文章编号:1003-6199(2017)01-0123-05

基于深度学习的医疗命名实体识别

张 帆†,王 敏

(湖南大学 电气与信息工程学院,湖南 长沙 410082)

摘 要:在较为深入地研究医疗文本实体识别的现有方法的基础上,设计一种基于深度学习的医疗文本实体识别方法。本文在医疗文本数据集上进行实体识别对比实验,所识别目标实体包含疾病,症状,药品,治疗方法和检查五大类。实验结果表明,设计的深度神经网络模型能够很好的应用到医疗文本实体识别,本文所设计的方法比传统算法(如CRF)具有较少人工特征干预及更高的准确率和召回率等优点。

关键词:实体识别;数据挖掘;深度学习;医疗信息

中图分类号:U491.14 **文献标识码**:A DOI:10.3969/j.issn.1003-6199.2017.01.025

Medical Text Entities Recognition Method Base on Deep Learning

ZHANG Fan†, WANG Min

(Hunan University, College of Electric and Information Engineering, Changsha, Hunan 410082)

Abstract: This article do deep research on many medical text entity recognition methods, and design a medical text entity recognition method base on deep learning. This article do entity recognition controlled experiments on medical text data set, the target entity contain diseases, symptoms, medicine, treatment and inspection. The experimental result show that our neural network model can be applied to medical text entity recognition. Our neural network model have higher precision, higher recall and less artificial features than traditional methods such as CRF.

Key words: sentity recognition; data mining; deep learning; medical information

1 引 言

医疗文本实体识别是医疗知识挖掘,医疗智能机器人,医疗临床决策支持系统等应用领域的重要基础工作。最近一大批在线医疗信息,社区及远程问诊网站及其应用迅猛发展。这些网站为病患者提供多元化的医疗信息获取渠道。同时产生大量疾病问答信息与医疗文本。这些信息将汇成一股非常可观的大数据。并且这些医疗文本中有大量真实的个人案例,潜藏着丰富的医疗价值。但是这些医疗文本大多处于一种非结构化的状态。为充分挖掘其中的价值,并为接下来医疗问答等应用打好基础工作,医疗文本实体识别是必不可少的步骤。

2 相关工作

命名实体识别这个概念是在MUC-6(Message Understanding Conference)会议被提出。命名实体识别主要任务是识别出文本中出现专有名称和有意义的数量短语并加以归类。通用的命名识别主要包含实体(组织名、人名、地名),时间表达式(日期、时间),数字表达式(货币值、百分数)等。在生物医学领域,目前比较集中的研究是针对医学文献中的基因、蛋白质、药物名、组织名等相关生物命名实体识别工作[1]。随着医疗系统的信息化,也出现大量针对电子病历进行的识别工作。

收稿日期:2016-11-01
基金项目:湖南省科技厅重点研发计划(2015WK3049)
作者简介:张 帆(1963—),男,湖南长沙人,副教授,研究方向:人工智能,嵌入式系统。
†通讯联系人,E-mail:562258948@qq.com

目前常用的命名实体识别方法分为两大类:基于规则和知识的方法与基于统计的方法。基于规则和知识的方法是一种最早使用的方法,这种方法简单,便利[2]。基于规则和知识方法缺点是需要大量的人工观察,可移植性较差。基于统计的方法将命名实体识别看作一个分类问题,采用类似支持向量机,贝叶斯模型等分类方法;同时也可以将命名实体识别看作一个序列标注问题,采用隐马尔可夫链、最大熵马尔可夫链、条件随机场等机器学习序列标注模型[3-6]。这些方法都需要人依靠逻辑直觉和训练语料中的统计信息手工设计出大量特征。这些统计学习方法识别性能很大程度上依赖于特征的准确度,所以要求团队中要有语言学专家。

加拿大多伦多大学的 Hinton 教授[7]提出深度学习的概念,在全球掀起一次热潮。深度学习通过模仿人脑多层抽象机制来实现对数据(图像、语音和文本等)的抽象表达,将特征学习和分类整合到一个统一的学习框架中,从而减少手工特征制定的工作量。最近几年来,深度学习在图像识别和语音识别等领域已经取得巨大成功。深度学习技术在原始字符集上提取同样也受到很多关注。因为深度学习技术可以在原始字符集上提取高级特征,所以本文利用深度学习技术在大量未标记医疗语料上无监督地学习到词特征、不用依赖人工设计特征,从而达到实体识别的目的。

针对实体识别这一任务,本文用到神经网络语言模型对词进行分布式表达。神经网络语言模型利用神经网络对词的概率分布进行估计、生成模型,从而得到词与词之间的关系;同时该模型是一种无监督训练模型,可以从大量未标记的非结构化文本中学习出词语的分布式表示,并且可以对词语之间的关系以及相似度进行建模。

神经网络语言模型(NNLM)[8]是 2003 年由 Bengio 提出,直至近年来由于硬件成本降低、文本数量急剧增加,神经网络语言模型开始逐渐被应用到多种自然语言处理任务中,并取得了不错的效果。纵观神经网络语言模型的演变过程,同样也说一个逐步完善和逐步应用的过程。2011 年 Mikolov 等[9]使用循环神经网络改进了 Bengio 的神经网络语言模型,该模型在语音识别上的应用性能要优于传统的 n-gram 语言模型。2011 年 Collobert 等[10]提出了一个统一的神经网络架构及其学习算法,并设计了 SENNA 系统可用于解决语言建模、词性标记、组块分析、命名实体识别、语义角色标记和句法分析等问题。2013 年 Zheng 等[11]在大规模未标记数据集上改进了中文词语的内在表示形式,并使用深度学习模型发现词语的深层特征以解决中文分词和词性标记问题。2016 年 Z Jiang 等[12]提出一种基于图的词向量表达,并将其应用到医疗文本挖掘中。2016 年 SR Gangireddy 等[13]提出一种自适应的 RNN 神经网络语言模型,并将其用到自然语音识别上。本文在前人研究基础上,利用神经网络语言模型构建了词的分布式特征,从而使医疗词汇的命名实体识别更加具有可应用价值。

3 算法模型设计

本文设计一种可以用于命名实体识别的深层神经网络架构,该架构的本质是构建具有多层的神经网络,学习出更有用的特征,从而提升识别的性能。比自然语言处理任务中常用模型如:条件随机场模型,SVM,贝叶斯模型,该架构具有两大优势:1. 传统的稀疏特征被稠密的分布式特征取代;2. 利用深度学习结构以发现更高级的特征。

3.1 命名实体识别的深层架构

本文的神经网络至少包含三层,第一层是输入层,第二层是隐含层,第三层是输出层。

该深层网络的输入是词分布式表达,输入的词向量也需要训练和优化模型参数;隐含层可以有多层,本文为提高训练速度,使用单层作为隐含层;输出层采用损失函数为二元交叉熵的逻辑分类器构成。

该架构主要思路是将实体识别看作一个分类问题。其输入是词向量表达与上下文词汇的词向量。这些词向量替代了传统机器学习方法人工定义的特征,将这些词向量输入到神经网络,然后通过隐含层将这些词向量转换为另外向量,再通过逻辑回归层进行分类,得到每个词的实体名概率,从而完成此实体识别工作(如图 1 所示)。

3.2 分布式表示

上文提到神经网络的输入是词向量。

对词特征和词性特征进行传统的特征表示,那么任意两个词语之间或者任意两个词性标记之间都是孤立的、没有联系的。对词特征和词性特征进行分布式表示,即把每个词语或者每个词性标记都表示为一个低维实数向量,那么任意两个词语之间或者任意两个词性标记之间的欧氏距离将更近。

词语特征的分布式表示可解决机器学习中的维数灾难和局部泛化限制等问题,相比于传统的特

征表示方式可以更深入地探索输入数据之间的固有联系,捕获其内部的语法、语义相似性。当遇到训练语料中未出现的词语或词性标记时,采用词语特征的分布式表达训练出的模型仍然能够有较好的表现。

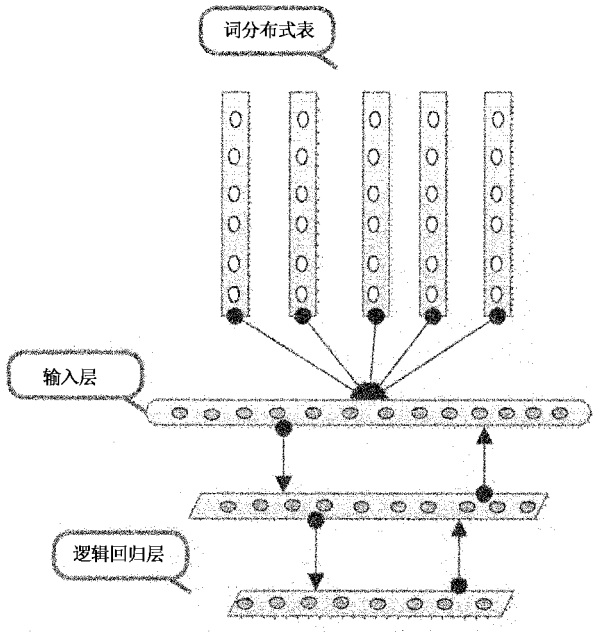

图 1　用于命名实体识别的神经网络架构图

3.3　前馈神经网络函数

一个词的实体识别需要考虑该词的上下文环境,这样识别准确度才能更高。本文神经网络输入层是窗口词向量,而不只是单个词的词向量。定义窗口大小为 C,当 C=1 时则表示输入是一个词向量。隐含层的输入是窗口词向量,是一个 C * M 的矩阵。C 为窗口大小,M 为词向量的维度。隐含层的输出作为逻辑回归层的特征。逻辑回归层将计算窗口的中心词为各个类别的概率。故本文网络架构的前馈神经网络函数如下:

$$z = W_{(x^1,x_2)} + b^1 \quad (1)$$

$$a = f(z) \quad (2)$$

$$h = g(U^T a + b^2) \quad (3)$$

向量 $_{(x^1,x_2)}$ 是输入窗口词向量,a 为隐含层的输出值,z 为输入的词窗口向量进行线性变换之后的特征向量。模型参数为 W,U,模型函数为 f 跟 g,b 为线性变换的标量,h 为每个类别的概率值。f 是激活函数,可以采用 sigmod 函数或者双曲正切函数。本文选用双曲正切函数(因为双曲正切函数的导数能被表达为双曲正切函数)。

$$\frac{d}{dx}tanhx = 1 - tanh^2 x \quad (4)$$

g 需要返回一个概率值,所以本文选用 sigmod 函数。

$$g(z) = sigmoid(z) = \frac{1}{1+e^{-z}} \quad (5)$$

下面,本文将解释该深度网络整体流程:

输入训练样本是由一系列输入对 (x^i, y^i) 组成。i 的范围从 1 到训练语料的总长度。每个 x^i 都是窗口词向量,$x^i = [x_{i-1}, x_i, x_{i+1}]$,$y^i$ 是实体概率,范围是从 0 − 1。定义 θ 为整个网络的参数,$\theta = (W, b^{(1)}, U, b^{(2)})$。通过这个定义,本文可以定义整个网络函数为:

$$h_\theta(x^i) = g(U^T f(W_{(x^1,x_2)} + b^1) + b^2) \quad (6)$$

3.4　随机化参数

在网络训练过程中,首先需要对参数进行随机化赋值。最有效的参数随机化策略是从范围域 $[-\in, \in]$ 随机选择数值为参数 W 赋值。本文利用每层神经网络的输入神经元数与输出神经元数计算得到 $\in$ 值。

$$\in = \frac{\sqrt{6}}{\sqrt{fanIn + fanOut}} \quad (7)$$

其中 $fanIn$ 是本层神经网络输入神经元个数,$fanOut$ 是本层神经网络输出神经元个数。

3.5　损失函数

在逻辑回归中,可以通过最大似然估计方法来计算参数,计算公式如下:

$$\gamma(\theta) = \log\prod_{i=1}^{m} p(y^{(i)} \mid x^i, \theta) =$$

$$\log\prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^i} (1 - h_\theta(x^{(i)}))^{1-y^i} =$$

$$\sum_{i=1}^{m} y^i \log h_\theta(x^i) + (1 - y^i)\log(1 - h_{(\theta)}(x^i)) \quad (8)$$

一般本文可以通过计算负最大似然函数的最小值来代替计算最大似然函数的最大值。神经网络的最终损失函数为:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^i \log h_\theta(x^i) - (1 - y^i)\log(1 - h_{(\theta)}(x^i))] \quad (9)$$

3.6　参数训练

对该深层架构的训练本质是在训练语料上计算模型中的未知参数,未知参数主要包括隐含层的若干参数,还包含逻辑回归层中的变换矩阵 $W \in R^{y\times n}$ 和偏置矩阵 $b \in R^{y\times n}$。训练神经网络需要用到反向传播算法和 SGD(随机梯度下降)算法。具体参数训练流程为:

第一步:随机初始化网络全部参数,包含隐含层、逻辑回归层参数。

第二步:随机挑选一个训练样本 (x^i, y^i),首先

进行前向传播，将隐含层的输出信息传递到逻辑回归层，将所提取的最高级特征映射到相应的标记信息上，利用数据的标记值对模型进行有监督训练，并不断调整连接权值，减小模型的目标预测标记与实际标记之间的概率误差。

第三步：反向传播，计算前向传播过程中目标预测标记与实际标记之间的概念误差，并将该误差从逻辑回归层向隐含层传播，并不断调整隐含层参数 $\theta=(W,b^{(i)})$。

4 医疗文本实体识别流程

针对在线医疗文本信息，本文主要考虑了 5 类命名实体：疾病、症状、药品、治疗方法和检查。具体实体识别流程如图 2 所示，主要包括数据爬取、数据处理、数据处理、词汇分布式特征训练、神经网络模型训练、实体识别和识别结果抽取。首先爬取胃癌、糖尿病、哮喘、高血压四种病相关医疗文本，对获取的医疗文本进行预处理，包括特殊符号的过滤、人工标注、分词、大小写转化等操作，然后，利用程序将所有数据划分为训练集和测试集两部分。将训练集放到模型中进行训练，随后再利用训练得到的参数测试模型识别效果。

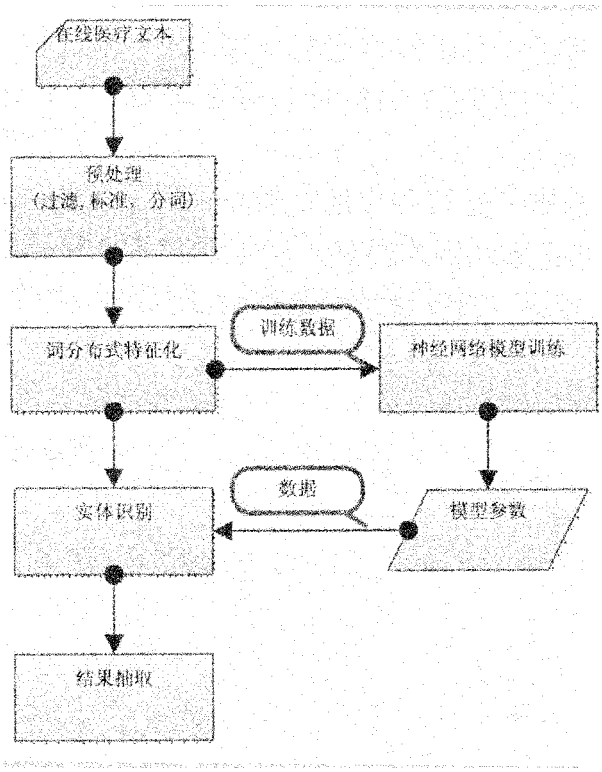

图 2 医疗命名实体识别整体流程

5 实验结果及分析

5.1 实验条件

本文在 Centos 系统环境下用 Java 实现相关代码，完成整个模型的构建与训练。其中使用一款开源工具包 word2vec 构建神经网络语言模型，word2vec 是 Tomas Mikolov 在 2013 年开开发出来的工具包。word2vec 使用 CBOW 模型（连续词袋模型）[14－16]。CBOW 模型是一种简化的 NNLM 模型，CBOW 去掉了最耗时的非线性隐层、且所有词共享隐层，可无监督地训练出词特征的分布式表示和词性特征的分布式表示。为验证本文算法效果，本文通过设置 2 组对比实验进行验证，两组对比实验如下：

实验 1 通过观察分析训练语料，手工构建特征集。这些特征集有符号特征，词性特征，形态特征，后缀特征，身体部位指示词特征与上下文特征等。在训练语料上使用这些特征集训练条件随机场模型，并利用得到的条件随机场模型在测试语料上进行命名实体识别，然后对识别结果进行评估，将实验标记为 CRF。

实验 2 在训练语料上无监督地学习出词的分布式表达和词性的分布式特征表达，并利用词的分布式表达和词性的分布式表达构建并训练 3 层网络架构。然后利用训练出来的深度神经网络在测试语料上进行命名实体识别，且对识别结果进行评估，将实验标记为 DBN。

5.2 实验结果

本实验使用 3 个指标来衡量命名实体识别的性能：正确率、召回率和 F 值。其计算公式如下：

$$\text{正确率}(P)=\frac{\text{系统正确识别的实体个数}}{\text{系统识别的实体个数}}\times 100\% \tag{10}$$

$$\text{召回率}(P)=\frac{\text{系统正确识别的实体个数}}{\text{文档中实体个数}}\times 100\% \tag{11}$$

$$\text{F－值}=\frac{2\times P\times R}{P+R}\times 100\% \tag{12}$$

对 2 组对比实验的结果进行正确率，召回率和 F－值的计算。CRF 实验构建条件随机场模型用于命名实体识别任务，其平均正确率、召回率、F－值分别为 79.61%、79.62%、79.61%。实验 DBN 构建深度学习网络架构用于实体识别任务，其平均正确率、召回率、F－值分别为 88.03%、82.34%、85.08%，相比于实验 CRF 分别提高了 8.42%、

2.72 %、5.47%。具体每个类别实验结果见表 1。

表 1　实验结果

实验	实体类型	正确率(%)	召回率(%)	F-值(%)
CRF	疾病	80.23	82.23	81.21
	症状	79.86	79.86	79.86
	药品	82.23	78.53	80.33
	治疗方法	77.74	77.96	77.84
	检查	78.02	79.53	78.76
DBN	疾病	90.12	85.17	87.57
	症状	86.32	81.81	84.00
	药品	89.42	80.01	84.45
	治疗方法	89.80	85.03	87.34
	检查	84.51	79.68	82.02

6　结　论

本文通过神经网络语言模型学习得到词特征的分布式表达和词性特征的分布式表达。并在词分布式表达基础上构建出一种深层架构,将该深层架构应用于医疗命名实体识别任务。实验表明该方法可以自动抽象出更高级特征,最大程度减少手工特征设计工作量。在医疗语料库上进行 2 组对比实验,取得总体上 88.03%的准确率和 82.34%的召回率,该实验结果表明该方法在命名实体识别任务中比条件随机场模型效果更好。

参考文献

[1] 胡双,陆涛,胡建华.文本挖掘技术在药物研究中的应用[J].医学信息学杂志, 2013,(8):49-53.

[2] 周昆. 基于规则的命名实体识别研究[D]. 合肥:合肥工业大学, 2010

[3] 阚琪. 基于条件随机场的命名实体识别及实体关系识别的研究与应用[D]. 北京:北京交通大学, 2015.

[4] 冯元勇,孙乐,张大鲲,等. 基于小规模尾字特征的中文命名实体识别研究[J]. 电子学报,2008,36(9): 1883-1838.

[5] 钟志农,刘方驰,吴烨,等. 主动学习与自学习的中文命名实体识别[J]. 国防科技大学学报,2014,4:82-88.

[6] 怀宝兴,宝腾飞,祝恒书,等. 一种基于概率主题模型的命名实体链接方法[J]. 软件学报,2014,9: 2076-2087.

[7] HINTON G E,SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[8] BENGIO Y, DUNCHARME R, VINCENT P,*et al*. A neural probabilistic language model [J]. The Journal of Machine Learning Research, 2003,3:1137-1155.

[9] MIKKOLOV T, KOMBRINK S, BURGET L,*et al*. Extensions of recurrent neural network language model[C] // 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE,2011: 5528-5331.

[10] COLLOBERT R, WESTON J, BOTTOU L,*et al*. Natural language processing (almost) from scratch[J]. The Journal of machine Learning Research, 2011, 12: 2493-2537.

[11] ZHENG Xiao-qing, CHEN Han-yang, XU Tia-yu. Deep Learning for chinese Word segmentation and POS Tagging [C] //EMNLP. 2013: 647-657.

[12] JIANG Zhen-chao,LI Li-shuang,HUANG De-gen. An Unsupervised Graph Based Continuous Word Representation Method for Biomedical Text Mining// IEEE/ACM Transactions on Computational Biology and Bioinformatics,2016, 13:634-642.

[13] GANGIREDDY S R,SWIETOJANSKI P,BELL P,*et al*. Unsupervised Adaptation of Recurrent Neural Network Language Models// Interspeech, 2016, 9: 2016-1342

[14] MIKOLOV T,CHEN K, CORRADO G,*et al*. Efficient estimation of word representations in vector space [J]. Neural Computation,2014, 14: 1771-1800.

[15] MIKOLOV T, SUTSKKEVER I, CHEN K,*et al*. Distributed representations of words and phrases and their compositionality[C] // Advances in Neural information Processing Systems. 2013: 3111-3119.

[16] ALEXEYBORISOV T K,MAARTEN DE R S CBOW: Optimizing Word Embeddings for Sentence Representations[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 10. 18653/v1/P16-1089.