基于命名实体识别和图嵌入技术的脑血管疾病相似病历研究

秦秋莉[1]，郭煜[1]，赵爽[1]，姜勇[2]
（1. 北京交通大学经济管理学院，北京 100044；
2. 首都医科大学附属北京天坛医院国家神经系统疾病临床医学研究中心，北京 100070）

摘　要　针对脑血管疾病电子病历信息挖掘不够充分、现有电子病历检索模型不能够挖掘文本深层次语义关系等问题，本文提出的一种“NER 和图嵌入技术”相似病历检索方法，首先，利用 Bi-LSTM-ATT-CRF 模型抽取疾病、症状、检查、治疗和身体部位 5 类实体；其次，以实体为点、实体共现关系为边，构建脑血管疾病实体网络；最后，通过 Node2Vec 算法进行图嵌入，通过网络随机游走挖掘电子病历的多度关系。该方法有效提升了医疗相似文本检测的效果。

关键词　电子命名实体识别技术，图嵌入技术，相似病历检索，注意力机制

中图分类号　C931.6

1　引言

脑血管疾病（CVD）是一种危害中老年人群生命健康的疾病[1]，它的治疗周期长、护理周期长、复发率高，由于其三大特性，现在已成为各个医疗救治单位的主要负担。但与此同时，脑血管疾病患者在长期治疗和护理过程中必然会产生大量的临床电子病历数据，其中包含脑血管疾病患者的个人体征、治疗、康复等相关信息，这些信息均由医院计算机系统统一存储与管理。那么，随着人工智能技术的发展，以电子病历为核心的信息挖掘、数据分析，为辅助诊疗和治疗脑血管疾病提供了新的解决思路。

首先，利用Bi-LSTM-ATT-CRF模型抽取疾病、症状、检查、治疗和身体部位5类实体；其次，以实体为点、实体共现关系为边，构建脑血管疾病实体网络；最后，通过 Node2Vec 算法进行图嵌入，通过网络随机游走挖掘电子病历的多度关系，一方面提升病历信息的覆盖率的效果，另一方面，实现对于脑血管电子病历的共享，输入症状、检查等信息，快速匹配和输入症状相似的病历，找出相关的脑血管疾病诊断案例，帮助各级医院的医生进行辅助诊断，从而避免因医生经验少而造成诊断失误的问题，同时通过比较分析相似的病历，可以获得脑血管疾病群体性的医疗健康方面的人口统计学特征，对于脑血管疾病的诊断、预测和治疗方案决策有较大的帮助。

2 相关研究综述

2.1 脑血管疾病辅助诊疗

脑血管疾病是对人类生命安全威胁最大的疾病，我国是脑血管疾病负担最重的国家之一，现有患者 1 494 万人，每年新增 330 万人[2]，因此，各个医院的脑血管疾病患者诊治需求量比较大，而专业医师较少。目前，通过传统的临床路径来提高脑血管疾病患者服务质量，对医生的专业技能要求较高。近年来，覆盖临床电子病历、影像、血液、基因以及患者结局等多维度数据集缺乏深度挖掘[3, 4]，基于临床数据的辅助诊疗技术逐渐成为研究热点，如何利用数据快速评估、诊疗、预后干预关乎每一位脑血管疾病患者的生命健康安全，这是当前亟待解决的问题，脑血管疾病领域专家对此问题展开了一系列的讨论。

第一个研究方向关注于医学影像数据，医学影像是脑血管疾病诊断的重要参考依据，机器学习等技术可以自动提取脑血管疾病组学特征，可以减少由人的主观性带来的判断错误。例如，Wu 等基于来自多个医疗中心的真实世界中患者的医学影像数据，通过 CNN 在相应的数据集上训练与测试，结果表明 Dice 系数达到 0.77~0.86[5]。由此可说明人工智能技术在医学影像方面有着较大的优势。第二个研究方向是脑血管疾病预后预测，Wang 等通过构建机器学习模型来实现对自发性脑出血患者功能结局（mRS 评分）的预测，结果显示在测试集中随机森林方法对脑血管疾病患者功能结局预测性能最佳[6]。随着以人工智能技术辅助诊疗的概念的提出，从脑血管疾病辅助诊疗的现有研究来看，专门针对脑血管疾病电子病历的研究较少，大多数研究的数据源是医学影像结构化数据，而对于文本数据的利用较少；辅助诊疗的方式比较单一，而更多的研究集中于脑血管疾病预后预测，比较单一。

2.2 命名实体识别

在医学领域，对于电子病历的文本挖掘成为研究热点，如药物适用病症的查询与识别，往往需要先提取非结构化文本中的药物实体和症状实体，再通过构建实体间的关系，来识别出药物适用症状，因此，中文电子病历命名实体识别对医学领域信息抽取、信息检索等有着重要的意义。为了抽取电子病历的信息，研究分为两个方向：从实体类别角度，Hu 等在 CCKS（China conference on knowledge graph and semantic computing）全国知识图谱与语义计算大会评测中将医学的实体分为疾病、症状和体征、检查和检验、治疗、身体部位 5 类实体[7]；从识别方法角度，传统的识别方法主要是基于规则、字典、领域专家进行医学命名识别，但是这种方法人工成本高，低效率。随着深度学习技术的发展，Mark 等为了降低构建实体识别模型的成本，提出了基于本体的深度学习方法提取疾病名称[8]，这种方法核心要点在于可以降低注释训练数据的成本。Ling 等提出一种较“新”的方法，即融合 CEMP（chemical entity mention recognition）和 GPRO（gene and protein related object recognition）的方法[9]，此种方法主要采用条件随机场与双向长短期记忆融合模型（Bi-LSTM-CRF）来从大量的专利数据中识别医学实体。张应成等提出了一种混合模型，即利用 CRF 条件随机场和 Bi-LSTM 双向长短时记忆网络相结合模型来抽取文本序列中的实体[10]，该模型相比于单一 CRF 模型，识别效果有提高。Bhaskaran 等明确了当输入文本序列不长时，采用基于 CNN 的模型和基于 Bi-LSTM 的模型，采用全局矢量表示进行嵌入方式[11]。此外，还通过反拟合方法，丰富了 Glove 的同义词和反义词。结果表明，首先，当输入文本序列不长时，采用全局矢量表示进行嵌入方式的可行性；其次，反拟合方法在

语义捕获中具有重要意义；最后，也指出了输入文本序列过长模型的识别效果会略差。随着技术的成熟，实体识别方法的研究成果逐渐成为实践的热点，接下来部分学者聚焦于构建实体识别系统。Ajees 和 Idicula 充分考虑了词性信息、词与后缀的嵌入表示、前后词的词性信息等具有各自不同的特点，并构建了一个基于神经网络的命名实体识别系统[12]。Peter 和 John 对比分析了三种命名实体识别系统的识别效果，一是基于 CRF 的习惯用法构建一个深度学习框架，使用了丰富的标记特征集合和词嵌入，并利用 Bi-LSTM 生成二序列标记；二是规避了大量的特征集合，使用字符嵌入和多 LSTM 层进行字符标记；三是综合所提到的第一种与第二种的结果，即 Bi-LSTM-CRF 模型。实验结果证实第三种系统的识别效果最好[13]。

2.3 电子病历检索

面对海量电子病历的信息，如何从电子病历信息中淘金从而达到辅助诊疗的目的？信息检索为解决此问题提供了新的思路。电子病历检索是跨学科、跨领域的课题，很多研究是将文本检索延伸到电子病历检索中。由于电子病历自身数据特点与传统文本检索任务不同，电子病历检索任务可描述为寻找相似病历的过程。

1. 电子病历检索方法

电子病历是非结构化与结构化数据结合的文本，检索应分为结构化检索和非结构化检索，结构化电子病历检索有两种，一是医生依据检索入口填写相应信息，电子病历系统自动生成结构化电子病历；二是利用电子病历信息抽取的方式，将非结构化的电子病历中的所需要的数据和查询的关键信息提取出来，按照一定的数据规则生成结构化电子病历[14]，其运作原理是通过 XML 的方式进行表示，然后再将提取的 XML 关键信息存储到数据库中，最后按照结构化查询语句的方式进行相应的查询，这种检索方式可快速定位到包含检索关键词的电子病历，并且每次查询的逻辑计算比较简单，但是这种检索方式忽略了文本中的语义信息。非结构化电子病历检索是用自定义输入信息，系统从电子病历文库中检索出相关度最大的病历文档，并按照一定方式排序返回。因此，两个电子病历查询的相关性是检索的核心，具体体现为相似度的计算和排序。

目前，将从电子病历文库中查询并生成排序的过程描述为[14]：从电子病历非结构化文本中识别医学命名实体，并建立丰富的实体词典；再对电子病历文本建立索引，查询时通过匹配索引，计算查询文本和病历库中文本的相似度并按照指定排序顺序返回。

2. 检索模型

在对电子病历检索模型的研究中，很多经典的文本检索模型被引入电子病历检索中[15-17]，常用的模型为向量空间模型、语言模型。向量空间模型主要是用 n 维特征空间向量来表示文档[18]。如表 1 所示，文档 $\boldsymbol{d}_i$ 含有 n 个特征词，则 $\boldsymbol{d}_i$ 表示为一个维度为 n 的向量，即 $\boldsymbol{d}_i=\{W_{i1},W_{i2},W_{i3},\cdots,W_{in}\}$，而表中文档 $\boldsymbol{d}_i$ 的每个特征词权重 W_{ij} 的计算有多种方式，如词频（TF），倒排文档频率（IDF）、词频–倒排文档频率（TF-IDF）。另外，相似度可通过非二值的权重值来计算，即一般将查询索引和文档的相关性转化为查询向量与文档向量的相似度，常采用余弦值来计算相似度，余弦值越大，查询结果越相似。但此方法也存在弊端，仅仅适用短文档。

表 1 向量模型文档表示

文档	特征 1	特征 2	...	特征 n
$\boldsymbol{d}_1$	W_{11}	W_{12}	...	W_{1n}
$\boldsymbol{d}_2$	W_{21}	W_{22}	...	W_{2n}
⋮	⋮	⋮	⋮	⋮
$\boldsymbol{d}_n$	W_{n1}	W_{n2}	...	W_{nn}

语言模型原理是为每一个文档建立语言模型，通过计算由该文档生成用户查询的文档的概率，再按照概率的高低排序后返回。其中，语言模型就是关键词在文档中概率分布情况的表示[19]。常用计算文本相似度的语言模型为 LDA 主题模型。LDA [20]是一种文档主题生成模型，即包含词、主题和文档三层结构的贝叶斯概率模型。生成模型，即可以认为文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到的，文档到主题服从多项式分布，主题到词服从多项式。LDA 的最终目的是识别主题，将文档−词汇矩阵转化为主题−词汇矩阵和文档−主题矩阵[21]。通过分析向量空间模型与 LDA 模型，发现这两种模型的弊端在于均没有办法挖掘文本之间潜在的语义关系。

3. 相似电子病历检索算法

脑血管疾病患者在临床治疗过程中会留下大量的非结构化电子病历数据，利用存在困难[22]。而电子病历数据中蕴含着大量潜在信息，对电子病历文本的挖掘依赖于 NLP（natural language processing，自然语言处理）技术，通过抽取患者电子病历中先验信息进行分析与挖掘，可以实现基于先验信息的疾病预测，再通过匹配相似病历，从症状、用药和治疗等不同维度进行分析，可以实现对医生的辅助诊疗。相似电子病历研究分为两派，一是相似检测技术的研究，如任民山和蔡红霞主要采用的大数量级知识文档的相似性检测技术是基于 Simhash 算法的文档相似性计算技术；特征的提取技术是通过计算权重的 TF-IDF 技术和分词技术 ICT-CLAS 实现的；文档相似度计算是依据海明距离计算的[23]。陈瑞东等提出一种基于网络流量相似性的僵尸网络识别方法，该方法是通过对每个数据特征进行模糊聚类，然后再判别每个模糊类别的特征边界，并基于最大隶属度原则、支持度和置信度筛选关联规则来确定具体的僵尸网络类型[24]。随着知识图谱技术的迅速发展，NLP 领域内已经积累了大量开放本体库，如 DBpedia、WordNet 等，可以基于这些开放本体库再结合上下文的信息进行相似文本推荐，从而可以很好地解决冷启动问题，达到提高相似文本检测算法性能的目的。其中，Xie 等提出 DeepWalk 算法融合深度学习技术和图嵌入方法，主要原理是将图中的节点在网络上的游走路径类比文本生成，对于一个节点序列采用 Skip-gram 和 Hierarchical、Softmax 方法对随机游走序列上的每一个局部窗口中的节点出现概率进行建模，再利用最大似然函数和随机梯度下降方法进行学习[25]。这种方式可以将网路中的节点映射到一个 K 维的空间，可以更好地发现节点之间的关系。

相似电子病历检测方法研究的最终目的是投入实践应用，即构建辅助诊疗系统，第二个方向是基于相似病历算法辅助诊疗系统研究，金文忠等针对肺部检测效能的问题，提出了基于 CNN 算法模型建立辅助诊断的人工智能系统，实验结果显示数据量级集越大，检测结果越准确[26]。牛琳等基于相似电子病历检测技术，设计并实现了乳腺 X 光片辅助诊断系统[27]。从医患角度来讲，具有大量临床经验的医生在做诊断时，除了可以利用自身的知识，还可以尽可能多地检索到相似患者的电子病历信息，降低了医生一些不必要的诊断失误，为医生辅助临床诊断；对于患者而言，可以更多地了解相似患者的诊治过程及发展状况，从而更好地把握自身的状况。

2.4 总结

综上 2.1、2.2、2.3 节，根据脑血管疾病辅助诊疗现状，现有研究大多专注于脑血管疾病预后预测，并未有出现专门针对脑血管疾病的相似电子病历检索的研究。另外，脑血管疾病患者治疗过程中积累了大量的电子病历，其数据存在形式是以大量非结构化的数据为主，处理较为困难，利用率低，而非结构化文本的挖掘依赖于 NLP 技术，而实体识别是 NLP 技术的重要一步。在实体识别方面，电子病历中可以抽取症状、检查、治疗等相关实体的内在联系，目前 Bi-LSTM-CRF 模型在实体识别领域得到充分的应用，但其也存在一定的局限性，由于解码时的向量表示的长度是固定的，当输入的文本序列较长时，Bi-LSTM-CRF 模型的性能会变得很差，所以模型有提升的空间。在相似电子病历检索方面，针对脑血管疾病的辅助诊疗现状集中于预后预测，本文将辅助诊疗扩展到相似病历检索，而常用的文本相似度计算集中于空间向量模型和 LDA 主题模型，但是这些模型无法挖掘文本之间潜在的语义关系。基于实体识别优势与图嵌入的优势，本文提出基于命名实体识别和图嵌入技术的脑血管疾病相似病历，通过实体识别构建实体网络，再利用图网络随机游走挖掘电子病历的多度关系，提升病历信息的覆盖率的效果，期望给脑血管疾病患者带来更大的益处。

3 研究设计

3.1 研究问题

通过对医院业务现状和相关研究现状进行分析，可得出现有研究存在的不足。

（1）在脑血管疾病领域，现有的模型对辅助诊疗的应用场景比较单一，并未实现对电子病历数据的充分挖掘。

（2）对于电子病历信息的挖掘，更多地在关注病历中出现的医学实体，电子病历中不同实体之间的共现关系也是同等重要的，而传统的二维表示方式不够直观。

（3）Bi-LSTM-CRF 模型在实体识别领域得到充分的应用，但模型仍具有提升空间。

（4）相似电子病历检索，相似度计算集中于空间向量模型和LDA主题模型，但是这些模型无法挖掘文本之间潜在的语义关系。

3.2 研究方法

本文采用的技术解决方案是将命名实体识别技术引入电子病历数据的处理中，通过实体识别构建实体网络，再利用图网络随机游走挖掘电子病历的多度关系，提升病历信息的覆盖率的效果。结合图嵌入技术构建相似的病历检索模型，从而实现将脑血管疾病电子病历辅助诊疗的应用场景从单一的相似度计算扩展到相似病历检索。具体过程如下。

（1）通过命名实体识别技术，识别出电子病历中的医学实体，构建脑血管疾病实体关系网络，深度挖掘网络中相似病历的隐藏信息，补充电子病历中的缺失信息。

（2）以词向量的形式输入命名实体识别网络，消除中文分词带来的误差；以基于深度学习的半监督的训练方式进行模型学习，只需要一小部分的标注语料进行模型训练，解决中文电子病历分词难和语料库缺失的问题。

（3）借助网络结构的优势，将命名实体识别技术和图嵌入技术结合，构建基于脑血管疾病命名实体网络和 Node2Vec 的相似病历检索模型，通过随机游走挖掘潜在信息，提高病历匹配准确率和病历关联信息覆盖率。

技术解决方案如图 1 所示。

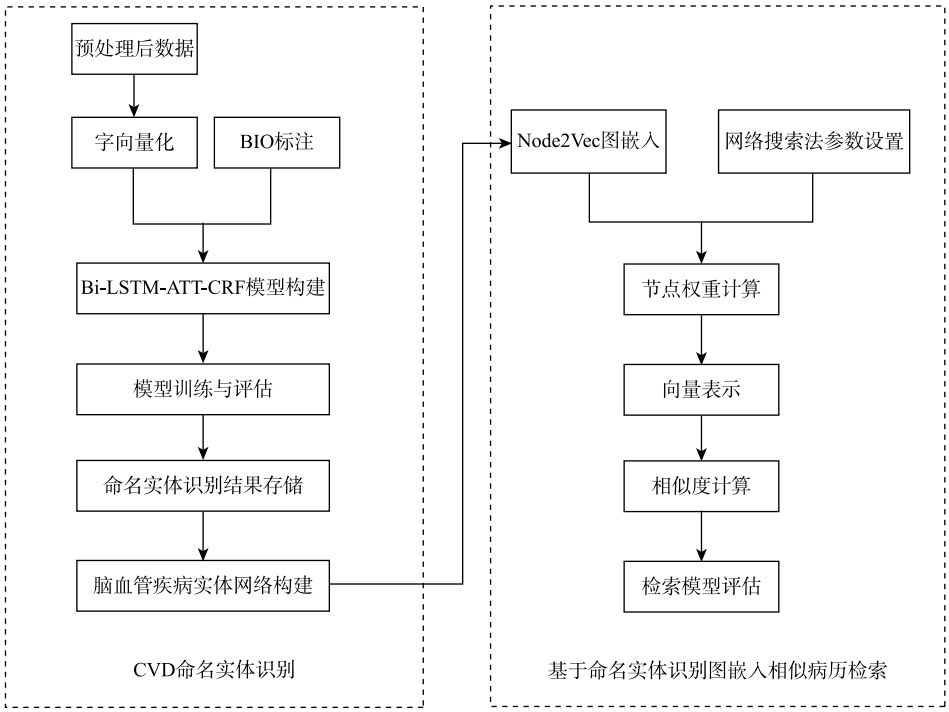

图 1 技术解决方案

3.3 引入注意力机制的命名实体识别模型构建

3.3.1 注意力机制

注意力机制是类似人脑资源分配模型，通过引入该机制，对于文本来说，使得关键词分配更多的注意力，对于其他部分分配更少的注意力，从而使得实体提取的效果更加集中于对于文本来说更关键的词汇。通过引入 Attention 机制，将 Attention 机制和 Bi-LSTM 结合，可以有效突出关键词的作用。例如，对于病历文本“2 天前患者自觉头痛、头晕较前加重，无视物旋转，无恶心、呕吐，伴腹泻”，在不加入 Attention 机制的时候，Bi-LSTM 模型本身关注的是文本的上下文信息，加入 Attention 之后，可以将句子的重点更加集中于“头痛”“头晕”等症状的描述中，Attention 机制通过计算权重可以实现这一功能。

本文模型运行原理是将 Bi-LSTM 层的输出通过 Attention 机制计算每个输入字向量的权重。本文的 Attention 机制[28]的相关计算公式为

$$a_{ki}=\frac{\exp\left(e_{ki}\right)}{\sum_{j=i}^{T}\exp\left(e_{kj}\right)} \tag{1}$$

$$e_{ki}=v\tan h\left(Wh_k+Uh_i+b\right) \tag{2}$$

$$C=\sum_{i=1}^{T}a_{kj}h_i \tag{3}$$

$$\boldsymbol{h}'_k=H\left(C,\boldsymbol{h}'_kX'\right) \tag{4}$$

式（1）给出的是注意力概率分布的语义编码，其中 a_{ki} 代表节点 i 对节点 k 的注意力概率权重；

式（2）给出了注意力权重的计算方式；式（3）中 C 代表语义编码；式（4）表示将语义编码结果和第 i 时刻的输出串联成一个新的向量 $\boldsymbol{h}_k'$，就是最终的特征向量，通过该特征向量可以表征关键词的语义信息。

3.3.2 Bi-LSTM-ATT-CRF

Bi-LSTM-CRF 模型在实体识别领域虽然取得了较好的效果[29, 30]，但是模型也存在一定的限制，此模型依赖于整个句子的信息，这种方式会产生信息的冗余，而冗余信息的噪声会影响实体识别的效果。在 Bi-LSTM-CRF 模型中引入 Attention 机制，即 Bi-LSTM-ATT-CRF 模型，它既可以结合上下文信息，还可以考虑标签的前后依赖关系，保证标签预测序列的有效性。本文提出的 Bi-LSTM-ATT-CRF 模型的网络结构如图 2 所示。

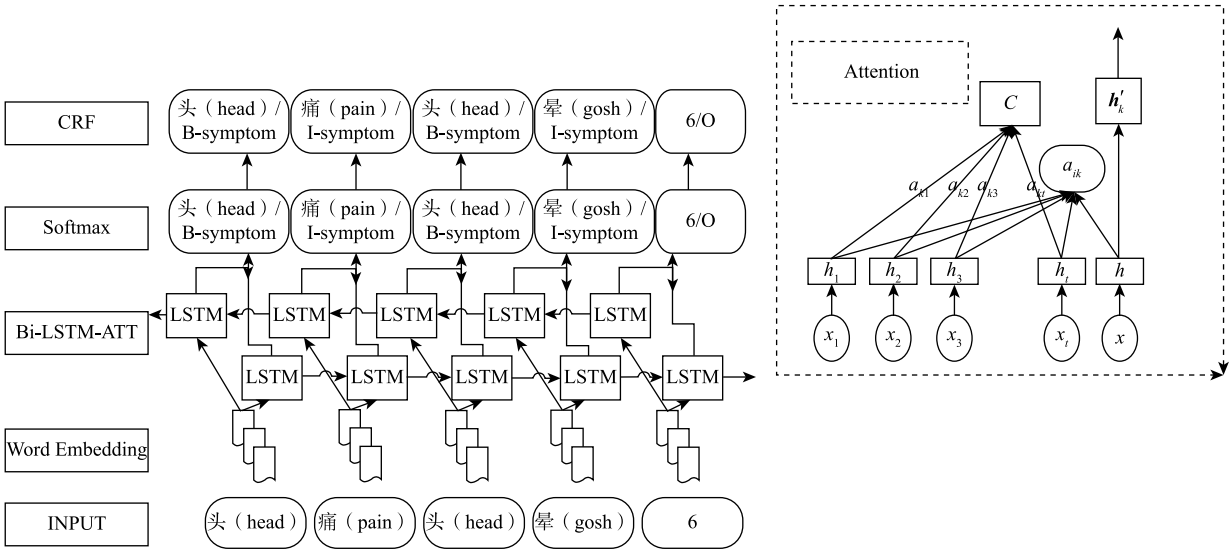

图 2 Bi-LSTM-ATT-CRF 模型结构

在 Bi-LSTM-ATT-CRF 的 Softmax 输出层后加入 CRF 层，引入状态转移矩阵 $\boldsymbol{A}$ 作为 CRF 层的参数，设矩阵 $\boldsymbol{L}$ 为 Bi-LSTM 的输出，其中 $A_{i,j}$ 表示时间序列上第 i 个状态转移到第 j 个状态的概率；$L_{i,j}$ 表示观察序列中第 i 个词被标为第 j 个标注的概率。本文的模型中使用最大似然估计作为代价函数，则待观察序 X 的待预测标注序列 $Y=(y_1,y_2,\cdots,y_n)$ 的输出计算公式为

$$s(X,Y)=\sum_{i=1}^{n}\left(A_{y_i,y_{i+1}}+L_{i,y_i}\right) \tag{5}$$

$$\log L(y|x)=s(X,Y)-\log_y \exp\left[s(X,Y')\right] \tag{6}$$

3.4 基于脑血管疾病命名实体网络和图嵌入技术的相似病历检索模型构建

在脑血管疾病领域中，对于电子病历信息的挖掘，更多地在关注病历中出现的医学实体，而同一电子病历中不同实体之间的共现关系也是非常重要的。传统的二维表示方式不够直观，而通过图结构，可以清晰直观地显示实体间的直接关联关系，另外，还可以通过对图中节点随机游走的挖掘，发现实体之间的间接关联。因此，不仅可以弥补病历覆盖信息稀疏的缺点，还进一步提高了挖掘实体间

关系的性能。

3.4.1 脑血管疾病命名实体网络构建

本文基于电子病历实体识别的已有成果分析[31]，将实体划分为身体部位、症状和体征、检查和检验、疾病、治疗。脑血管疾病命名实体网络构建的目的一方面是直观表达脑血管疾病领域不同实体在同一电子病历的共现关系，另一方面是通过计算病历相似度，将命名实体关系网络通过图嵌入的方式映射到 K 维空间中，实现间接关系节点的挖掘。因此，本文构建的命名实体网络以实体为节点，以两个实体在同一个病历文本中的共现关系为边。具体实体网络构建流程：脑血管疾病命名实体识别，再提取病历实体共现关系，最后再进行实体权重的计算。

1. 脑血管疾病命名实体识别

脑血管疾病领域的命名实体识别部分是通过调用本文提出的命名实体识别算法，将原始数据中命名实体全部识别出来，按照表 2 中指定的格式进行存储。

表 2 实体识别结构表结构

字段名	类型	限制	描述
ID	int	Primary Key、Not Null、Auto Increment	实体编号
Case_ID	String	Not Null	实体所属文本编号
Entity_Source	int	Not Null	实体来源
Entity_Type	String	Not Null	实体类型
Entity	String	Not Null	实体名称

2. 病历实体共现关系提取

病历实体共现关系的提取，通过实体的 Case_ID、Entity_Source、Entity_Type 和 Entity 的连接关系提取实体在同一份病历中的共现关系，共现关系的权重用贡献次数表示，以此作为命名实体网络的边。因此，通过构建脑血管疾病命名实体识别网络，点信息为实体信息，边信息为实体共现关系，网络可视化用 Gephi 实现。

3. 实体权重计算

在电子病历相似度计算中，很多经典的文本中词语权重的计算被引进来。在空间向量模型中，最常用来计算权重的方法是词频−倒排文档频率（TF-IDF）；而在图结构中，衡量节点的重要程度的测度指标较多，如度中心性、中介中心性、接近中心性、特征向量中心度。其中，度中心性反映网络中一个节点与所有其他节点相联系的程度；中介中心性以经过某个节点的最短路径数据来刻画节点重要性；接近中心性反映网络中某一节点与其他节点的接近程度；特征向量中心度通过某节点的邻居节点的数量来反映该节点的重要程度。在本文构建的网络中，主要考虑的是节点和其邻居节点的关系，因此，选用特征向量中心度作为节点的权重。

3.4.2 基于 Node2Vec 算法的图嵌入

Node2Vec 算法是在 DeepWalk 基础上的一种改进算法，它通过改进节点随机游走的方式优化了 DeepWalk 算法。具体的实现方式是通过参数 p 和 q 来控制随机游走序列进行深度搜索和广度搜索的概率。现有的特征学习方法无法足够地捕捉出网络中节点的多种连接模式，而在搜索相邻节点时增加搜

索的灵活性是提升特征学习算法的关键。Node2Vec 通过对输入数据的训练，输出在 K 维空间上的一系列特征表达，这些特征表达最大化了节点的邻接节点的似然估计。其主要思想是，通过特定的游走方式进行采样，对于每个点都会生成对应的序列。再将这些序列视为文本导入 Word2Vec 中的 CBOW 或者 Skip-Gram 模型，即可得到每个节点的向量。

通过图嵌入过程，可以得到所有实体在指定网络和维度的向量表达，图 3 是数据经过图嵌入训练出来的在 128 维度上的表达，实体 id 为 3，对应的实体为“头痛”。其他所有实体的表示方式与此类似。

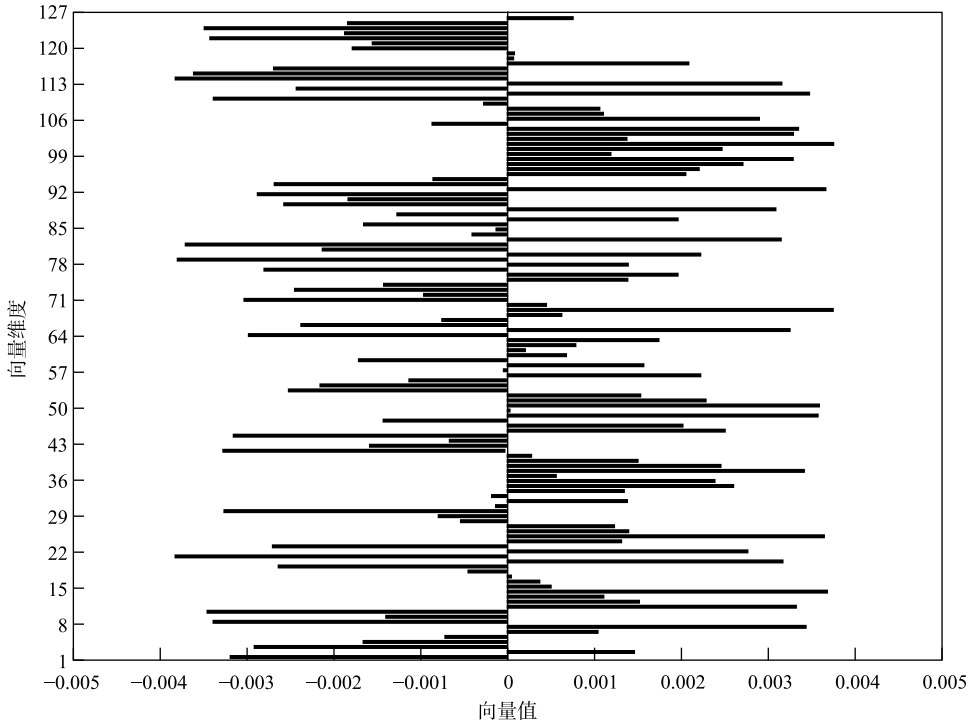

图 3 “头痛”实体图嵌入示例图

3.4.3 相似度计算

篇章向量的计算分为两个步骤：第一步是篇章之间词相似度计算，第二步是词语加权篇章向量计算。假设两个文本分别为 D_1 和 D_2，经过提取特征后，可以表示为 $D_1=(x_1,x_2,\cdots,x_n)$，$D_2=(y_1,y_2,\cdots,y_m)$，n 和 m 是两个文本的实体数量，则 D_1 中第 i 个单词与 D_2 中第 j 个单词的相似度计算如下：

$$\text{similarity}_{w\_i\_j}=\frac{w_i\times w_j\times \text{Vec}(x_i)\times \text{Vec}(y_i)^{\text{T}}}{\sqrt{\text{Vec}(x_i)^2}\times\sqrt{\text{Vec}(y_i)^2}} \tag{7}$$

其中，$\text{Vec}(x_i)=[v_{i1},v_{i2},\cdots,v_{i128}]$，$\text{Vec}(y_j)=[v_{j1},v_{j2},\cdots,v_{j128}]$，即通过 Node2Vec 图嵌入的词向量表达，$w_i$ 和 w_j 分别是词 i 与词 j 的权重，即前文中提到节点的特征中心向量值。则文本向量表示为

$$\text{Vec}(D_1)=[\text{similarity}_{w\_1\_1},\text{similarity}_{w\_1\_2},\cdots,\text{similarity}_{w\_n\_m}]$$

$$\text{Vec}(D_2)=[\text{similarity}_{w\_1\_1},\text{similarity}_{w\_2\_1},\cdots,\text{similarity}_{w\_n\_m}]$$

本文的相似度计算使用余弦相似度来计算。假设两个文本分别为 D_1 和 D_2，经过命名实体识别特征提取后，分别可以表示为 $D_1=(x_1,x_2,\cdots,x_n)$，$D_2=(y_1,y_2,\cdots,y_m)$，n 和 m 分别为两个文本的实体数量，

每个词的向量表示为 $\mathrm{Vec}(x_i)=[v_1,v_2,\cdots,v_k]$，$\mathrm{Vec}(x_i)$ 通过 Node2Vec 图嵌入的词向量表达，则相似度计算如下：

$$\mathrm{similarity}=\frac{\mathrm{Vec}(D_1)\times\mathrm{Vec}(D_2)^{\mathrm{T}}}{\sqrt{\mathrm{Vec}(D_1)^2}\times\sqrt{\mathrm{Vec}(D_2)^2}} \tag{8}$$

3.4.4 相似度排序

在排序算法中，最为简单的是冒泡和插入排序，但是这两种算法的时间复杂度为 $O(n^2)$，而在复杂度为 $O[n\log(n)]$ 的算法如快速排序和归并排序中，都需要将所有的节点进行排序，而在相似病历检索中，只需要返回相似度前 n 的文档，对剩余文档的排序并不需要考虑。因此可以采取最大堆排序算法。

4 实验设计

4.1 数据来源

本文实验数据由北京天坛医院的协同防治云平台获取的 1 200 份脑血管疾病电子病历数据、28 000 份脑血管疾病寻医问药官网检索数据、1G 的中文维基百科数据三部分构成。其中，电子病历数据主要用于构建脑血管疾病命名实体网络和验证相似病历检索模型的性能，在北京天坛医院专业医生的指导下，以 50 例相似病历作为正样本，50 篇非相似病历作为负样本，通过分别计算相似病历返回结果中正样本的比率、负样本的比率来验证模型的准确性；脑血管疾病在线医学文本主要扩充脑血管疾病命名实体网络中的实体；维基百科数据用于构建预训练字向量和语义网。

4.2 数据处理

4.2.1 数据预处理

本文使用的电子病历需要对接天坛医院的协同防治云平台，获取电子病历数据，将获取到数据后按照制定格式进行存储，数据预处理主要为数据脱敏、数据特殊符号替换；脑血管疾病在线医疗文本的数据需要用 Python 语言爬取，预处理主要为网页内容解析和数据清洗，其中数据清洗工作主要包括缺失值的处理和特殊符号的清洗与转换。

4.2.2 BIO 标注

实体识别过程中需要对字符串中的实体进行标注，BIO 标注是将每类实体标注为“B-X”“I-X”“O”的格式。其中，“B-X”表示 X 实体的头部，“I-X”表示 X 实体的中间位置，“O”表示不属于任何类型。BODY 代表身体部位、SIGNS 代表症状、CHECK 代表检查、DISEASE 代表疾病、TREATMENT 代表治疗。

4.2.3 文本向量化

在脑血管疾病实体识别领域，需要将向量化的文本输入网络中，本文向量化使用 Word2Vec 结合电子病历语料进行字向量的训练，结合 1G 的维基百科数据，使用 Word2Vec 的 CMOW 模型训练成为维度为 100 的词嵌入查询表，词嵌入查询表可以将样本中的每个字转换成词向量；将实体的位置映射成实数向量。将词向量和位置向量拼接起来，即输入字序列的向量化表示。

4.3 Bi-LSTM-ATT-CRF 模型参数

本文的 Bi-LSTM-ATT-CRF 模型的训练环境为 Linux 环境，模型使用 Tensorflow 框架实现，框架的训练使用反向传播算法。对于 Bi-LSTM-ATT-CRF 模型的 LSTM 层，每个方向的 LSTM 隐藏层的神经元数目均为 100，即 Embedding 维数为 100；模型训练采用随机梯度下降方法，为了防止梯度消失和梯度爆炸，梯度阈值（clip）为 5，训练数据片大小（batch_size）为 64，学习率（lr）为 0.001，Dropout 过程的概率为 0.5；另外，对于 Bi-LSTM-ATT-CRF 模型的 CRF 层，采用 Tensorflow 默认的 CRF 设置，默认的链长为 512。迭代训练表明此时模型的精度最高。

4.4 脑血管疾病命名实体识别结果分析

4.4.1 评价标准

命名实体识别领域最常用的指标[32]是 Precision 准确率，Recall 召回率，F_1 值。即 P 是被正确识别出来的命名实体识别比率，R 是在测试集合中命名实体被正确识别出来的比率，F_1 值为 P 和 R 的调和平均数，是模型的综合评价指标。其中，当 P、R 值越高，准确率与召回率越高，但是事实上这两者在某些情况下是矛盾的，因此，常用 F_1 值来评价模型的综合性能。具体公式如下：

$$P = N_r / N_s \left(0 \leqslant P \leqslant 1\right) \tag{9}$$

$$R = N_r / N_{\text{all}} \left(0 \leqslant R \leqslant 1\right) \tag{10}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \left(0 \leqslant F_1 \leqslant 1\right) \tag{11}$$

式（9）中给出的 N_r 表示识别出正确实体的数量；N_s 为识别出的实体总数；式（10）中 N_{all} 为文本中实体总数。

4.4.2 实体识别对比实验

命名实体识别实验对比结果如表 3 所示。

表 3 命名实体识别实验对比结果

模型	Category	Precision	Recall	F_1
CRF	BODY	70.05%	74.95%	72.42%
	EXAMINATION	83.57%	85.99%	84.76%
	DISEASE	58.33%	69.96%	63.62%
	SYMPTOM	88.65%	91.11%	89.86%
	TREATMENT	50.11%	63.41%	55.98%
	OVERALL	70.14%	77.08%	73.45%
Bi-LSTM	BODY	81.17%	90.92%	85.77%
	EXAMINATION	90.36%	93.97%	92.13%
	DISEASE	78.55%	90.44%	84.08%
	SYMPTOM	91.58%	94.47%	93.00%
	TREATMENT	66.13%	92.06%	76.97%
	OVERALL	81.56%	92.37%	86.63%

续表

模型	Category	Precision	Recall	F_1
Bi-LSTM-CRF	BODY	84.00%	81.05%	82.50%
	EXAMINATION	93.73%	95.16%	94.44%
	DISEASE	84.62%	75.00%	79.52%
	SYMPTOM	89.38%	98.70%	93.81%
	TREATMENT	73.03%	75.58%	74.28%
	OVERALL	85.10%	85.10%	85.10%
Bi-LSTM-ATT-CRF	BODY	86.03%	85.18%	85.60%
	EXAMINATION	95.09%	94.23%	94.66%
	DISEASE	87.20%	80.23%	83.57%
	SYMPTOM	91.20%	98.23%	94.58%
	TREATMENT	74.02%	79.01%	76.43%
	OVERALL	86.71%	87.38%	87.04%

从实体类别角度分析，上述的四种模型，检查类、症状类实体有较高的识别率，一方面，说明病历中常见的症状和检查具有高的相似性；另一方面，检查类实体与症状类实体构成比较简单，结构性比较强。例如，检查类实体包含“检查”“检验”等关键词。另外，疾病类和治疗类实体识别效果较低，其原因在于这两类实体的数量在整个数据集中偏少。从 P、R、F_1 值分析，在准确率 P 上，Bi-LSTM-ATT-CRF 模型均高于其他三个模型；在召回率 R 上，Bi-LSTM-ATT-CRF 模型相比于其他模型有所波动，尤其是对比 Bi-LSTM 模型，虽然在 R 值有所降低，但是 P 值提升了，模型的整体性能与 P、R 值两者均有关系，因此，F_1 值更具有参考力。从 F_1 看，Bi-LSTM-ATT-CRF 模型均高于其他模型。从对比实验分析，F_1 值是测度模型综合性能的指标，对比 CRF、Bi-LSTM、Bi-LSTM-CRF 模型，Bi-LSTM-ATT-CRF 模型的 F_1 值均高于其他模型，特别是比 Bi-LSTM-CRF 高 1.9%，充分说明了引入注意力机制的有效性，即 Attention 机制通过赋予每个输出向量不同的权重，突出了关键词作用，减少了无效向量的干扰。

综上所述，本文提出的 Bi-LSTM-ATT-CRF 模型在电子病历实体识别领域有更好的效果，因此，此模型抽取疾病、症状、检查、治疗和身体部位 5 类实体有更高的精准性，进而保证以实体为中心点，实体共现关系为边来构建脑血管疾病实体网络有更高的可信度。接下来，通过 Node2Vec 算法进行图嵌入，将脑血管疾病实体网络转化为向量表达，从而实现基于图搜索的思想，利用网络随机游走的方式更深入地挖掘电子病历多度关系，最终达到提升病历信息的覆盖率的效果。

4.5 相似病历检索模型训练

4.5.1 实体识别及关系提取

利用 Bi-LSTM-ATT-CRF 命名实体识别模型，遍历所有文本，进行命名实体识别，其中实体分布如图 4 所示，本文使用的实验数据识别出 107 149 个实体（无去重），重复数据删除后识别出 9 022 个实体。电子病历和在线医疗文本中对身体部位的描述最多，其次是检查和检验以及症状和体征，而治疗类于疾病类型较少；基于实体识别的结果提取实体共现关系，共现关系权重分布如图 5 所示，其中共现关系的权重用共现次数表示，共提取到 5 175 条关系对。共现次数在 0~50 的占比最多，共现次数大于

1 000 和介于 500~1 000 的最少，但这种共现实体是在文本相似度比较计算时的重点关注对象。

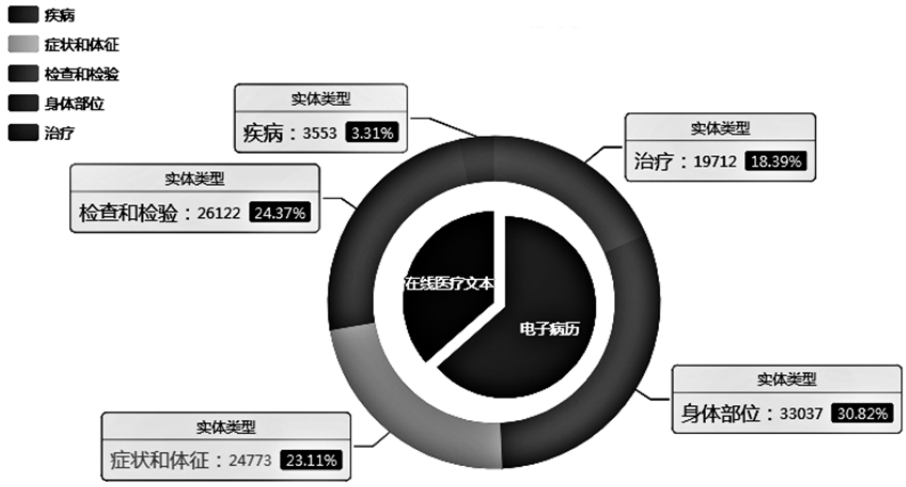

图 4　命名实体类型分布

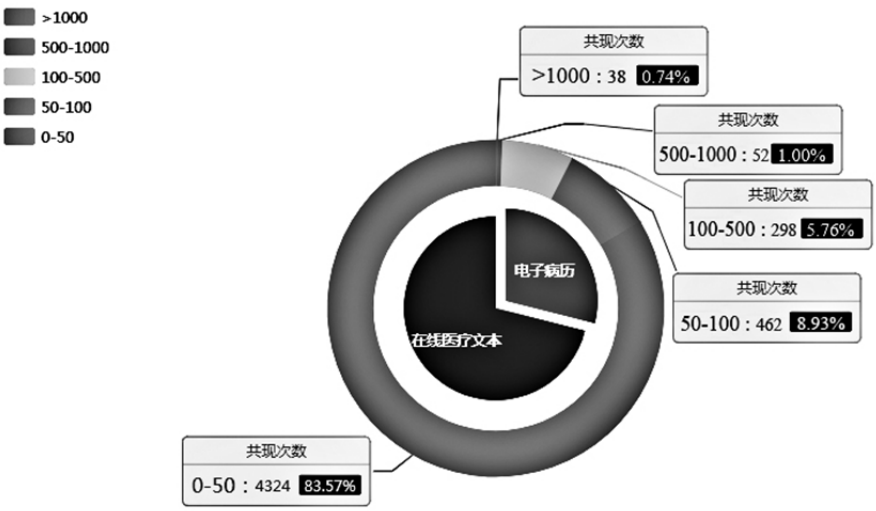

图 5　共现关系权重分布

4.5.2　脑血管疾病命名实体网络表示学习

通过提取命名实体之间的共现关系，由 gephi 构建基础命名实体网络图，脑血管疾病命名实体网络图如图 6 所示，可以看出实体被聚为 4 簇，分别是检查及检验实体、症状和体征类、身体部位实体、治疗实体。在脑血管疾病领域，同一类实体的关联性强，如症状之间伴随出现，治疗方式与检查方式也是协同进行的。因此，从实体类别维度挖掘病历的相似性，可以为电子病历辅助诊疗带来更加直观和有效的帮助。

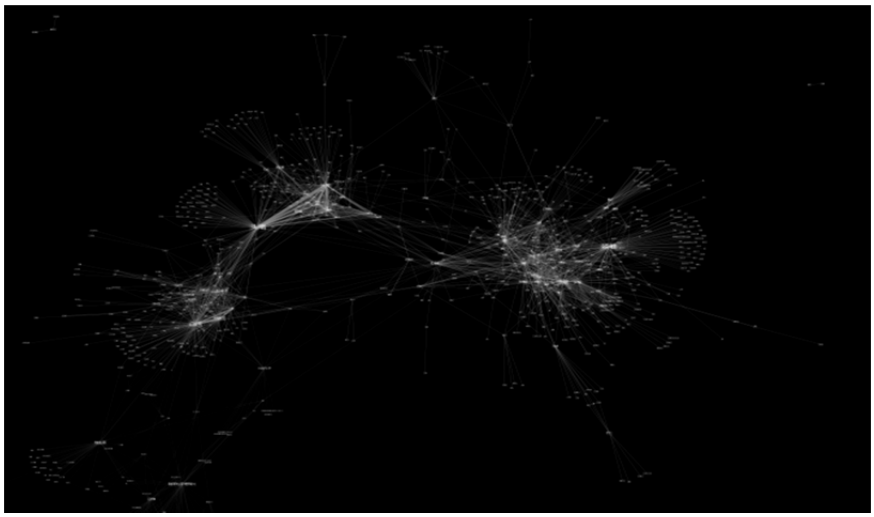

图 6 脑血管疾病命名实体网络图

本文的命名实体网络主要关注实体与邻居节点之间的关系，因此使用广度优先搜索最为合适。利用网格搜索方法，确定 Node2Vec 的主要参数为 p=0.8，q=0.2，步长=10，Node2Vec 学习的文本篇章向量表示如图 7 所示。

Case_id	dim1	dim2	dim3	dim4	dim5	dim6	...	dim128
3	0.001376109	−0.001231826	−0.000552627	−0.003184685	−0.000153786	0.000689239	...	−0.001495734
1	0.00210065	0.000858748	−1.32E−05	−0.002922465	−0.000905336	0.000646856	...	−0.00129372
2	−0.00218982	−0.003541881	−0.000179324	−0.002377921	0.002354986	−0.001348659	...	0.001269625
4	0.002519123	0.001692382	0.000991087	−0.001700501	−0.002174082	−0.001251478	...	0.000948104
7	0.001539345	0.00132265	−0.002089603	−0.001687471	0.001404917	−0.002338786	...	0.001836667
6	0.00028173	0.00336015	−0.001746403	0.002258092	0.00114974 3	0.001220995	...	0.003224136
13	−0.001969829	−0.001849476	0.00150835	−0.00201558	−0.000215571	0.000834102	...	−0.001668105
8	−0.002510364	−0.002111377	0.000951374	0.003902945	0.000694668	0.001503932	...	−0.001658908
20	0.001840288	−0.001222124	−0.001662075	−0.002773228	0.001218107	−0.003612828	...	−0.003535946
17	0.001208052	0.001572681	−0.002855825	−0.002225816	−0.003258857	0.001723028	...	−0.003877001
21	0.002589502	0.002884065	−0.001920423	−0.002372888	0.000716164	−0.003390536	...	−0.003790991

图 7 病历文本网络表示学习向量

4.5.3 实体权重确定

本文使用了常用的多种方法进行实体权重的确定，在其他条件不变的情况，对比了常用 TF-IDF、PageRank 算法、平均值算法和特征中心向量，预测准确率结果对比如表 4 所示。本文提出的使用特征中心向量作为权重的方式可以有效提升相似病历检索模型的效果，所以选择特征中心向量作为网络节点的权重计算文本相似度。

表 4 预测准确率结果对比

计算方法	准确率
TF-IDF	79.6%
PageRank	79.2%
平均值	77.9%
特征中心向量	89.2%

4.5.4 相似病历对比实验

本文选取的评价指标准确率（Precision）是返回的结果中相似电子病历所占的比例，召回率（Recall）是返回的相似电子病历占所有相似电子病历的比例，F_1是准确率与召回率的调和平均值。

表 5 相似病历对比实验结果

计算方法	Precision	Recall	F_1
NER+Node2Vec	89.2%	87.4%	88.2%
Space Vector Model	80.3%	82.1%	81.2%
LDA	84.2%	81.9%	83.0%
NER+LDA	86.5%	85.7%	86.1%

注：表中NER+LDA是用NER进行关键词的提取

（1）将命名实体识别技术和图嵌入技术融合的模型效果优于传统的空间向量模型，比 LDA 主题模型和基于命名实体识别的改进的 LDA 主题模型好，验证了所提模型的有效性。

（2）基于命名实体识别的 LDA 主题模型优于传统的 LDA 主题模型。结果表明，通过命名实体识别来确定关键文本信息并减少噪声干扰，也可以帮助提高相似文本检索的准确性。

经过前文的研究，基于命名实体识别和 Node2Vec 算法的相似病历检索模型具有良好的效果，其结果可为医生的诊治提供帮助。

5 总结与展望

本文研究结果表明，在命名实体识别方面，通过对条件随机场、Bi-LSTM 和 Bi-LSTM-CRF 模型的比较，证明了引入注意力机制的 Bi-LSTM-ATT-CRF 模型在医学命名实体识别领域的适用性。在相似病历检测方面，与 LDA 主题模型进行对比实验，证明通过融合命名实体识别技术和图嵌入技术可以有效提高医学相似文本检测的效果，这丰富了医疗领域命名实体识别和文本相似检测的研究视角与研究方法。在应用层面，结合具体的行业应用需求，构建了脑血管疾病电子病历辅助诊疗系统，一方面，可以快速匹配脑血管疾病症状相近的病历表现，找到相关的诊疗案例，帮助医生进行辅助诊疗，避免不必要的诊断失误，提高医生及医院的工作效率；另一方面，在相似病历的比较分析中，获取到了脑血管疾病医疗健康方面的统计学特征，可以为脑血管疾病的预测、防控和预后提供支持。

参 考 文 献

[1] Wu S，Wu B，Liu M，et al. Stroke in China：advances and challenges in epidemiology，prevention，and management[J]. Lancet Neurol，2019，18（4）：394-405.

[2] Wang W，Jiang B，Sun H，et al. Prevalence，incidence and mortality of stroke in China：results from a nationwide population-based survey of 480,687 adults[J]. Circulation，2017，135（8）：759-771.

[3] 王拥军，李子孝，丁玲玲. 人工智能在卒中诊疗的研究和应用：曙光初现，任重道远[J]. 中国卒中杂志，2020，15（3）：223-227.

[4] 李子孝，刘涛，丁玲玲，等. 机器学习在脑血管病诊疗应用中的研究进展[J]. 中国卒中杂志，2020，15（3）：283-289.

[5] Wu O，Winzeck S，Giese A-K，et al. Big data approaches to phenotyping acute ischemic stroke using automated lesion

segmentation of multi-center magnetic resonance imaging data[J]. Stroke，2019，50（7）：1734-1741.

[6] Wang H L，Hsu W Y，Lee M H，et al. Automatic machine-learning-based outcome prediction in patients with primary intracerebral hemorrhage[J]. Front Neurol，2019，（10）：910.

[7] Hu J L，Shi X，Liu Z J，et al. HITSZ CNER：a hybrid system for entity recognition from Chinese clinical text[C]//CEUR Workshop Proceedings. Chengdu：The Technical Committee on Language and Knowledge Computing of the Chinese Information Processing Society of China，2017：25-30.

[8] Mark A M，Peter N，Ernest M. Ontology boosted deep learning for disease name extraction from Twitter messages[J]. Journal of Big Data，2018，5（1）：1-19.

[9] Ling L，Yang Z H，Yang P，et al. A neural network approach to chemical and gene/protein entity recognition in patents[J]. Journal of Cheminformatics，2018，10（1）：65.

[10] 张应成，杨洋，蒋瑞，等. 基于 BiLSTM-CRF 的商情实体识别模型[J]. 计算机工程，2019，45（5）：308-314.

[11] Bhaskaran S K，Sreejith C，Rafeeque P C. Neural networks and conditional random fields-based approach for effective question processing[J]. Procedia Computer Science，2018，（143）：211-218.

[12] Ajees A P，Idicula S M. A named entity recognition system for malayalam using neural networks[J]. Procedia Computer Science，2018，（143）：962-969.

[13] Peter C，John B. Chemlistem：chemical named entity recognition using recurrent neural networks[J]. Journal of Cheminformatics，2018，10（1）：1-9.

[14] 张佳玥. 电子病历检索中时序语义相似度研究[D]. 北京邮电大学硕士学位论文，2018.

[15] 邹涛，王继成，杨文清，等. 文本信息检索技术[J]. 计算机科学，1999，（9）：72-75.

[16] 王斌. 文本检索综述[J]. 数字图书馆论坛，2006，（8）：1-9，35.

[17] 丁志均，杨青，张会兵，等. 基于非结构化文本检索模型综述[J]. 计算机应用研究，2017，34（6）：1601-1608，1612.

[18] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001，18（9）：23-26.

[19] 刘江华. 一种基于 k-means 聚类算法和 LDA 主题模型的文本检索方法及有效性验证[J]. 情报科学，2017，35（2）：16-21，26.

[20] 王振振，何明，杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学，2013，40（12）：229-232.

[21] 张璐，芦天亮，杜彦辉. 基于 WMF_LDA 主题模型的文本相似度计算[J]. 计算机应用研究，2019，36（10）：2916-2919，2951.

[22] Garg R，Oh E，Naidech A，et al. Automating is chemic stroke subtype classification using machine learning and natural language processing[J]. Journal of Stroke and Cerebrovascular Diseases，2019，28（7）：2045-2051.

[23] 任民山，蔡红霞. 基于 Simhash 算法的海量文本相似性检测方法研究[J]. 计量与测试技术，2018，45（4）：78-80.

[24] 陈瑞东，赵凌园，张小松. 基于模糊聚类的僵尸网络识别技术[J]. 计算机工程，2018，44（10）：46-50.

[25] Xie J，Yang Z，Neubig G，et al. Neural cross-lingual named entity recognition with minimal resources[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing，2018：369-379.

[26] 金文忠，陆耀，汪阳. 基于人工智能的胸部 CT 智能辅助诊断系统在 LDCT 数据集上的应用研究[J]. 中国医学计算机成像杂志，2018，24（5）：373-377.

[27] 牛琳，张雨薇，张露馨. 基于 SVM 算法的乳腺 X 光片辅助诊断系统的设计与实现[J]. 软件工程，2018，21（8）：42-45.

[28] Daniluk M，Rocktschel T，Welbl J，et al. Frustratingly short attention spans in neural language modeling[C]. ICLR，

2017.
[29] 吴俊，程垚，郝瀚，等. 基于BERT嵌入BiLSTM-CRF模型的中文专业术语抽取研究[J]. 情报学报，2020，39（4）：409-418.
[30] 曹依依，周应华，申发海，等. 基于 CNN-CRF 的中文电子病历命名实体识别研究[J]. 重庆邮电大学学报（自然科学版），2019，31（6）：869-875.
[31] Xia Y，Wang Q. Clinical named entity recognition：ECUST in the CCKS-2017 shared task 2[C]. CEUR Workshop Proceedings. Chengdu：The Technical Committee on Language and Knowledge Computing of the Chinese Information Processing Society of China，2017：43-48.
[32] 杨锦锋，于秋滨，关毅，等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报，2014，40（8）：1537-1562.

Research on Similar Medical Records of Cerebrovascular Diseases with Named Entity Recognition and Graph Embedding Technology

QIN Qiuli[1]，GUO Yu[1]，ZHAO Shuang[1]，JIANG Yong[2]
（1. School of Economics and Management，BJTU，Beijing 100044，China；
2. China National Clinical Research Center for Neurological Diseases，Beijing 100070，China）

Abstract Aiming at the problems of insufficient mining of electronic medical record information for cerebrovascular diseases and the inability of existing electronic medical record retrieval models to mine deep semantic relations of text，this paper proposes a "NER and graph embedding technology" similar medical record retrieval method. Firstly，this paper uses the Bi-LSTM-ATT-CRF model to recognize five types of entities，which include disease，symptoms，examination，treatment，and body parts. The model achieved good recognition effect. Secondly，taking entity as the point and entity co-occurrence relationship as the edge to construct the network structure diagram of cerebrovascular disease entity. Then Node2Vec algorithm is used to embed the graph，and the multi-degree relationship of electronic medical record is mined through random walk of network. Finally，compared with the traditional methods，the model is proved to be effective.

Key Words Electronic Named Entity Recognition Technology，Graph Embedding Technology，Similar Medical Record Retrieval，Attention Mechanism

作者简介

秦秋莉（1972—），女，北京交通大学经济管理学院副教授、硕士生导师，研究方向包括企业信息化理论与实践、数据挖掘与数据分析、ERP 应用与实践、电子商务、社交网络分析，研究领域包括交通信息化、医疗信息化、教育信息化等；E-mail：qlqin@bjtu.edu.cn。

郭煜（1995—），女，北京交通大学经济管理学院 2016 年级硕士研究生，研究方向为数据挖掘、自然语言处理；E-mail：18801261527@163.com。

赵爽（1996—），女，北京交通大学经济管理学院 2019 年级硕士研究生，研究方向为数据挖掘、医学文本挖掘；E-mail：19120630@bjtu.edu.cn。

姜勇（1978—），男，首都医科大学附属北京天坛医院国家神经系统疾病临床医学研究中心大数据中心负责人，副研究员，研究方向为脑血管病流行病学及临床研究和大数据分析方法学研究；E-mail：jiangyong@ncrcnd.org.cn。