

基于 Word2vec 与 BP 神经网络的病历症状自动分类研究^{*}

叶 辉 卓奕荣 肖志毅 吴瑞源

(广州中医药大学 广州 510016)

〔摘要〕 分别运用 Word2vec 和 BP 神经网络算法以及两者混合对 300 份长病历数据进行症状信息的分类抽取,旨在解决医学病历症状信息自动分类问题,结果表明采用两种方法混合处理能够得到更好的效果。

〔关键词〕 BP 神经网络; Word2vec; 文本分类; 病历症状信息

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2018.11.013

Study on the Automatic Classification of Symptoms in Medical Records Based on Word2vec and BP Neural Network YE Hui, ZHUO Yirong, XIAO Zhiyi, WU Ruiyuan, Guangzhou University of Chinese Medicine, Guangzhou 510016, China

〔Abstract〕 By making use of Word2vec, BP neural network algorithm and the method mixing the both, respectively, the classification extraction on data of 300 long medical records based on data of symptoms has been implemented, which aims to solve the automatic classification of data of symptoms in medical records. Results show that effect of mixing the two methods in process is better.

〔Keywords〕 BP neural network; Word2vec; text classification; symptom information of medical records

1 引言

文本病历信息抽取属于自然语言处理中信息抽取子领域。自然语言处理可以分成 3 个层级,分别是文本理解、信息抽取和信息检索。在以往病历电

子化过程中电子病历通常只是简单地记录患者所有文字信息并将其保存在文本文档中,通常包含大量长篇无序的文本信息,这种电子病历称为非结构化文本病历,见表 1。当前面临的问题是医疗机构无法对以往的非结构化电子病历进行数据分析和挖掘,从而浪费数据资源,采用人工整理非结构化数据则工作量巨大且效率低下,因此研究一种针对非结构化电子病历文本信息自动分类提取的方法尤为重要。本文尝试先研究出针对非结构化电子病历信息中某种特定字段(如症状)进行自动分类提取的方法,从而推广到对所有字段都能有效提取的方法,将来可对自动分类提取形成的文本信息再进行命名实体识别、医疗关系识别等信息抽取,逐渐形成更深层的结构化电子病历。

〔收稿日期〕 2018-09-04

〔作者简介〕 叶辉,讲师,发表论文 12 篇。

〔基金项目〕 2017 国家重点研发计划“云计算和大数据”重点专项(项目编号:2017YFB1002300); 2016 广东省高水平大学建设青年创新人才项目(项目编号:2016KQNCX024)。

表 1 非结构化病历混杂信息对应相关含义分析

病历文本信息	医疗相关含义
患者主诉肠梗阻，患者上周肠梗阻又复发	症状描述
脉沉滑，微弦，舌尖红，苔中度腻	查体情况
调肝胃肠，理气健脾，兼治月经病	治疗方案
单角子宫，卵巢囊肿切除术	既往病史

2 相关研究

浙江大学的李莹^[1]以家族史病历信息结构化为例，运用条件随机场（Conditional Random Field, CRF）进行命名实体识别，共有 3 部分工作，即人物和医学问题的识别、人物和医学问题的关联以及抽取后的人物和医学问题的表示。条件随机场由美国 Lafferty 等人提出，结合最大熵模型和隐马尔可夫模型的特点，是一种无向图模型，曲春燕^[2]实现基于 3 种有监督学习方法的病历命名实体识别系统并引入扩展特征与组合分类器算法对系统性能进行提升。崔园^[3]展示一种新的基于短语匹配的中文电子病历分类算法。该方法能够充分地运用信息量更大的短语匹配而不是独立的字词匹配，保留文档结构信息。通过用向量空间表示匹配的短语，每个病历记录被表示为 1 个向量，向量中每个元素表示 1 个短语在记录中出现的次数。所有向量组成病历数据集将其作为自组织神经网络的输入数据。但以上，研究一般只关注词汇和关系的抽取，在输入大量混杂的长文本时效果可能不理想，而且对此类信息使用上分类可能产生混乱。因此本文首先研究如何在混杂的文本信息中分类出病历症状文本信息。

3 研究方法

3.1 利用 Word2vec 训练词向量

3.1.1 Word2vec 模型 病历症状抽取属于自然语言处理（Natural Language Processing, NLP）问题，用计算机处理 NLP 问题首先需要将自然语言以计算机能理解的形式输入，而自然语言中有意义的最小单位是词。Word2vec 是一种快速有效地训练词向量模型的方法。该模型目的在于从大量文档数据中学习高质量的词向量，经检验通过 Word2vec 训练出来的词向量在语义和语法的近义词分辨上均有

很好的表现。Word2vec 训练过程，见图 1。Word2vec 模型的效果与语料库关系密切，因区别于一般的 NLP 任务，本文研究内容属于医学领域，故仅含日常用语的语料库难以满足需求。本次研究对北京某医院提供的 300 份长病历数据进行预处理，每份病历记录此患者多次看病情况并构建病历语料库。构建语料库时对属于症状的词标记为“ZZ”，不属于症状的词标记为“N”。如“胸痛”为 ZZ，“费用”为 N。Word2vec 包括两种训练模型，分别为 CBOW 和 Skip_gram^[4]，本文采用 Skip_gram 模型，使用该特征词来预测其上下文的词。

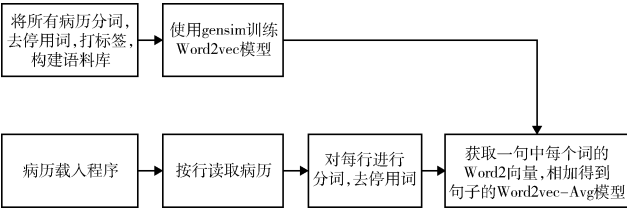


图 1 Word2vec 训练症状词向量

3.1.2 词向量训练 本次研究选取 300 份长病历数据进行预处理并构建病历语料库。对所有病历的症状项进行分词，去掉标点符号，进行 Word2vec 词向量训练，训练方法采用 Skip_gram，构建维度为 50 的词向量。训练时将每个句子分词，得出每个词向量后再采用 Word2vec - Avg 模型如下所示：

$$Vec(d) = \sum Wt \div N(d)$$

其中 Wt 代表特征词 t 的词向量，N(d) 表示句子 d 的词数，将每个词的向量相加，看做是每个句子的输入，由于每个词语的 Word2vec 词向量为 50 维，每个句子的输入形式也是 50 维的向量。如语料中出现“胸闷”，经过 Word2vec 训练后得到结果，见图 2。如病历中出现“胸闷气短，头晕，倦怠乏力，口干口苦”，经词向量叠加处理后的 Word2vec 表示，见图 3。

[0.17612274 0.41569802 -0.00271717 0.19015966 0.12620656 0.41469917
0.17191374 -0.14107068 0.3527513 0.28661543 -0.43499714 0.3358684
0.15915449 -0.19935423 0.48917794 -0.03673224 -0.01307207 0.26057363
0.04506967 -0.07426202 -0.08434106 0.24548465 -0.48532188 -0.3445684
0.3146759 -0.25184676 0.06827361 -0.3018324 0.28817815 -0.06188123
0.34987947 0.3370362 0.19434667 -0.35100338 -0.23151952 -0.517401
-0.1474834 0.05269756 0.0953 0.30382124 0.16035742 0.16858512
0.36434844 0.012738 -0.6431712 -0.08127119 -0.492918 0.14721267
0.07907856 -0.20503876]

图 2 “胸闷”的 Word2vec 词向量

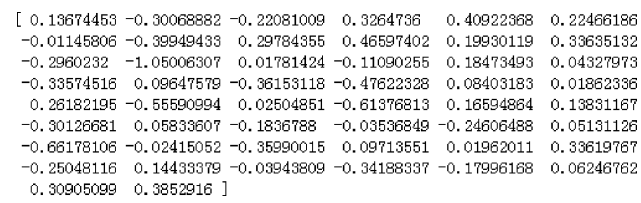


图3 句子的 Word2vec 表示

3.2 词向量输入 BP 神经网络进行模型训练

反向传播（Back Propagation，BP）是一种按误差逆传播算法训练的多层前馈网络，能学习和存贮大量的输入－输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。其学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小^[5]。对训练集的病历文本每一句单独处理，对 BP 神经网络设置 50 个输入神经元，每个神经元对应句子向量的 1 个维度，以其作为输入对神经网络进行训练，见图 4。神经网络采用 Sigmoid 函数激活，由于是二分类问题，输出层也采用 Sigmoid 函数，输出属于症状的句子标签为 1，不属于症状的句子标签为 0。

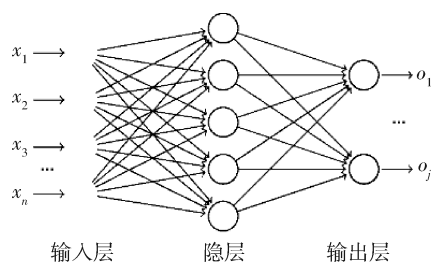


图4 Word2vec 训练后作为输入层输入给 BP 神经网络

4 实验设计

4.1 数据处理

将病历进行预处理，90% 的病历作为训练集，10% 作为测试集，其中属于症状的句子标签为 1，不属于症状的句子标签为 0。

4.2 基于 Word2vec 词向量计算句子相似度方法

此方法以下简称 w2v 方法，即使用 Word2vec 模型表示待判断是否属于症状的句子，表示结果仍

为 50 维的向量，而在训练 Word2vec 模型时为每个属于症状的词都打上“ZZ”的标签，不属于症状的词打上“N”的标签，所以“N”和“ZZ”会有其 Word2vec 词向量且维度也为 50 维。根据 Word2vec 训练原理，“N”和“ZZ”都会带上症状和非症状的语义，所以可以将计算句子相似度的问题转换成两个向量的计算问题。使用上文训练的 50 维 Word2vec 词向量对测试集进行验证。参考 NLP 中常用的句子相似度方法，其中的句子表示模型采用 Word2vec－Avg 模型。实验证明该表示方式不会造成维数灾难且效果理想。句子相似度则采用余弦相似度的方法，因已经为症状的词打上“ZZ”标签，为非症状的词打上“N”标签，故只需要计算每句 Word2vec－Avg 模型以及“ZZ”与“N”的 Word2vec 词向量相似度即可，公式如下：

cosθ = (Σ(Ai × Bi)) / (sqrt(ΣAi²) × sqrt(ΣBi²))

若句子和“ZZ”的相似度高于“N”的相似度则认为句子属于症状，反之则不属于。本研究采用 Word2vec 训练方法，采用 Skip－gram，Windows 大小取 3，Word2vec 向量维度取 50，采用 Hierarchical Softmax 优化模型，Min－count 设为 1。对句子“胸闷气短，头晕，倦怠乏力，口干口苦”训练后计算的相似度输出，见表 2。

表2 Word2vec 判断属于症状的近义句子

句子对标签距离	计算值	是否症状
对 ZZ 标签的距离	0.7557123	是
对 N 标签的距离	0.0000145	否

4.3 以 one－hot－Representation 做输入的神经网络方法

此方法以下简称 BP 方法，输入是每个句子的 one－hot－Representation 向量，对所有病历的“特征”项进行分词，去掉标点符号，形成词袋，取出词袋中频率最高的前 800 个词作为输入。在训练集中如该句含 800 个词中的词，相对应的输入就标记为“1”，否则标记为“0”。如前 3 个词语为

“双下肢”、“无力”、“抽搐”，句子中仅包含“抽搐”，则输入为 [0, 0, 1]。

4.4 以 Word2vec 做输入的 BP 神经网络方法

此方法以下简称 w2v + BP，以每个句子的 Word2vec 向量表示作为 BP 神经网络输入层输入，属于症状的句子标签为 1，不属于则标签为 0，使用训练集对网络进行训练，训练得到网络的权重和偏置后对测试集进行测试。

5 结果与分析

实验 1 使用基于 Word2vec 的句子相似度计算法进行病历症状抽取，实验 2 对用 one-hot-Representation 做输入的 BP 神经网络对测试集进行验证，实验 3 使用以 Word2vec 词向量做输入的 BP 神经网络对测试集进行验证，具体方法比较，见表 3。

表 3 单一算法和混合算法比较 (%)

方法	P	R	F
w2v	58.4	79.1	67.2
BP	87.0	71.4	78.4
w2v + BP	71.7	96.4	81.8

可见使用 Word2vec 做输入的 BP 神经网络对病历混杂的信息中症状分类效果较优。对于病历中症状的分类，甚至其他信息的分类，使用以 Word2vec

做输入的 BP 神经网络效果较单一使用神经网络效果更好。

6 结语

本文使用 300 份长篇病历构建语料库，先训练 Word2vec 词向量，然后构建 BP 神经网络，以 Word2vec 词向量作为输入，使用 BP 神经网络进行病历中症状的分类。经测试以 Word2vec 词向量作为输入的 BP 神经网络的病历医学症状分类效果好，优于单一计算句子相似度的方法和单独使用 BP 神经网络的方法。本研究为将来医疗文本信息分类和进一步结构化病历信息提供一种可行的解决方法。

参考文献

1 李莹. 文本病历信息抽取方法研究 [D]. 杭州: 浙江大学, 2009.

2 曲春燕. 中文电子病历命名实体识别研究 [D]. 哈尔滨: 哈尔滨工业大学, 2015.

3 崔园. 数据挖掘在中文病历分类中的应用 [J]. 计算机与数字工程, 2011, 39 (3): 160-163.

4 刘建平. word2vec 原理 CBOW 与 Skip-Gram 模型基础 [EB/OL]. [2018-07-15]. <https://www.cnblogs.com/pinard/p/7160330.html>.

5 王雪丽, 宋启祥. 基于 BP 神经网络的文本分类算法研究与设计 [J]. 通化师范学院学报, 2018 (2): 70-73.

关于《医学信息学杂志》启用
“科技期刊学术不端文献检测系统”的启事

为了提高编辑部对于学术不端文献的辨别能力，端正学风，维护作者权益，《医学信息学杂志》已正式启用“科技期刊学术不端文献检测系统”，对来稿进行逐篇检查。该系统以《中国学术文献网络出版总库》为全文比对数据库，可检测抄袭与剽窃、伪造、篡改、不当署名、一稿多投等学术不端文献。如查出作者所投稿件存在上述学术不端行为，本刊将立即做退稿处理并予以警告。希望广大作者在论文撰写中保持严谨、谨慎、端正的态度，自觉抵制任何有损学术声誉的行为。

《医学信息学杂志》编辑部