基于深度学习的用户健康词表构建方法研究*

尹延鑫 李传富

（安徽中医药大学 合肥 230012） （安徽中医药大学第一附属医院 合肥 230012）

〔摘要〕 对基于深度学习的用户健康词表构建方法进行系统探索，阐述实验工具、过程及结果，对利用深度学习方法构建用户健康词表的适用性、结果影响因素进行探讨，指出该方法具有较高适用性但尚待提高应用成熟度。

〔关键词〕 深度学习；文本挖掘；用户健康词表；Word2vec

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.08.006

Study on the Construction Method of Consumer Health Vocabulary Based on Deep Learning *YIN Yanxin, Anhui University of Chinese Medicine, Hefei 230012, China; LI Chuanfu, The First Affiliated Hospital of Anhui University of Chinese Medicine, Hefei 230012, China*

〔**Abstract**〕 The paper makes a systematic exploration of the construction method of Consumer Health Vocabulary (CHV) based on Deep Learning (DL), elaborates the experimental tools, process and results, discusses the applicability and results influencing factors of constructing CHV by using DL method, and points out that this method has high applicability but still needs to improve its application maturity.

〔**Keywords**〕 Deep Learning (DL); text mining; Consumer Health Vocabulary (CHV); Word2vec

1 引言

随着物质生活水平不断提高，公众的健康重视程度和健康信息需求持续提升。大数据背景下互联网成为获得医疗信息的重要渠道。根据中国科学技术协会2019 年第1 季度发布的《中国网民科普需求搜索行为报告》[1]，“健康与医疗”主题在全部参与分析主题中占比35.63%，位居第2。随着计算机及互联网应用发展，健康信息交流活动遭遇新困难。主要体现在健康信息用户[2]信息搜索和信息内容理解的问题和阻碍，用户使用信息与源信息之间存在表达差异。在信息获取过程中该差异表现为医生与患者之间沟通不畅，以及患者使用检索系统查询相关医疗知识时实际检索结果与预期检索结果存在出入。目前的检索语言体系无法满足社会需要，造成用户对专业医学术语理解与使用困难、检索系统无法理解用户表达的信息等交流障碍。用户健康

〔收稿日期〕 2019－12－13

〔作者简介〕 尹延鑫，硕士研究生；通讯作者：李传富，主任医师。

〔基金项目〕 合肥市自主创新政策“借转补”资金项目（医疗卫生项目）“基于深度学习的肺癌智能诊断系统研制与应用示范”（项目编号：YW201710120004）。

词表（Consumer Health Vocabulary，CHV）[3]可辅助实现健康信息用户与医生、检索系统之间的良性信息互动[4]。

2 研究方法

2.1 建立语料库

可利用八爪鱼搜集器对医药健康网站部分用户健康用词进行搜集，以此作为语料库来源；通过用户问卷调查等方式收集用词建立语料库。

2.2 中文分词

利用 Jieba 分词工具通过神经网络理论[5]模拟人脑词汇处理过程对原始语料进行分词，过滤并确定用户健康用词有效词。

2.3 词向量训练[6]

利用 Word2vec 工具通过选择合适语言模型对分词结果进行词向量训练，得到用户健康用词的词向量模型。完成构建后使用词向量模型通过余弦相似度与专业医疗健康用词建立对应联系（实验中所有词汇处理工具的调用均通过 Python 代码实现）。

3 实验

3.1 实验工具

3.1.1 语料搜集器 使用八爪鱼采集器（爬虫工具），选定采集模式，输入目标语料的数据来源网址、要采集页面元素并对其设定采集要求（如采集文本、采集链接、循环点击等），实现数据全自动采集。

3.1.2 分词工具 Python 编辑器的 Jieba 分词工具是深度学习方法在自然语言处理领域的实践之一。通过 Python 调用 Jieba 分词工具包以添加自定义词和自定义词库实现不同需求的分词要求。Jieba 分词工具包含全模式、精准模式、搜索引擎 3 种分词模式，各具优势，可视具体需要选择。Jieba 分词工具还可实现关键词提取、根据词汇出现频率排序、标注词性、合并同义词等功能。本实验选用精准模式分词，分词要求仅限于实现原始语料的基本分词、去除停用词干扰。

3.1.3 模型训练工具 本次实验选择在 Anaconda 的 Jupyter Notebook 中运行 Python3 代码，调用 Gensim 工具包中的 Word2vec。

3.2 实验过程

3.2.1 语料抓取 使用八爪鱼采集器对 39 问医生-39 健康网频道内科模块中的用户健康用词进行搜集，作为实验语料。通过对网页内提问语句元素进行自动爬取获得“地中海贫血是什么原因”、“坐太久了头晕想吐是贫血吗”、“舌头发白不知道怎么回事”等 4 810 条有关内科健康的提问语句，以疑问句为主，主要构成为用户症状描述 + 病情提问。

3.2.2 分词处理 将爬取的用户健康用词以文本文件（Text File，TXT）格式保存，使用 Jupyter Notebook 调用 Jieba 分词工具分别上传语料、停用词表与《内科医学名词中英文对照表》。输入分词代码对原始语料进行初步中文分词处理，见图 1。经过 Jieba 分词处理，原始语料库中语句划割成若干个独立词汇，如“贫血、请问、地中海、原因”；“体能、贫血、头晕、月经、头痛”；“太久、贫血、头晕”；“产后检查、缺铁性、贫血、呼吸”等。经筛查发现分词结果与预期实验用分词文本存在一定差别，具体表现为：（1）存在误差词。除了医药健康方面词汇外，存在包括语气词、形容词和地名等与实验不相关的误差词。（2）专业名词误分。如将斯利安叶酸片分成“斯利安”和“叶酸片”，在中文表述中“斯利安”和“叶酸片”可指代同一样药物，而“斯利安叶酸片”是该药物的标准名称。

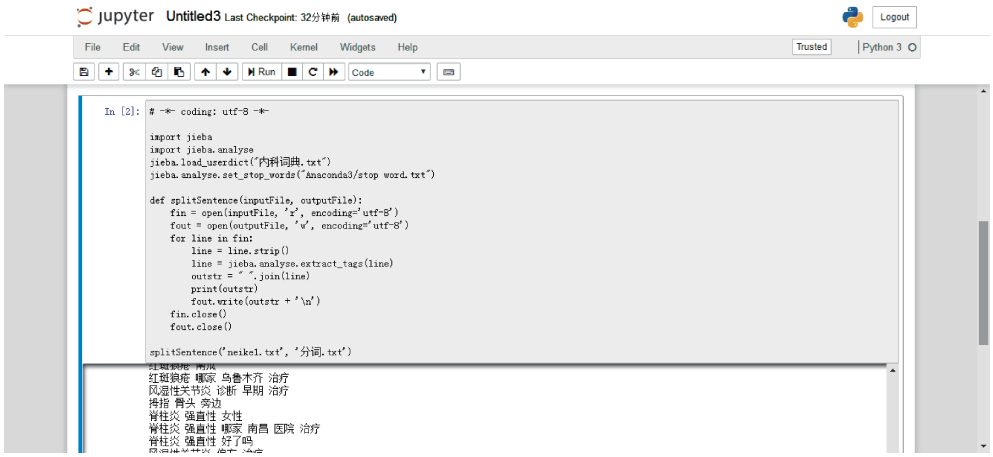

图1 Jieba 分词代码及部分分词结果

3.2.3 模型训练 经过 Jieba 分词的分词语料可作为 Word2vec 的输入文本训练词向量，训练代码运行成功后可得到有关内科用户健康词表的词向量模型。编写代码并设置 Word2vec 内词向量模型的训练参数，然后开始训练。由于模型训练的目的是测定指定医学用词出现概率，选择连续词袋（Continuous Bag－of－Words，CBOW）算法更合适，未对算法进行指定设置，而是默认算法选择，即 sg = 0。由于本次实验中的语料不够充分，为避免因参数设置过大导致部分词汇被过滤遗漏，尽可能保证对所有词汇都进行训练处理，故参数设置均较小，见图2、图3。

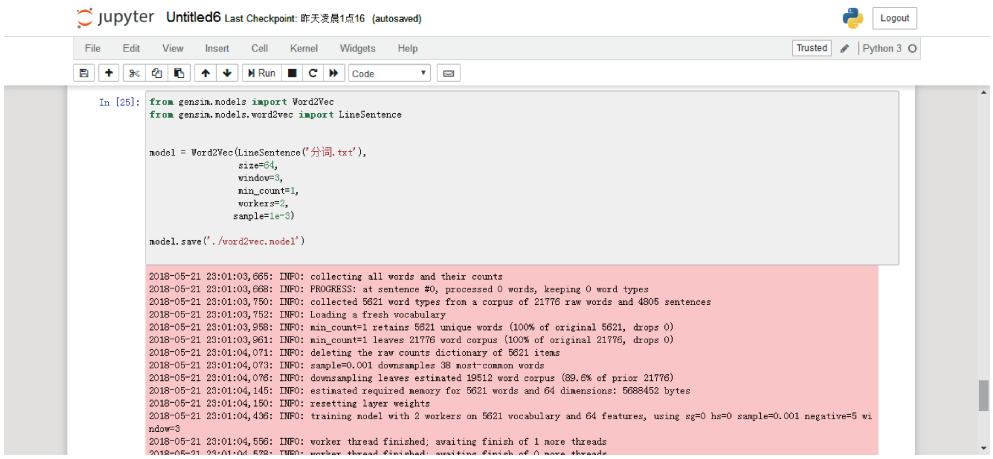

图3 部分模型训练结果

3.3 实验结果

于《内科医学名词中英文对照表》中选取 100 个中文内科医学名词作为种子词，在 Python 中调用在上一实验步骤中已设置并训练好的 Word2vec 的词向量模型文件，计算在种子词中有无与模型中相近似的词并按照相似性倒序排列，以此为依据得到种子词在用户健康用词中的同义词。如通过运行代码输入“呼吸困难”，在词向量中共有 10 个返回项，分别为“呼吸”——0.518 497 765 064 239 5、“身体”——0.500 917 851 924 896 2、“患者”——0.493 879 109 621 048、“胸闷”——0.485 214 829 444 885 25、“老想”——0.480 314 433 574 676 5、“一点”——0.480 262 249 708 175 66、“头晕”——0.470 977 306 365 966 8、“昨晚”——0.466 171 026 229 858 4、“气短”——0.459 644 794 464 111 33、“早上”——0.457 675 486 803 054 8，见图 4。根据余弦相似度排序判断该模型中与种子词最接近的为“呼吸”。通过测试，发现 100 个内科医学名词中只有 41 个词语存在返回值且通过人工审查，使用模型得到的种子词的相似词大部分余弦相似度都低于 0.5 且并不具备同义词的实际意义，见表 1。

```
In [39]: model.most_similar(["呼吸困难"])

C:\Users\Administrator\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: DeprecationWarning: Call to deprecated `most_similar` (Method w
ill be removed in 4.0.0, use self.wv.most_similar() instead).
  """Entry point for launching an IPython kernel.

Out[39]: [('呼吸', 0.5184977650642395),
          ('身体', 0.5009178519248962),
          ('患者', 0.4938791096210481),
          ('胸闷', 0.4852148294448852 5),
          ('老想', 0.4803144335746765),
          ('一点', 0.4802622497081756 6),
          ('头晕', 0.4709773063659668),
          ('昨晚', 0.4661710262298584),
          ('气短', 0.4596447944641113 3),
          ('早上', 0.4576754868030548)]
```

图 4 输入“呼吸困难”后模型生成的同义词列表及其余弦相似度

表 1 参与训练的种子词返回值及其同义词最高余弦相似度

种子词	返回值	最高余弦相似度
支气管炎	1	0.480 927 676
感染	1	0.429 849 923
咳嗽	1	0.862 440 705
肥大	1	0.512 385 011
增生	1	0.483 724 296
炎症	1	0.413 125 664
呼吸困难	1	0.518 497 765
肺气肿	1	0.401 118 129
通气	1	0.456 187 874
肺结核	1	0.427 801 4
增生	1	0.483 724 296
皮质醇	1	0.507 722 02
低血压	1	0.513 588 25
心动过速	1	0.395 836 532
醛固酮	1	0.488 989 651
水肿	1	0.604 066 372
苍白	1	0.481 152 564
心悸	1	0.400 969 088
心绞痛	1	0.499 257 982
气短	1	0.687 229 037

续表 1

白蛋白	1	0.414 371 997
分析	1	0.469 103 664
贫血	1	0.829 434 991
厌食	1	0.478 800 774
抗生素	1	0.439 013 839
抗体	1	0.428 892 553
胆红素	1	0.504 170 537
活检	1	0.420 718 789
骨髓	1	0.464 626 044
毛细血管	1	0.538 528 621
化疗	1	0.610 338 807
昏迷	1	0.506 724 179
代偿	1	0.405 768 335
并发症	1	0.451 637 685
虚弱	1	0.476 588 607
出汗	1	0.585 767 09
头痛	1	0.569 009 304
血尿	1	0.471 328 676
出血	1	0.502 587 259
疱疹	1	0.493 185 312
黏膜	1	0.485 648 751

4 讨论

4.1 深度学习方法构建用户健康词表的适用性

采用 Word2vec 训练词向量模型方法构造用户健康词表，验证了其可行性与局限性，说明可以通过深度学习理论及相关技术实现用户健康信息需求与健康信息资源之间的匹配，为用户解决部分专业医学检索需求。利用深度学习的神经网络拟合目标函数构造语言模型，可完成非医学专业用户词和医学专业术语之间的“映射”，进而实现用户健康词表在健康信息用户与医生、检索系统之间信息互动的“桥梁”工具功能，即深度学习方法在构建用户健康词表方面具备较高适用性。

4.2 用户健康词表构建结果影响因素

4.2.1 *原始语料多样性及数据规模* 深度学习理论中，所有在神经网络基础上模拟人脑信息处理的操作和模型构建，都必须以大规模正规数据为基础，即给予所构造“神经网络系统”规范且充分的学习资料，供其总结学习信息处理规律。在实验中存在大量种子词未检索到满足条件的用户健康用词，其主要原因包括以下两方面：一是前期爬取的内科相关用户健康用词语料数据规模较小，无法全面体现普通用户关于内科健康的用词习惯；二是所爬取语料资源内容形式不规范，存在类似"PCR"（Polymerase Chain Reaction，聚合酶链式反应）、"MCHC"（Mean Corpusular Hemoglobin Concerntration，红细胞平均血红蛋白浓度）等英文缩写或疾病代称以及“地贫”（地中海贫血）等中文简称，导致后期模型训练获取同义词出现误差。

4.2.2 *分词程序的分词结果* 利用 Jieba 分词程序对原始语料进行简单分词处理后，两方面原因导致分词结果未完全满足用户健康用词标准划分的分词预期：一是中文表达的多样性导致分词结果出现误分、多分（如将斯利安叶酸片分成“斯利安”和“叶酸片”）；二是分词过程中选用的停用词表覆盖范围有限，导致分词结果中保留了部分与实验所测试用户健康用词不相关的无意义的词（如“一会儿”、“晚上”等）。此外分词程序的分词结果对后期模型训练查找同义词无返回值、增加程序筛选时间产生一定影响。

5 结语

调用 Word2vec 工具包，借助词向量模型训练，以词向量模型中内科医学专业名词与非专业用户健康词表之间返回的余弦值为依据建立二者对应关系。根据实验结果中医学专业用语在用户健康用词词向量模型中余弦值的反馈，可以认为深度学习理论在用户健康词表构建方面具有较高适用性，可实现医学专业用语与用户健康用词之间“映射”关系，但存在医学专业术语无词向量模型返回值及返回值无意义等问题。说明目前深度学习技术在用户健康词表构建方面发展成熟度不足，词表构建尚未达到高度智能化、完全自动化，需人工筛选介入，该领域仍有广阔研究发展空间。

参考文献

1 中国科学技术协会. 中国网民科普需求搜索行为报告（2019 年第一季度）[EB/OL]. [2019 - 12 - 10]. http://www.cast.org.cn/art/2019/4/26/art_1281_94546.html.

2 张泰瑞，陈渝. 基于 LDA 模型因素提取的健康信息用户转移行为研究 [EB/OL]. [2019 - 12 - 12]. https://doi.org/10.13266/j.issn.0252 - 3116.2019.21.007.

3 Gu Gen, Zhang Xingting, Zhu Xingeng, et al. Development of a Consumer Health Vocabulary by Mining Health Forum Texts Based on Word Embedding: semiautomatic approach [J]. JMIR Medical Informatics, 2019, 7 (2): e12704.

4 庹兵兵. 中文用户健康词汇表构建研究 [D]. 武汉：华中科技大学，2017.

5 林奕欧，雷航，李晓瑜，等. 自然语言处理中的深度学习：方法及应用 [J]. 电子科技大学学报，2017，46 (6)：913 - 919.

6 吕建新，郑伟，马林，等. 基于词向量语义扩展的网络文本特征选择方法研究 [J]. 情报科学，2019，37 (12)：47 - 51.