11896/j.issn.1002-137X.2015.07.004]