(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114996388 A
(43)申请公布日 2022.09.02

(21)申请号 202210841464.X

(22)申请日 2022.07.18

(71)申请人 湖南创星科技股份有限公司
地址 410000 湖南省长沙市长沙高新开发区岳麓西大道1698号麓谷科技创新创业园B1栋7楼、8楼及B2栋7楼、8楼

(72)发明人 段志强 文建全 彭炜 任强 黄刊迪 谢春梅

(74)专利代理机构 长沙智德知识产权代理事务所(普通合伙) 43207
专利代理师 陈铭浩

(51)Int.Cl.
G06F 16/31(2019.01)
G06F 16/36(2019.01)
G06F 40/289(2020.01)
G16H 10/60(2018.01)

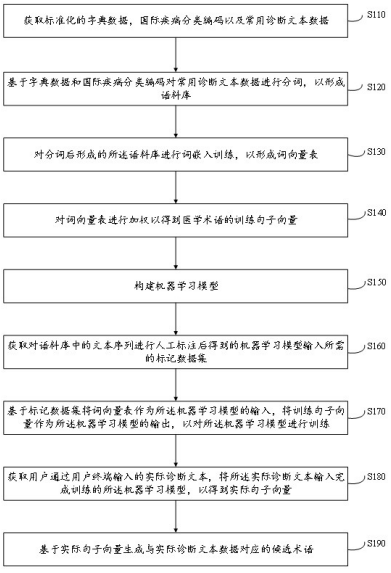
权利要求书2页 说明书7页 附图1页

(54)发明名称

一种诊断名称标准化的智能匹配方法及系统

(57)摘要

本发明公开了一种诊断名称标准化的智能匹配方法及系统,本发明提出的诊断名称标准化的智能匹配方法提出了一种基于自然语言处理和医学知识库,由自然语言诊断名称到IDC9-CM-3、ICD10标准诊断编码及名称的智能匹配技术,以提高医学文本的结构化水平;本发明可以高效快速的实现对医师输入的杂乱无序的诊断术语达到标准有序的后结构化目标,能够更加高效的帮助医师将手写的电子病历进行结构化匹配;在医院的实际挂号和会诊过程中发挥落地能力,降低因医师主观疏忽原因导致的误诊事故,从而推进数字医疗产业化。



CN 114996388 A

1.一种诊断名称标准化的智能匹配方法,其特征在于,包括:

获取标准化的字典数据,国际疾病分类编码以及常用诊断文本数据;

基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词,以形成语料库;

对分词后形成的所述语料库进行词嵌入训练,以形成词向量表;

对词向量表进行加权以得到医学术语的训练句子向量;

构建机器学习模型;

获取对语料库中的文本序列进行人工标注后得到的机器学习模型输入所需的标记数据集;

基于标记数据集将词向量表作为所述机器学习模型的输入,将训练句子向量作为所述机器学习模型的输出,以对所述机器学习模型进行训练;

获取用户通过用户终端输入的实际诊断文本,将所述实际诊断文本输入完成训练的所述机器学习模型,以得到实际句子向量;

基于实际句子向量生成与实际诊断文本数据对应的候选术语。

2.根据权利要求1所述的一种诊断名称标准化的智能匹配方法,其特征在于,所述基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词,以形成语料库,包括:

获取医学术语表;

基于医学术语表生成自定义词典;

基于字典数据、国际疾病分类编码和自定义词典对常用诊断文本数据进行分词,以形成语料库。

3.根据权利要求1所述的一种诊断名称标准化的智能匹配方法,其特征在于,所述基于实际句子向量生成与实际诊断文本数据对应的候选术语,之后还包括:

对人工干预过后的候选术语采用协同过滤算法进行过滤,以生成初始推荐结果。

4.根据权利要求3所述的一种诊断名称标准化的智能匹配方法,其特征在于,所述对人工干预过后的候选术语采用协同过滤算法进行过滤,以生成初始推荐结果,包括:

获取用户通过用户终端输入的选择数量;

将所述候选术语的数量设置为所述选择数量。

5.根据权利要求3所述的一种诊断名称标准化的智能匹配方法,其特征在于,所述对人工干预过后的候选术语采用协同过滤算法进行过滤,以生成初始推荐结果,之后还包括:

采用云原生向量数据库对初始推荐结果进行向量检索以得到倒排索引;

根据索引编号获取初始推荐结果对应的key-value数据库中存储的具体术语缓存数据;

将具体术语缓存数据推荐给用户。

6.根据权利要求5所述的一种诊断名称标准化的智能匹配方法,其特征在于,所述将具体术语缓存数据推荐给用户,包括:

将具体术语缓冲数据通过用户终端的显示界面以下拉框的形式进行显示。

7.根据权利要求1所述的一种诊断名称标准化的智能匹配方法,其特征在于,所述基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词,以形成语料库,还包括:

通过Python的第三方分词模块基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词,以形成语料库。

8.根据权利要求1所述的一种诊断名称标准化的智能匹配方法，其特征在于，所述基于标记数据集将词向量表作为所述机器学习模型的输入，包括：

将词向量表的文本序列中的各个单词，通过词向量表映射为对应的词向量序列；

基于所述标记数据集将所述词向量序列作为所述机器学习模型的输入。

9.根据权利要求1所述的一种诊断名称标准化的智能匹配方法，其特征在于，所述标记数据集包括标记名；所述标记名和所述文本序列一一对应。

10.一种诊断名称标准化的智能匹配系统，其特征在于，应用于如权利要求1-9中任一项所述的诊断名称标准化的智能匹配方法；所述系统包括：

标准术语字典建模模块，用于获取标准化的字典数据，国际疾病分类编码及常用诊断文本数据；基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库；对分词后形成的所述语料库进行词嵌入训练，以形成词向量表；对词向量表进行加权以得到医学术语的训练句子向量；构建机器学习模型；基于标记数据集将词向量表作为所述机器学习模型的输入，将训练句子向量作为所述机器学习模型的输出，以对所述机器学习模型进行训练；

推荐模块，用于获取对语料库中的文本序列进行人工标注后得到的机器学习模型输入所需的标记数据集；基于实际句子向量生成与实际诊断文本数据对应的候选术语；

用户画像模块，用于获取用户通过用户终端输入的实际诊断文本，将所述实际诊断文本输入完成训练的所述机器学习模型，以得到实际句子向量。

一种诊断名称标准化的智能匹配方法及系统

技术领域

[0001] 本发明涉及医疗大数据技术领域，具体涉及一种诊断名称标准化的智能匹配方法及系统。

背景技术

[0002] 随着信息技术的发展和医疗信息化的建设，现在医院都在逐渐使用结构化报告系统进行报告书写，但当前的前结构化的病历模板不够灵活，医生使用不方便导致利用率较低；而且在实际应用中，将医生随手写的电子病历进行结构化又需要有一定的医学知识门槛，故急需开发一种能够更加高效的帮助医师将手写的电子病历进行结构化匹配，从而自动生成结构化电子病历的技术。

发明内容

[0003] 本发明的主要目的是提供一种诊断名称标准化的智能匹配方法及系统，旨在解决急需开发一种能够更加高效的帮助医师将手写的电子病历进行结构化匹配的技术的问题。

[0004] 本发明提出的技术方案为：

一种诊断名称标准化的智能匹配方法，包括：

获取标准化的字典数据，国际疾病分类编码以及常用诊断文本数据；

基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库；

对分词后形成的所述语料库进行词嵌入训练，以形成词向量表；

对词向量表进行加权以得到医学术语的训练句子向量；

构建机器学习模型；

获取对语料库中的文本序列进行人工标注后得到的机器学习模型输入所需的标记数据集；

基于标记数据集将词向量表作为所述机器学习模型的输入，将训练句子向量作为所述机器学习模型的输出，以对所述机器学习模型进行训练；

获取用户通过用户终端输入的实际诊断文本，将所述实际诊断文本输入完成训练的所述机器学习模型，以得到实际句子向量；

基于实际句子向量生成与实际诊断文本数据对应的候选术语。

[0005] 优选的，所述基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库，包括：

获取医学术语表；

基于医学术语表生成自定义词典；

基于字典数据、国际疾病分类编码和自定义词典对常用诊断文本数据进行分词，以形成语料库。

[0006] 优选的，所述基于实际句子向量生成与实际诊断文本数据对应的候选术语，之后

还包括：

对人工干预过后的候选术语采用协同过滤算法进行过滤，以生成初始推荐结果。

[0007] 优选的，所述对人工干预过后的候选术语采用协同过滤算法进行过滤，以生成初始推荐结果，包括：

获取用户通过用户终端输入的选择数量；

将所述候选术语的数量设置为所述选择数量。

[0008] 优选的，所述对人工干预过后的候选术语采用协同过滤算法进行过滤，以生成初始推荐结果，之后还包括：

采用云原生向量数据库对初始推荐结果进行向量检索以得到倒排索引；

根据索引编号获取初始推荐结果对应的key-value数据库中存储的具体术语缓存数据；

将具体术语缓存数据推荐给用户。

[0009] 优选的，所述将具体术语缓存数据推荐给用户，包括：

将具体术语缓冲数据通过用户终端的显示界面以下拉框的形式进行显示。

[0010] 优选的，所述基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库，还包括：

通过Python的第三方分词模块基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库。

[0011] 优选的，所述基于标记数据集将词向量表作为所述机器学习模型的输入，包括：

将词向量表的文本序列中的各个单词，通过词向量表映射为对应的词向量序列；

基于所述标记数据集将所述词向量序列作为所述机器学习模型的输入。

[0012] 优选的，所述标记数据集包括标记名；所述标记名和所述文本序列一一对应。

[0013] 本发明还提出一种诊断名称标准化的智能匹配系统，应用于如上述中任一项所述的诊断名称标准化的智能匹配方法；所述系统包括：

标准术语字典建模模块，用于获取标准化的字典数据，国际疾病分类编码及常用诊断文本数据；基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库；对分词后形成的所述语料库进行词嵌入训练，以形成词向量表；对词向量表进行加权以得到医学术语的训练句子向量；构建机器学习模型；基于标记数据集将词向量表作为所述机器学习模型的输入，将训练句子向量作为所述机器学习模型的输出，以对所述机器学习模型进行训练；

推荐模块，用于获取对语料库中的文本序列进行人工标注后得到的机器学习模型输入所需的标记数据集；基于实际句子向量生成与实际诊断文本数据对应的候选术语；

用户画像模块，用于获取用户通过用户终端输入的实际诊断文本，将所述实际诊断文本输入完成训练的所述机器学习模型，以得到实际句子向量。

[0014] 通过上述技术方案，能实现以下有益效果：

本发明提出的诊断名称标准化的智能匹配方法提出了一种基于自然语言处理和医学知识库，由自然语言诊断名称到ICD10标准诊断编码及名称的智能匹配技术，以提高医学文本的结构化水平；本发明可以高效快速的实现对医师输入的杂乱无序的振动术语达到标准有序的后结构化目标，能够更加高效的帮助医师将手写的电子病历进行结构化匹

配;在医院的实际挂号和会诊过程中发挥落地能力,降低因医师主观疏忽原因导致的误诊事故,从而推进数字医疗产业化。

附图说明

[0015] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图示出的结构获得其他的附图。

[0016] 图1为本发明提出的一种诊断名称标准化的智能匹配方法第一实施例的流程图。

具体实施方式

[0017] 应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。

[0018] 本发明提出一种诊断名称标准化的智能匹配方法及系统。

[0019] 如附图1所示,在本发明提出的一种诊断名称标准化的智能匹配方法的一实施例中,本实施例包括如下步骤:

步骤S110:获取标准化的字典数据,国际疾病分类编码(例如ICD10标准诊断编码)以及常用诊断文本数据。

[0020] 具体的,这里的常用诊断文本数据为用于作为后续模型训练的常见病历诊断历史数据;包括病名、症状、患者信息、药物使用情况、检查数据等。

[0021] 此外,获取标准化的字典数据的过程中,需要针对临床版、北京版、广东版等各种版本进行去重、纠错等数据预处理,即字典治理,以得到标准化的字典数据。

[0022] 步骤S120:基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词,以形成语料库。

[0023] 步骤S130:对分词后形成的所述语料库进行词嵌入训练,以形成词向量表。

[0024] 具体的,通过Python中的Gensim模块对分词后形成的所述语料库进行词嵌入训练,以形成词向量表。

[0025] 步骤S140:对词向量表进行加权以得到医学术语的训练句子向量。

[0026] 步骤S150:构建机器学习模型。

[0027] 步骤S160:获取对语料库中的文本序列进行人工标注后得到的机器学习模型输入所需的标记数据集。

[0028] 具体的,对语料库中的文本序列进行人工标注,这里的标注参考临床医师的指导,以得到机器学习模型输入所需的标记数据集;从而为监督学习的机器学习模型提供学习标注。

[0029] 步骤S170:基于标记数据集将词向量表作为所述机器学习模型的输入,将训练句子向量作为所述机器学习模型的输出,以对所述机器学习模型进行训练。

[0030] 步骤S180:获取用户通过用户终端(例如个人电脑)输入的实际诊断文本,将所述实际诊断文本输入完成训练的所述机器学习模型,以得到实际句子向量。

[0031] 步骤S190:基于实际句子向量生成与实际诊断文本数据对应的候选术语。

[0032] 本发明提出的诊断名称标准化的智能匹配方法提出了一种基于自然语言处理和

医学知识库，由自然语言诊断名称到ICD10标准诊断编码及名称的智能匹配技术，以提高医学文本的结构化水平；本发明可以高效快速的实现对医师输入的杂乱无序的振动术语达到标准有序的后结构化目标，能够更加高效的帮助医师将手写的电子病历进行结构化匹配；在医院的实际挂号和会诊过程中发挥落地能力，降低因医师主观疏忽原因导致的误诊事故，从而推进数字医疗产业化。

[0033] 在本发明提出的一种诊断名称标准化的智能匹配方法的第二实施例中，基于第一实施例，步骤S120，包括如下步骤：

步骤S210：获取医学术语表。

[0034] 步骤S220：基于医学术语表生成自定义词典。

[0035] 步骤S230：基于字典数据、国际疾病分类编码和自定义词典对常用诊断文本数据进行分词，以形成语料库。

[0036] 具体的，针对某些专业术语可能分词错误的情况，根据搜集到的医学术语表，如ICD-10编码术语表、ICD-9-CM-3编码术语表等，整理相关词汇形成自定义词典，从而提升分词的准确性。

[0037] 在本发明提出的一种诊断名称标准化的智能匹配方法的第三实施例中，基于第一实施例，步骤S190，之后还包括如下步骤：

步骤S310：对人工干预过后的候选术语采用协同过滤算法进行过滤，以生成初始推荐结果。

[0038] 例如：根据目标疾病(如多发性骨髓瘤)的临床名称，制定最宽泛的抽提字段，如“骨髓瘤”、“卡勒”、“骨髓癌”、“骨髓病”、“203.0”、“C90.0”和“M97320/3”等，将全部潜在目标疾病患者都进行抽提。然后结合计算机辅助筛选开始匹配工作，每一轮匹配结果都会进行计算机辅助的人工干预(及核查工作)，核查的原则是每一例确诊目标患者，必须有明确的诊断词条(与标准词列表即纳入词库相同，这里的纳入词库即是已经核查正确的确诊目标患者所对应的诊断词条所对应的词库)，每一例确诊非目标患者，必须符合明确的排除词条(与非目标疾病的排除词库相同；这里的排除词库是由在人工干预过程中人工判断出的非多发性骨髓瘤诊断的词语组成的)，循环反复，直至记录全部明确界定为止。

[0039] 当每次人工核查显示当前版本的标准词表准确率低于95％时，将此次核查得到的新词加入生成新版本的标准词列表，重新对编辑距离与相似度进行计算后进行下一轮人工核查；直至人工核查结果显示词表准确率达到95％，迭代停止。

[0040] 这里的准确率是指：在利用形成的纳入词库和排除词库对医保数据库里的患者进行是否为多发性骨髓瘤患者的判断；通过判断，医保数据库中一部分患者将被准确判定为多发性骨髓瘤患者，其人数记为a，一部分患者会被确定为非多发性骨髓瘤患者，其人数记为b，还会剩下一部分患者仍然需要进一步判断，其人数记为c。则准确率记为a/(a+c)。

[0041] 在本发明提出的一种诊断名称标准化的智能匹配方法的第四实施例中，基于第三实施例，步骤S310，包括如下步骤：

步骤S410：获取用户通过用户终端输入的选择数量。

[0042] 步骤S420：将所述候选术语的数量设置为所述选择数量。

[0043] 具体的，用户可以自行选择生成的候选术语的数量，以满足不同用户的使用习惯。具体的，这里的用户为专业医生，故专业医生可以根据实际业务场景调整相似度的顺序，将

最符合的候选术语置顶,从而起到督选的作用。

[0044] 在本发明提出的一种诊断名称标准化的智能匹配方法的第五实施例中,基于第三实施例,步骤S320,之后还包括如下步骤:

步骤S510:采用云原生向量数据库(例如Milvus框架)对初始推荐结果进行向量检索以得到倒排索引。

[0045] 步骤S520:根据索引编号获取初始推荐结果对应的key-value数据库(例如Redis)中存储的具体术语缓存数据。

[0046] 步骤S530:将具体术语缓存数据推荐给用户。

[0047] 具体的,这里的具体术语即为最终返回给用户的与用户输入的实际诊断文本对应的匹配的结构化文本,用户选择这些结构化文本从而使得输入的诊断文本更加标准,避免出现诊断文本错误描述。

[0048] 在本发明提出的一种诊断名称标准化的智能匹配方法的第六实施例中,基于第五实施例,步骤S530,包括如下步骤:

步骤S610:将具体术语缓冲数据通过用户终端的显示界面以下拉框的形式进行显示。

[0049] 具体的,这里将具体术语缓冲数据通过用户终端的显示界面(显示屏)以下拉框的形式进行显示;从而更加便捷的将具体术语缓存数据推荐给用户。

[0050] 在本发明提出的一种诊断名称标准化的智能匹配方法的第七实施例中,基于第一实施例,步骤S120,还包括如下步骤:

步骤S710:通过Python的第三方分词模块(例如Jieba分词模块)基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词,以形成语料库。

[0051] 具体的,jieba库是一个简单实用的中文自然语言处理分词库。jieba分词属于概率语言模型分词。概率语言模型分词的任务是:在全切分所得的所有结果中求某个切分方案S,使得P(S)最大。

[0052] Jieba分词支持三种分词模式:

全模式:把句子中所有的可以成词的词语都扫描出来, 速度非常快,但是不能解决歧义。

[0053] 精确模式:试图将句子最精确地切开,适合文本分析。

[0054] 搜索引擎模式,在精确模式的基础上,对长词再次切分,提高召回率,适合用于搜索引擎分词。

[0055] 在本发明提出的一种诊断名称标准化的智能匹配方法的第八实施例中,基于第一实施例,步骤S170中的基于标记数据集将词向量表作为所述机器学习模型的输入的步骤,包括如下步骤:

步骤S810:将词向量表的文本序列中的各个单词,通过词向量表映射为对应的词向量序列。

[0056] 步骤S820:基于所述标记数据集将所述词向量序列作为所述机器学习模型的输入。

[0057] 将分好词的诊断文本使用现有的gensim方法包进行词向量的训练,首先使用编辑距离对相似词语进行第一次筛选匹配,选出字面上最相似的词语,再对筛选出的结果进行

第二次筛选，使用余弦距离对词语之间关联性进行计算，经过组合距离计算，最终得到最优的相似词语。具体步骤如下：

1、设置编辑距离，求解出实际诊断文本数据与纳入词库最相似的正序排序，并将其切分成相应的词集合。编辑距离主要用来计算两个字符串的相似度，其定义如下：

设有字符串A和B，B为模式串，现给定以下操作：从字符串中删除一个字符；从字符串中插入一个字符；从字符串中替换一个字符。通过以上三种操作，将字符串A编辑为模式串B所需的最小操作数称为A和B的最短编辑距离，编辑距离越小，说明两个字符串越相似。反之，越不相似。

[0058] 2、将上述步骤的结果使用余弦距离求得两个字符串之间的关联性，并设置阈值，(具体实施时，经过多次调试，阈值可设置为0.6，匹配的速度和准确率相对最佳)如果小于阈值则两个字符串之间的关联性为0，即认为无关联，将关联的词距加和正序排序，求解出次优先级的相似词。

[0059] 具体的，余弦距离余弦相似度用向量空间中两个向量的夹角的余弦值来衡量两个文本间的相似度，相比距离度量，余弦相似度更加注重两个向量在方向上的差异。

[0060] 在本发明提出的一种诊断名称标准化的智能匹配方法的第九实施例中，基于第一实施例，所述标记数据集包括标记名；所述标记名和所述文本序列一一对应；

具体的，语料库中的文本序列交给专业医师进行人工标注，将各个文本序列对应设置标记名；诸如：“多发性骨髓瘤”、“疑似多发性骨髓瘤”、“原发性骨髓瘤”等的标记名；通过进行标记，以提升对后续的机器学习模型的训练精准度。

[0061] 本发明还提出一种诊断名称标准化的智能匹配系统，本系统应用于如上述中任一项所述的诊断名称标准化的智能匹配方法；所述系统包括：

标准术语字典建模模块，用于获取标准化的字典数据，国际疾病分类编码及常用诊断文本数据；基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库；对分词后形成的所述语料库进行词嵌入训练，以形成词向量表；对词向量表进行加权以得到医学术语的训练句子向量；构建机器学习模型；基于标记数据集将词向量表作为所述机器学习模型的输入，将训练句子向量作为所述机器学习模型的输出，以对所述机器学习模型进行训练；

推荐模块，用于获取对语料库中的文本序列进行人工标注后得到的机器学习模型输入所需的标记数据集；基于实际句子向量生成与实际诊断文本数据对应的候选术语；

用户画像模块，用于获取用户通过用户终端输入的实际诊断文本，将所述实际诊断文本输入完成训练的所述机器学习模型，以得到实际句子向量。

[0062] 上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

[0063] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本发明各个实施例所述的方法。

[0064] 上面结合附图对本发明的实施例进行了描述，但是本发明并不局限于上述的具体

实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本发明的启示下，在不脱离本发明宗旨和权利要求所保护的范围情况下，还可做出很多形式，这些均属于本发明的保护之内。


获取标准化的字典数据，国际疾病分类编码以及常用诊断文本数据
S110
基于字典数据和国际疾病分类编码对常用诊断文本数据进行分词，以形成语料库
S120
对分词后形成的所述语料库进行词嵌入训练，以形成词向量表
S130
对词向量表进行加权以得到医学术语的训练句子向量
S140
构建机器学习模型
S150
获取对语料库中的文本序列进行人工标注后得到的机器学习模型输入所需的标记数据集
S160
基于标记数据集将词向量表作为所述机器学习模型的输入，将训练句子向量作为所述机器学习模型的输出，以对所述机器学习模型进行训练
S170
获取用户通过用户终端输入的实际诊断文本，将所述实际诊断文本输入完成训练的所述机器学习模型，以得到实际句子向量
S180
基于实际句子向量生成与实际诊断文本数据对应的候选术语
S190

图1