



(10) 授权公告号 CN 113348515 B

(45) 授权公告日 2024. 09. 10

(21) 申请号 201980090257.1

(22) 申请日 2019.11.22

(65) 同一申请的已公布的文献号

申请公布号 CN 113348515 A

(43) 申请公布日 2021.09.03

(30) 优先权数据

62/771,308 2018.11.26 US

(85) PCT国际申请进入国家阶段日

2021.07.26

(86) PCT国际申请的申请数据

PCT/EP2019/082288 2019.11.22

(87) PCT国际申请的公布数据

W02020/109177 EN 2020.06.04

(73) 专利权人 阿尔戈泰克系统有限公司

地址 以色列赖阿南纳

(72) 发明人 O·阿尔默

(74) 专利代理机构 北京市金杜律师事务所

11256

专利代理师 董莘

(51) Int.Cl.

G16H 15/00 (2006.01)

(56) 对比文件

US 2018293227 A1, 2018.10.11

审查员 桑晓慧

权利要求书5页 说明书21页 附图4页

(54) 发明名称

用于匹配放射报告中的医学概念的系统和
方法

(57) 摘要

一种确定医学概念集合中的哪些概念与输入文本有关的方法,包括:a) 为每个概念创建查询集合,每个查询是概念中的两个字的串;b) 对于每个查询,确定输入文本是否包括该查询的所有字,并且计算指示查询与输入文本之间的匹配程度的子得分;c) 对于有足够多的查询的字在输入文本中足够靠近的每个概念,取决于子得分来计算得分;以及d) 取决于概念的得分,确定已经被计算得分的概念中的哪些概念与输入文本有关,而哪些概念与输入文本不相关。

1. 一种通过在计算机上执行指令来自动确定医学概念集中的哪些概念与输入文本有关的方法,所述方法包括:

a) 针对具有两个或更多个字的每个概念创建一个或多个查询的集合,每个查询是两个字的串,所述两个字的串是所述概念中的所述字的子串,所述串中的所述两个字的顺序与所述概念中的所述字相同;

b) 对于具有两个或更多个字的所述概念的选定子集中的每个概念,针对每个查询,确定所述输入文本是否包括该查询的所有所述字,并且如果所述输入文本包括该查询的所有所述字,则计算指示所述查询与所述输入文本之间的匹配程度的子得分;

c) 对于所述选定子集中的如下每个概念,其中针对该概念,超过最小数目的所述查询的所有字根据准则在所述输入文本中足够靠近,取决于针对该概念的所述查询的所述子得分,计算用于指示所述概念与所述输入文本之间的匹配程度的得分,其中所述查询中的字在所述输入文本中足够靠近的所述准则指定所述字之间的最大距离;以及

d) 应用一个或多个规则来确定所述选定子集中的被计算出得分的所述概念中哪些概念与所述输入文本有关,而哪些概念与所述输入文本不相关,所述规则中的至少一些规则取决于所述概念的所述得分。

2. 根据权利要求1所述的方法,还包括:至少针对所述选定子集中具有两个或更多个字的如下每个概念,计算匹配得分,该概念的所述两个或更多个字中有超过最小数目的所述字在所述输入文本中,并且针对该概念的超过最小数目的所述查询的所有字根据所述准则在所述输入文本中足够靠近,所述匹配得分指示所述概念与所述输入文本之间的根据词袋方法的匹配程度,并且其中计算针对每个概念的所述得分包括:取决于该概念的所述匹配得分以及针对该概念的所述查询的所述子得分来计算所述得分。

3. 根据权利要求2的方法,其中计算针对每个概念的所述得分包括:计算所述匹配得分和针对所述查询的所述子得分的加权和。

4. 根据权利要求2所述的方法,其中字的所述最小数目对于具有两个字的概念为2,对于具有三个字的概念为2或3,对于具有四个字的概念为2、3或4,对于具有五个字的概念为3或4,对于具有六个字的概念为3、4或5。

5. 根据权利要求2所述的方法,还包括:当在所述输入文本中找到一个字时,向只有该一个字的所述概念分配得分,其中取决于所述概念的所述得分的所述规则被应用于只有该一个字的所述概念和具有两个或更多个字的所述概念。

6. 根据权利要求1或权利要求5所述的方法,其中所述一个或多个规则指定:当已经计算出得分的两个概念的字具有足够大的重叠时,具有较低得分的所述概念与所述输入文本不相关,并且如果概念具有的计算出的得分高于如下针对任何其他概念的计算出的得分,其中所述概念的字与所述任何其他概念的字具有足够大的重叠,则所述概念与所述输入文本相关。

7. 根据权利要求1的方法,其中所述最小数目包括第一最小数目,所述第一最小数目在针对该概念创建的查询的数目的35%到65%之间。

8. 根据权利要求1所述的方法,其中计算所述得分包括计算所述查询的所述子得分的加权和,其中较低权重被赋给如下查询,所述查询具有所述概念中的相距更远的字。

9. 根据权利要求1所述的方法,其中概念的所述选定子集至少排除如下概念,针对所述

概念,在所述输入文本中没有找到被定义为概念的强制字的一个或多个字,所述强制字是必须出现在所述输入文本中的字。

10.根据权利要求9所述的方法,其中在所述概念集中的所有字中最稀有的所述概念中的所述字被定义为强制字。

11.根据权利要求9所述的方法,其中作为疾病名称的所述概念中的任何单数字被定义为强制字。

12.根据权利要求1的方法,其中概念的所述选定子集至少排除包括针对身体部位的字和描述所述身体部位的位置或方向的字的那些概念,针对那些概念,描述身体部位的所述位置或方向的所述字在所述输入文本中与针对所述身体部位的所述字相距一个以上的字。

13.根据权利要求1所述的方法,其中,所述输入文本中的顺序与在所述查询中不同的字的所述最大距离更小,并且与在所述概念中间隔更近的字相比,在所述概念中间隔更远的字的所述最大距离更大。

14.根据权利要求13的方法,其中在所述概念中相邻并且在所述查询和所述输入文本中顺序相同的字的所述最大距离在10与25之间。

15.根据权利要求1所述的方法,还包括准备所述概念集合并且预处理所述输入文本,包括:

a)从医学概念数据库中提供初始概念集合;

b)通过扩展椎骨字母数字名称以包括“椎骨”和“脊椎”并且将所述字母数字名称中的字母替换为所述字母代表的身体区域,来修改所述初始概念集合,所述身体区域表示颈椎、胸椎或腰椎;以及

c)通过扩展椎骨字母数字名称以包括“椎骨”和“脊椎”并且将所述字母数字名称中的字母替换为所述字母代表的所述身体区域,来预处理所述输入文本,所述身体区域表示颈椎、胸椎或腰椎。

16.一种计算机存储产品,具有至少一个计算机存储介质,所述至少一个计算机存储介质中存储有指令,所述指令使一个或多个计算机执行根据权利要求1所述的方法。

17.一种计算机存储介质,其中存储有指令,所述指令用于使计算机执行根据权利要求1所述的方法。

18.一种体现在计算机可读介质中的计算机产品,用于执行根据权利要求1所述的步骤。

19.一种用于自动标识在医学报告中找到医学概念集合中的哪些概念的系统,所述系统包括:

a)第一数据库,提供对一个或多个医学报告的访问;

b)拆分器模块,能够访问所述第一数据库并且将医学报告划分成输入文本;

c)第二数据库,提供对所述医学概念集合的访问;

d)处理器模块,能够访问所述输入文本和所述第二数据库,并且被配置为:针对每个输入文本

1)针对具有两个或更多个字的每个概念创建一个或多个查询的集合,每个查询是两个字的串,所述两个字的串是所述概念中的所述字的子串,所述串中的所述两个字的顺序与所述概念中的所述字相同;

2) 对于具有两个或更多个字的所述概念的选定子集中的每个概念, 针对每个查询, 确定所述输入文本是否包括该查询的所有所述字, 并且如果所述输入文本包括该查询的所有所述字, 则计算指示所述查询与所述输入文本之间的匹配程度的子得分;

3) 对于所述选定子集中准则的如下每个概念, 其中针对该概念, 超过最小数目的所述查询的所有字根据准则在所述输入文本中足够靠近, 取决于针对该概念的所述查询的所述子得分, 计算用于指示所述概念与所述输入文本之间的匹配程度的得分; 以及

4) 应用一个或多个规则来确定所述选定子集中的被计算出得分的所述概念中哪些概念与所述输入文本有关, 而哪些概念与所述输入文本不相关, 所述规则中的至少一些规则取决于所述概念的所述得分; 以及

e) 输出模块, 输出被确定为与所述医学报告的所述输入文本有关的所述概念。

20. 一种通过在计算机上执行指令来自动确定医学概念集合中的哪些概念与输入文本有关的方法, 所述方法包括:

a) 对于所述集合中的每个概念, 应用一个或多个准则来确定所述概念是否是所述输入文本的可能匹配, 其中对于所述概念中的至少一些概念, 所述准则不要求在所述输入文本中找到所述概念的所有字, 但要求在所述输入文本中找到被指定为所述概念的强制字的一个或多个字, 所述强制字是必须出现在所述输入文本中的字; 以及

b) 应用一个或多个规则来确定作为所述输入文本的可能匹配的所述概念中的哪些概念与所述输入文本有关, 而哪些概念与所述输入文本不相关。

21. 根据权利要求20所述的方法, 还包括针对作为所述输入文本的可能匹配的每个概念, 计算指示所述概念与所述输入文本之间的匹配程度的得分, 其中所述一个或多个规则中的至少一些规则取决于所述概念的得分和作为所述输入文本的可能匹配的任何其他概念的所述得分。

22. 根据权利要求20的方法, 其中所述概念的在所述集合中的所有所述概念中最不常见的字被指定为所述概念的强制字。

23. 根据权利要求20所述的方法, 其中概念的如下字被指定为所述概念的强制字, 所述字是针对疾病的单字项。

24. 根据权利要求20的方法, 其中对于包括针对身体部位的字和描述所述身体部位的位置或方向的字的概念, 这两种字都被指定为所述概念的强制字, 并且所述准则进一步要求所述字在所述输入文本中彼此相邻。

25. 一种通过在计算机上执行指令来自动确定医学概念集合中的哪些概念与输入文本有关的方法, 所述方法包括:

a) 对于所述集合中的每个概念, 应用准则来确定所述概念是否是所述输入文本的可能匹配;

b) 对于作为所述输入文本的可能匹配的每个概念, 计算指示所述概念与所述输入文本之间的匹配程度的得分; 以及

c) 当已经计算出得分的两个概念的字具有足够大的重叠时, 确定具有较低得分的所述概念与所述输入文本不相关, 并且如果概念的计算得分高于如下任何其他概念的计算出的得分, 其中所述概念的字与所述任何其他概念的字具有足够大的重叠, 则确定所述概念与所述输入文本相关,

其中对于所述集合中的所述概念中的至少一些概念,计算所述得分包括:计算匹配得分,计算match_phrase得分并且合并所述匹配得分和所述match_phrase得分以获取所述得分,所述匹配得分取决于在所述输入文本中找到所述概念中的所述字的数目,而不取决于那些字在所述输入文本中的顺序,并且所述match_phrase得分取决于在所述输入文本中找到所述概念中的所述字的数目,并且取决于那些字在所述输入文本中的顺序。

26.一种用于自动标识在医学报告中找到医学概念集合中的哪些概念的系统,所述系统包括:

- a) 第一数据库,提供对一个或多个医学报告的访问;
- b) 拆分器模块,能够访问所述第一数据库并且将医学报告划分成输入文本;
- c) 第二数据库,提供对所述医学概念集合的访问;
- d) 处理器模块,能够访问所述输入文本和所述第二数据库,并且被配置为:针对每个输入文本

1) 对于所述集合中的每个概念,应用一个或多个准则来确定所述概念是否是所述输入文本的可能匹配,其中对于所述概念中的至少一些概念,所述准则不要求在所述输入文本中找到所述概念的所有字,但确实要求在所述输入文本中找到被指定为所述概念的强制字的一个或多个字,所述强制字是必须出现在所述输入文本中的字;以及

2) 应用一个或多个规则来确定作为所述输入文本的可能匹配的所述概念中的哪些概念与所述输入文本有关,而哪些概念与所述输入文本不相关;以及

e) 输出模块,输出被确定为与所述医学报告的所述输入文本有关的所述概念。

27.一种用于自动标识在医学报告中找到医学概念集合中的哪些概念的系统,所述系统包括:

- a) 第一数据库,提供对一个或多个医学报告的访问;
- b) 拆分器模块,能够访问所述第一数据库并且将医学报告划分成输入文本;
- c) 第二数据库,提供对所述医学概念集合的访问;
- d) 处理器模块,能够访问所述输入文本和所述第二数据库,并且被配置为:针对每个输入文本

1) 对于所述集合中的每个概念,应用准则来确定所述概念是否是所述输入文本的可能匹配;

2) 对于作为所述输入文本的可能匹配的每个概念,计算指示所述概念与所述输入文本之间的匹配程度的得分;以及

3) 当已经计算出得分的两个概念的字具有足够大的重叠时,确定具有较低得分的所述概念与所述输入文本不相关,并且如果概念的计算出的得分高于如下任何其他概念的计算得分,其中所述概念的字与所述任何其他概念的字具有足够大的重叠,则确定所述概念与所述输入文本相关;

其中对于所述集合中的所述概念中的至少一些概念,计算所述得分包括:计算匹配得分,计算match_phrase得分并且合并所述匹配得分和所述match_phrase得分以获取所述得分,所述匹配得分取决于在所述输入文本中找到所述概念中的所述字的数目,而不取决于那些字在所述输入文本中的顺序,并且所述match_phrase得分取决于在所述输入文本中找到所述概念中的所述字的数目,并且取决于那些字在所述输入文本中的顺序,以及

e) 输出模块,输出被确定为与所述医学报告的所述输入文本有关的所述概念。

用于匹配放射报告中的医学概念的系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求以Or Almer的名义于2018年11月26日临时提交的题为“SYSTEM AND METHOD FOR MATCHING MEDICAL CONCEPTS IN RADIOLOGICAL REPORTS”、申请号为62/771,308的美国临时申请的权益,该临时申请整体并入本文。

技术领域

[0003] 本公开总体上涉及自然语言处理领域,并且具体地涉及标识在医学报告中找到预定义医学概念集合中的哪些医学概念。更具体但非排他地,本公开涉及用于针对放射学报告进行这一操作的方法。

背景技术

[0004] 存在多种搜索算法用于自动确定在给定输入文本中是否找到给定概念。例如,该概念可以是输入到搜索引擎中的搜索项、字串,并且输入文本可以是在搜索引擎正在搜索的大量网页之一上找到的文本。在某些搜索算法中,不需要以完全相同的形式在输入文本中找到搜索项,以便搜索算法返回肯定结果,但搜索项中的某些字可能会从输入文本中遗漏,或者在输入文本中,搜索项的字可能以不同顺序出现,和/或在它们之间有其他字。在这种情况下,搜索算法可以为该搜索项和输入文本计算得分,该得分指示匹配的良好程度,并且当搜索引擎完成搜索它正在搜索的网页时,它可以提供为该搜索项找到正面结果的网页的列表,这些网页按照得分顺序排序。

[0005] 例如,ElasticSearch是一种在[https://www\(dot\)elastic\(dot\)co/products/elasticsearch](https://www(dot)elastic(dot)co/products/elasticsearch)有描述的商用搜索引擎,其中可以在[https://www\(dot\)elastic\(dot\)co/guide/en/elasticsearch/guide/current/index\(dot\)html](https://www(dot)elastic(dot)co/guide/en/elasticsearch/guide/current/index(dot)html)找到用户指南。ElasticSearch通过使用不同类型的查询来分析文本字段,诸如文档的标题字段,以计算文档与搜索项的相关性得分。在“匹配”查询中,如果在文本字段中找到搜索项中的超过最小数目或超过最小百分比的字,则文本字段被认为与搜索项匹配,并且如果在文本字段中找到搜索项中的更多字,则“匹配”相关性得分更高。ElasticSearch的“匹配”查询是输入文本字段中的字顺序无关紧要的“词袋”方法的一个示例。

[0006] 在ElasticSearch中的“match_phrase(短语匹配)”查询中,如果在文本字段中找到搜索项的所有字,并且搜索项中字的相对位置与文本字段中相同字的相对位置相距不太远,则文本字段被认为与搜索项匹配。字的位置相距多远是指搜索项中字的位置必须进行多少次改变(其中每次将字的位置在任一方向上改变1)以使字与其在文本字段中的位置相比具有相同位置。例如,如果搜索项由两个字组成,并且在文本字段中找到这两个字彼此相邻但顺序相反,则位置改变次数为2,因为这两个字中的每个都必须经历一次位置改变。如果在文本字段中找到这两个字的顺序与搜索项中的顺序相同,但它们之间有n个其他字,则所需要的位置改变次数为n。为了使文本字段被认为是搜索项的匹配,所需要的位置改变次数必须不大于最大数目(称为“slop”的参数)。如果所需要的位置改变次数较小,则“match_

phrase”查询可能具有较高相关性得分。可以将“match”相关性得分和“match_phrase”相关性得分相加,以获取文本字段与搜索项的整体相关性得分。

[0007] 已知搜索算法将文本字段和/或搜索项中的字替换为其词干。例如,ElasticSearch具有英语抽词干特征,其可以对英语文本执行这个操作。

[0008] Savova等人的“Mayo clinical Text Analysis and Knowledge Extraction System(cTAKES):architecture,component evaluation and applications”(J.Am.Med.Inform.Assoc.2010:17:507-513(doi:10.1136/jamia.2009.001560))描述了用于从电子病历临床自由文本中提取信息的cTAKES系统,并且提供了具有这个目标的其他系统的历史。

[0009] 在[https://browser\(dot\)ihtsdotools\(dot\)org](https://browser(dot)ihtsdotools(dot)org)找到的SNOMED CT医学词典具有大量医学概念的结构化列表,包括可以用于诊断疾病的身体部位、疾病和症状。每个概念由一个或多个字构成的串来定义。在某些情况下,多个概念指代同一事物,在这种情况下,给出优选术语和替代术语。概念可以具有亲代和子代,其指代更一般和更具体的概念,例如肺是肺部分的亲代,肺部分是肺叶的亲代。疾病概念与身体部位概念具有发现部位关系,例如肺癌与肺具有发现部位关系。

发明内容

[0010] 本发明的一些实施例的一个方面涉及一种用于确定医学概念集合中的哪些医学概念与来自医学报告的输入文本有关的方法,其中如果在输入文本中找到概念中的足够大量的字对彼此足够靠近,则认为概念是输入文本的可能匹配。

[0011] 根据本公开的一个方面,提供了一种通过在计算机上执行指令来自动确定医学概念集合中的哪些概念与输入文本有关的方法,该方法包括:

[0012] a) 为具有两个或更多个字的每个概念创建一个或多个查询的集合,每个查询是两个字的串,所述两个字的子串是概念中的字的子串,串中的两个字的顺序与概念中的字相同;

[0013] b) 对于具有两个或更多个字的概念的选定子集中的每个概念,对于每个查询,确定输入文本是否包括查询的所有字,并且如果输入文本包括查询的所有字,则计算指示查询与输入文本之间的匹配程度的子得分;

[0014] c) 对于选定子集中超过最小数目的查询的所有字根据准则在输入文本中足够靠近的每个概念,取决于概念的查询的子得分,计算用于指示概念与输入文本之间的匹配程度的得分;以及

[0015] d) 应用一个或多个规则来确定选定子集中的概念中已经被计算得分的哪些概念与输入文本有关,而哪些概念与输入文本不相关,规则中的至少一些规则取决于概念的得分。

[0016] 可选地,该方法还包括至少计算选定子集中具有两个或更多个字的每个概念的匹配得分,两个或更多个字中有超过最小数目的字在输入文本中,并且概念的超过最小数目的查询的所有字根据准则在输入文本中足够靠近,匹配得分指示概念与输入文本之间根据词袋方法的匹配程度,并且其中计算每个概念的得分包括取决于概念的匹配得分以及概念的查询的子得分来计算得分。

[0017] 可选地,计算每个概念的得分包括计算匹配得分和查询的子得分的加权和。

[0018] 可选地,最小字数对于具有两个字的概念为2,对于具有三个字的概念为2或3,对于具有四个字的概念为2、3或4,对于具有五个字的概念为3或4,对于具有六个字的概念为3、4或5。

[0019] 可选地,该方法还包括当在输入文本中找到一个字时,为只有单字概念分配得分,其中取决于概念的得分的规则被应用于只有单字概念和具有两个或更多个字的概念。

[0020] 可选地,该一个或多个规则指定:当已经被计算得分的两个概念的字中具有足够大的重叠时,具有较低得分的概念与输入文本不相关,并且如果概念的计算得分高于任何其他概念的计算得分并且概念的字与任何其他概念的字具有足够大的重叠,则概念确实与输入文本有关。

[0021] 可选地,第一最小数目在为概念而创建的查询数的35%到65%之间。

[0022] 可选地,计算得分包括计算查询的子得分的加权和,其中较低权重被给予在概念中具有相距更远的字的查询。

[0023] 可选地,选定概念子集至少排除在输入文本中没有找到被定义为概念的强制字的一个或多个字的概念。

[0024] 可选地,在概念集合中的所有字中最稀有的概念中的字被定义为强制字。

[0025] 可选地,作为疾病名称的概念中的任何单数字被定义为强制字。

[0026] 可选地,选定概念子集至少排除包括身体部位的字和描述身体部位的位置或方向的字的概念,概念的描述身体部位的位置或方向的字在输入文本中与身体部位的字相距一个以上的字。

[0027] 可选地,查询中的字在输入文本中足够靠近的准则指定字之间的最大距离,对于在输入文本中与在查询中顺序不同的字,最大距离较小,并且与对于在概念中间隔较近的字相比,对于在概念中间隔较远的字,最大距离较大。

[0028] 可选地,在概念中相邻并且在查询和输入文本中顺序相同的字的最大距离在10到25之间。

[0029] 可选地,该方法还包括准备概念集合并且预处理输入文本,包括:

[0030] a) 从医学概念数据库中提供初始概念集合;

[0031] b) 通过扩展椎骨字母数字名称以包括字“椎骨”和“脊椎”并且将字母数字名称中的字母替换为字母代表的身体区域(颈椎、胸椎或腰椎)来修改初始概念集合;以及

[0032] c) 通过扩展椎骨字母数字名称以包括字“椎骨”和“脊椎”并且将字母数字名称中的字母替换为字母代表的身体区域(颈椎、胸椎或腰椎)来预处理输入文本。

[0033] 根据本公开的另一方面,提供了一种计算机存储产品,该计算机存储产品具有至少一个计算机存储介质,至少一个计算机存储介质中存储有指令,该指令引起一个或多个计算机执行本发明的示例性方法。

[0034] 根据本公开的另一方面,提供了一种其中存储有助于引起计算机执行本发明的示例性方法的指令的计算机存储介质。

[0035] 根据本公开的另一方面,提供了一种体现在计算机可读介质中的计算机产品,该计算机产品用于执行本发明的示例性方法的步骤。

[0036] 根据本公开的另一方面,提供了一种用于自动标识在医学报告中找到医学概念集

合中的哪些概念的系统,该系统包括:

[0037] a) 第一数据库,提供对一个或多个医学报告的访问;

[0038] b) 拆分器模块,能够访问第一数据库并且将医学报告分成输入文本;

[0039] c) 第二数据库,提供对医学概念集合的访问;

[0040] d) 处理器模块,能够访问输入文本和第二数据库,针对每个输入文本被配置为:

[0041] 1) 为具有两个或更多个字的每个概念创建一个或多个查询的集合,每个查询是作为概念中的字的子串的两个字的串,串中的两个字的顺序与概念中的字相同;

[0042] 2) 对于具有两个或更多个字的概念的选定子集中的每个概念,对于每个查询,确定输入文本是否包括查询的所有字,并且如果输入文本包括查询的所有字,则计算指示查询与输入文本之间的匹配程度的子得分;

[0043] 3) 对于选定子集中超过最小数目的查询的所有字根据准则在输入文本中足够靠近的每个概念,取决于概念的查询的子得分,计算用于指示概念与输入文本之间的匹配程度的得分;以及

[0044] 4) 应用一个或多个规则来确定选定子集中的概念中已经被计算得分的哪些概念与输入文本有关,而哪些概念与输入文本不相关,规则中的至少一些规则取决于概念的得分;以及

[0045] e) 输出模块,输出被确定为与医学报告的输入文本有关的概念。

[0046] 根据本公开的另一方面,提供了一种通过在计算机上执行指令来自动确定医学概念集合中的哪些概念与输入文本有关的方法,该方法包括:

[0047] a) 对于集合中的每个概念,应用一个或多个准则来确定概念是否是输入文本的可能匹配,其中对于概念中的至少一些概念,准则不要求在输入文本中找到概念的所有字,但确实要求在输入文本中找到被指定为概念的强制字的一个或多个字;以及

[0048] b) 应用一个或多个规则来确定作为输入文本的可能匹配的概念中的哪些概念与输入文本有关,而哪些概念与输入文本不相关。

[0049] 可选地,该方法还包括为作为输入文本的可能匹配的每个概念计算指示概念与输入文本之间的匹配程度的得分,其中一个或多个规则中的至少一些规则取决于概念的得分和作为输入文本的可能匹配的任何其他概念的得分。

[0050] 可选地,在集合中的所有概念中最不常见的概念的字被指定为概念的强制字。

[0051] 可选地,作为疾病的单字项的概念的字被指定为概念的强制字。

[0052] 可选地,对于包括身体部位的字和描述身体部位的位置或方向的字的概念,这两个字都被指定为概念的强制字,并且准则进一步要求字在输入文本中彼此相邻。

[0053] 根据本公开的另一方面,提供了一种通过在计算机上执行指令来自动确定医学概念集合中的哪些概念涉及输入文本的方法,该方法包括:

[0054] a) 对于集合中的每个概念,应用准则来确定概念是否是输入文本的可能匹配;

[0055] b) 对于作为输入文本的可能匹配的每个概念,计算指示概念与输入文本之间的匹配程度的得分;以及

[0056] c) 确定:当已经被计算得分的两个概念的字中具有足够大的重叠时,具有较低得分的概念与输入文本不相关,并且如果概念的计算得分高于任何其他概念的计算得分并且概念的字与任何其他概念的字具有足够大的重叠,则概念确实与输入文本有关。

[0057] 可选地,对于集合中的概念中的至少一些概念,计算得分包括计算匹配得分,计算match_phrase得分并且组合匹配得分和match_phrase得分以获取得分,匹配得分取决于在输入文本中找到概念中的字的数目,而不取决于这些字在输入文本中的顺序,并且match_phrase得分取决于在输入文本中找到概念中的字的数目,并且取决于这些字在输入文本中的顺序。

[0058] 根据本公开的另一方面,提供了一种用于自动标识在医学报告中找到医学概念集合中的哪些概念的系统,该系统包括:

[0059] a) 第一数据库,提供对一个或多个医学报告的访问;

[0060] b) 拆分器模块,能够访问第一数据库并且将医学报告分成输入文本;

[0061] c) 第二数据库,提供对医学概念集合的访问;

[0062] d) 处理器模块,能够访问输入文本和第二数据库,针对每个输入文本被配置为:

[0063] 1) 对于集合中的每个概念,应用一个或多个准则来确定概念是否是输入文本的可能匹配,其中对于概念中的至少一些概念,准则不要求在输入文本中找到概念的所有字,但确实要求在输入文本中找到被指定为概念的强制字的一个或多个字;以及

[0064] 2) 应用一个或多个规则来确定作为输入文本的可能匹配的概念中的哪些概念与输入文本有关,而哪些概念与输入文本不相关;以及

[0065] e) 输出模块,输出被确定为与医学报告的输入文本有关的概念。

[0066] 根据本公开的另一方面,提供了一种用于自动标识在医学报告中找到医学概念集合中的哪些概念的系统,该系统包括:

[0067] a) 第一数据库,提供对一个或多个医学报告的访问;

[0068] b) 拆分器模块,能够访问第一数据库并且将医学报告分成输入文本;

[0069] c) 第二数据库,提供对医学概念集合的访问;

[0070] d) 处理器模块,能够访问输入文本和第二数据库,针对每个输入文本被配置为:

[0071] 1) 对于集合中的每个概念,应用准则来确定概念是否是输入文本的可能匹配;

[0072] 2) 对于作为输入文本的可能匹配的每个概念,计算指示概念与输入文本之间的匹配程度的得分;以及

[0073] 3) 确定:当已经计算得分的两个概念的字中具有足够大的重叠时,具有较低得分的概念与输入文本不相关,并且如果概念的计算得分高于任何其他概念的计算得分并且概念的字与任何其他概念的字具有足够大的重叠,则概念确实与输入文本有关;以及

[0074] e) 输出模块,输出被确定为与医学报告的输入文本有关的概念。

[0075] 除非另有定义,否则本文中使用的所有技术和/或科学术语与本发明所属领域的普通技术人员通常理解的含义相同。尽管在本发明的实施例的实践或测试中可以使用与本文中描述的那些方法和材料相似或等效的方法和材料,但下面描述示例性方法和/或材料。如果发生冲突,则以专利说明书(包括定义)为准。此外,材料、方法和示例仅是说明性的,并不旨在进行必要的限制。

[0076] 本发明的实施例的方法和/或系统的实现可以涉及以手动、自动或其组合方式来执行或完成选定任务。此外,根据本发明的方法和/或系统的实施例的实际仪器和设备,若干选定任务可以使用操作系统通过硬件、软件或固件或其组合来实现。

[0077] 例如,根据本发明的实施例用于执行选定任务的硬件可以实现为芯片或电路。作

为软件,根据本发明的实施例的选定任务可以实现为由计算机使用任何合适操作系统执行的多个软件指令。在本发明的示例性实施例中,根据如本文中描述的方法和/或系统的示例性实施例的一个或多个任务由数据处理器执行,诸如用于执行多个指令的计算平台。可选地,数据处理器包括用于存储指令和/或数据的易失性存储器、和/或用于存储指令和/或数据的非易失性存储器,例如磁性硬盘和/或可移动介质。可选地,还提供了网络连接。可选地,还提供了显示器和/或用户输入设备,诸如键盘或鼠标。

附图说明

[0078] 本发明的前述和其他目的、特征和优点将从以下对本发明的实施例的更具体描述中变得明显,如附图所示。附图的元素不必相对于彼此成比例。现在具体详细地参考附图,强调,细节是通过示例的方式并且出于对本发明的实施例的说明性讨论的目的而示出的。在这点上,结合附图进行的描述使本领域技术人员清楚可以如何实践本发明的实施例。

[0079] 在附图中:

[0080] 图1A示出了根据本发明的示例性实施例的用于标识医学报告中的医学概念的系统的框图;

[0081] 图1B示出了根据本发明的任何示例性实施例的由图1A的系统的处理器模块使用的方法的高级流程图;

[0082] 图2示出了根据本发明的示例性实施例的由图1A的系统的处理器模块使用的方法的更详细的流程图;以及

[0083] 图3、图4和图5示出了根据本发明的不同示例性实施例的由图1A的系统的处理器模块使用的方法的高级流程图。

具体实施方式

[0084] 以下是优选实施例的详细描述,参考附图,附图中的相同的附图标记在若干图中的每个中表示相同的结构元件。

[0085] 本公开总体上涉及自然语言处理领域,并且具体地涉及标识在医学报告中找到预定义医学概念集合中的哪些医学概念。更具体但非排他地,本公开涉及用于针对放射学报告进行这一操作的方法。

[0086] 本文中描述的系统(包括用于处理医学报告的计算机系统)和/或方法(例如,由一个或多个处理器执行的代码指令)解决了自动标识在关于患者的医学报告中提到的医学概念的技术问题,与在医学报告中自动标识医学概念的现有技术方法相比,遗漏概念更少并且误报更少。该系统和方法通过以下中的一项或多项解决了这个问题:1) 不要求概念的所有字都存在于输入文本中以便将概念视为可能匹配,即使在进行其中字顺序很重要的“match_phrase”测试时;2) 要求概念中的某些强制字存在于输入文本中来将概念视为可能匹配,即使并非所有字都必须存在;以及3) 为作为输入文本的可能匹配的每个概念计算得分,并且如果存在与该概念显著重叠的具有更高得分的另一概念,则消除作为可能匹配的该概念。

[0087] 本文中描述的系统(包括用于处理医学报告的计算机系统)和/或方法(例如,由一个或多个处理器执行的代码指令)改进了自然语言处理技术领域内的基础技术过程。本文

中描述的系统和/或方法通过以下中的一项或多项改进了标识医学报告中的医学概念的过程:1) 不要求概念的所有字存在于输入文本中以便将概念视为可能匹配;2) 要求必须存在某些强制字;3) 为每个可能匹配计算得分,并且在它们与更高得分的可能概念显著重叠时消除该可能匹配。

[0088] 本文中描述的系统(包括用于处理医学报告的计算机系统)和/或方法(例如,由一个或多个处理器执行的代码指令)提高了执行标识医学报告中的医学概念的代码指令的计算单元的性能。性能的提高增加了医学报告的每个句子中实际存在的被成功标识的医学概念的数目,同时减少了误报、被错误标识为在医学报告的句子中找到的医学概念的数目。性能的提高至少部分通过以下方式来实现:不要求输入文本中存在概念的所有字,这增加了成功找到的概念的数目;要求存在概念的某些强制字,这减少了由于对遗漏字的不良猜测而导致的误报;以及为每个可能概念找到得分并且消除与其他较高得分的可能概念显著重叠的可能概念,这减少了由于在仅存在较窄概念时标识较宽概念而导致的误报。

[0089] 本文中描述的系统(包括用于处理医学报告的计算机系统)和/或方法(例如,由一个或多个处理器执行的代码指令)与现实生活中的物理组件相关联,因为它们处理存储在数据存储介质中的医学报告,并且因为医学报告由医生撰写以用于描述使用医疗诊断设备(例如,医疗成像设备,诸如CT扫描设备和MRI设备)对真实患者进行的医疗检查。该系统和方法通过以下方式改善患者护理:使得可以与必须由医学报告的人类阅读者手动标识医学概念时相比更快地,并且与自动标识医学报告中的医学概念的现有技术的计算机系统相比,更准确地从医学报告中检索医学信息。

[0090] 本文中描述的系统(包括用于处理医学报告的计算机系统)和/或方法(例如,由一个或多个处理器执行的代码指令)提供了一种标识医学报告中的医学概念的独特的、特定的和先进的技术。

[0091] 因此,本文中描述的系统和/或方法与计算机技术和物理硬件密不可分,以克服在医学报告的自然语言处理中出现的实际技术问题。

[0092] 在详细解释本发明的至少一个实施例之前,应当理解,本发明在其应用方面不一定限于在以下描述中给出和/或在附图和/或示例中说明的组件和/或方法的构造和布置的细节。本发明能够有其他实施例或者能够以各种方式实践或执行。

[0093] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括其上具有用于使处理器执行本发明的各方面的计算机可读程序指令的(多个)计算机可读存储介质。

[0094] 计算机可读存储介质可以是保留和存储指令以供指令执行设备使用的有形设备。计算机可读存储介质可以是例如但不限于电子存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备、或前述各项的任何合适的组合。

[0095] 计算机可读存储介质的更具体示例的非详尽列表包括:便携式计算机软盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式光盘只读存储器(CD-ROM)、数字多功能磁盘(DVD)、记忆棒、软盘、和前述各项的任何合适的组合。本文中使用的计算机可读存储介质不应当被解释为暂态信号本身,诸如无线电波或其他自由传播的电磁波、通过波导或其他传输介质传播的电磁波(例如,通过光缆的光脉冲)、或通过电线传输的电信号。

[0096] 本文中描述的计算机可读程序指令可以从计算机可读存储介质下载到相应计算/处理设备,或者经由网络(例如,因特网、局域网、广域网和/或无线网络)下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光传输光纤、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或网络接口从网络接收计算机可读程序指令并且转发计算机可读程序指令以存储在相应计算/处理设备内的计算机可读存储介质中。

[0097] 用于执行本发明的操作的计算机可读程序指令可以是以一种或多种编程语言的任何组合编写的汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或源代码或对象代码,上述编程语言包括诸如Smalltalk、C++等面向对象的编程语言以及诸如“C”编程语言或类似编程语言等传统过程编程语言。

[0098] 计算机可读程序指令可以完全在用户的计算机上,部分在用户的计算机上,作为独立软件包,部分在用户的计算机上并且部分在远程计算机上,或者完全在远程计算机或服务器上执行。在后一种情况下,远程计算机可以通过任何类型的网络(包括局域网(LAN)或广域网(WAN))连接到用户的计算机,或者可以与外部计算机建立连接(例如,使用互联网服务提供商通过互联网进行连接)。在一些实施例中,包括例如可编程逻辑电路系统、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)在内的电子电路系统可以通过利用计算机可读程序指令的状态信息个性化电子电路系统来执行计算机可读程序指令,以便执行本发明的各方面。

[0099] 本文中参考根据本发明的实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图而描述了本发明的各方面。应当理解,流程图和/或框图的每个框以及流程图和/或框图中的框的组合可以由计算机可读程序指令来实现。

[0100] 这些计算机可读程序指令可以被提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器以产生机器,使得经由计算机或其他可编程数据处理装置的处理器执行的该指令能够创建用于实现流程图和/或框图框中指定的功能/动作的装置。这些计算机可读程序指令还可以存储在计算机可读存储介质中,该计算机可读存储介质可以指导计算机、可编程数据处理装置和/或其他设备以特定方式起作用,使得其中存储有指令的计算机可读存储介质包括制品,该制品包括实现在流程图和/或框图中指定的功能/动作的各方面的指令。

[0101] 计算机可读程序指令也可以被加载到计算机、其他可编程数据处理装置或其他设备上,以引起一系列操作步骤在计算机、其他可编程装置或其他设备上被执行以产生计算机实现的过程,使得在计算机、其他可编程装置或其他设备上执行的指令实现流程图和/或框图中指定的功能/动作。

[0102] 附图中的流程图和框图示出了根据本发明的各种实施例的系统、方法和计算机程序产品的可能实现的架构、功能和操作。在这点上,流程图或框图中的每个框可以表示指令的模块、分段或部分,其包括用于实现(多个)指定逻辑功能的一个或多个可执行指令。

[0103] 在一些替代实现中,框中标注的功能可以不按照图中标注的顺序出现。例如,根据所涉及的功能,连续示出的两个框实际上可以基本同时执行,或者这些框有时可以以相反的顺序执行。还应当注意,框图和/或流程图说明的每个框、以及框图和/或流程图说明中的框组合可以由执行指定功能或动作的基于专用硬件的系统来实现,或者由专用硬件和计算

机指令的组合来实现。

[0104] 本公开的一些实施例的目的是使用执行自然语言处理的计算机系统自动标识从医学概念的综合数据库中选择的由医生在检查患者之后编写的报告中提到的医学概念。在一些实施例中,相比于出于该目的而使用现有方法的计算机,计算机在标识报告中提到的概念的更高百分比以及不会错误地标识报告中未找到的很多医学概念方面具有更好的性能。通过这两种措施同时实现改进的性能是具有挑战性的,因为医生通常会遗漏他们认为阅读报告的另一位医生会理解的概念的字。找到这些概念需要猜测被遗漏的字,但猜测字也可能导致错误地标识出在报告中没有真正找到的更多概念。成功实现这些目标的一个有用的衡量准则是参数F1,参数F1是已标识概念的存在百分比与已标识概念的实际存在百分比的调和平均值。

[0105] 本公开的一些实施例的另一目的是实现这样的改进性能,例如由F1测量,同时足够快地处理医学报告,使得该方法可以作为用于放射学报告的更大的自然语言处理系统的第一阶段。例如,在一些实施例中,使用频率为3.40GHz并且具有4个内核和8个逻辑处理器的Intel®Core™ i7-6700CPU,计算机每秒可以处理至少100份报告。

[0106] 本公开的一些实施例的另一目的是在算法不具有对每个概念的类似AI的深刻理解的情况下实现这种改进性能。提供对每个概念的如此深入的类似AI的理解可能会使算法的开发变得非常费力。

[0107] 这些目标仅作为说明性示例给出,并且这些目标可以是本发明的一个或多个实施例的示例。由本领域技术人员固有地实现的其他期望的目标和优点可以发生或变得很清楚。本发明由所附权利要求限定。

[0108] 计算机系统的有用特征

[0109] 本发明的一些实施例的一个方面涉及一种计算机系统,该计算机系统通过在计算机上执行指令来自动确定在医学报告(例如,放射学报告中)找到医学概念集合中的哪个医学概念,发明人找到的三个特征中的一个或多个通常可以有效减少误报,并且不会遗漏太多真正存在的概念。这些特征可以通过使用计算机自动标识报告中的概念的两种方法来解释。如果在输入文本中找到概念的所有字或几乎所有字(无论字顺序如何),则以ElasticSearch中的匹配测试为例的词袋方法认为该概念与输入文本匹配,例如医学报告的一个句子。这种方法在查找存在的概念方面往往相当有效,但往往会产生大量误报,查找并不真正存在的概念。如果在输入文本中找到与在概念中相同或几乎相同顺序的概念的所有字,则由cTAKES算法使用并且以ElasticSearch中的match_phrase测试为例的另一种方法认为该概念与输入文本匹配。这种方法在消除误报方面往往比词袋方法要好一些,但在寻找存在的概念方面则差一些。发明人已经发现,在本发明的一些实施例中,采用以下特征中的任何一个或两个、或者更好的是以下所有三个特征的方法可以在消除误报和找到真正存在的概念方面给出更好的结果:

[0110] 1) 即使在字顺序很重要的match_phrase测试中也不需要所有字都存在。当撰写报告的医生从概念中省略了他们认为对报告的人类读者来说很清楚的字时,这允许找到其他概念。

[0111] 2) 要求存在某些强制字,无论是在匹配测试还是match_phrase测试中,即使所有字都不必存在。这减少了由于猜测通常不会被撰写报告的医生遗漏的重要字而导致的误报

数目。

[0112] 3) 从匹配测试或match_phrase测试中计算得分,或者更好地从两个测试的组合中计算得分,并且当存在得分更高的另一概念与得分较低的概念显著重叠时,使用得分来消除得分较低的概念。这可以通过在仅存在较窄概念时排除相关较宽概念并且在存在没有遗漏字或遗漏字较少的相似概念时排除具有遗漏字的概念来减少误报。

[0113] 发明人已经发现,在某些情况下,与单独的匹配测试或匹配短语测试相比,包括所有这三个特征的算法产生更少的误报,并且找到更大百分比的真正存在的概念。

[0114] 不需要match_phrase测试中存在概念的所有字的第一特征可选地如下实现。每个概念是字符串。对于两个或更多个字的每个概念,生成一个或多个查询,每个查询由概念的字的一个字的子串组成,可选地与在概念中的顺序相同。查询不需要是概念的字的一个适当的子串,但在2字概念的情况下,查询可以与概念相同。如果至少最小数目的查询在输入文本中具有它们的两个字,并且足够靠近,则该概念可选地被认为是输入文本的可能匹配。可选地,最小数目取决于该概念的查询总数,例如,最小数目是查询总数的百分比,例如35%或者40%、45%、50%、55%、60%或65%的查询。可选地,还可以存在必须满足的其他准则,以便将概念视为输入文本的可能匹配,例如要求存在的强制字。如果该概念被认为是输入文本的可能匹配,则为每个查询找到查询子得分,如果在输入文本中找到查询的更接近的两个字,并且如果这两个字在输入文本中的顺序与在概念中的顺序相同,则查询子得分更高。然后基于查询子得分(例如,查询子得分的加权平均)找到概念的整体match_phrase得分,并且如果概念与得分较高的另一概念具有足够重叠,则认为该概念不涉及输入文本。可选地,加权平均向两个字在概念上相距更远的查询给予较低权重。

[0115] 要求输入文本中存在某些强制字的第二特征可选地如下实现。可选地,在概念集合中的所有概念的字中作为概念中最不常见的字的的概念中的字是必须在输入文本中找到的强制字。可选地,作为疾病名称的任何单数字都是必须在输入文本中找到的强制字。可选地,指代身体部位的方向或位置的任何字必须在输入文本中与其指代的身体部位的字相邻,以便将该概念视为输入文本的可能匹配。

[0116] 可选地,除了通过查看在输入文本中找到哪些查询以及它们的查询子得分是多少来找到哪些概念是输入文本的可能匹配,还测试概念以根据词袋方法查看它们是否是输入文本的可能匹配,这称为匹配测试。在匹配测试中,如果在输入文本中找到概念中的足够数目的字,则该概念被认为是可能匹配,其中最小字数可以取决于概念中的字数。如果概念根据这个准则被认为是输入文本的可能匹配,则可选地找到该概念的匹配得分,可选地仅取决于输入文本中存在的概念的字数,而不管输入文本中字的顺序。

[0117] 可选地,将匹配得分与上述match_phrase得分结合,例如求出匹配得分与match_phrase得分的加权和,以求出该概念的总分,并且该总分用于决定当两个概念具有足够的重叠时,要保留哪个概念。可选地,仅在根据具有足够多的查询与输入文本匹配的准则(“match_phrase测试”)并且根据在输入文本中存在概念的足够多的字的准则(“匹配测试”)而认为概念是可能匹配时,才认为概念是输入文本的可能匹配。要求概念同时通过匹配测试和match_phrase测试可能会使概念更可能涉及输入文本,即使在输入文本中找不到该概念的某些字。替代地,仅进行匹配测试,或仅进行match_phrase测试,或进行这两个测试但概念只需通过测试之一即可以被认为是输入文本的可能匹配。

[0118] 可选地,匹配测试也应用于对其不应用match_phrase测试并且不计算match_phrase得分的单字概念,因为在单字概念中没有两字查询。如果在输入文本中找到一个字,并且可选地其匹配得分和总分始终相同,则单字概念通过匹配测试。可选地,对于单字概念,match_phrase得分为零。可选地,将单字概念与与其重叠的其他多字概念(例如,与也具有该字并且通过匹配测试和match_phrase测试的其他概念)的得分进行比较。可选地,在这种情况下,多字概念的得分总是较高,并且单字概念被排除在外。

[0119] 可选地,通过对字进行词干提取(即,将字替换为词干)来对概念集合进行预处理。可选地,在预处理中从概念中去除停用词。可选地,当若干医学术语具有相同含义时,概念中的字被替换为优选同义词。例如,诸如“肝脏的”等形容词被替换为对应名词(在这种情况下是“肝脏”)。可选地,扩展概念中的首字母缩略词。可选地,由字母数字名称所指代的椎骨添加了字“椎骨”和“脊椎”,并且该字母被替换为其所指代的脊椎区域(“颈椎”、“胸椎”或“腰椎”)。另外地或替代地,本文中针对概念描述的任何或所有预处理过程被应用于输入文本。

[0120] 应当注意,本文中描述和要求保护的方法并非针对抽象思想,而是针对改进用于自然语言处理的计算机技术的方法,并且权利要求并未给出标识医学报告中的医学概念的所有方法,而是仅限于实现该结果的特定方法。例如,至少一些权利要求可以仅限于从概念创建两字查询的方法,并且通过查看有多少两字查询是输入文本的匹配来对每个概念执行match_phrase测试。至少一些权利要求可以限于输入文本中必须存在概念的某些强制字的方法,以便将概念视为输入文本的可能匹配。被认为比现有技术具有新颖性和创造性的这些方法的这些元素特别适用于使用计算机技术来标识在医学报告中找到的医学概念,并且与使用人脑的自然语言处理能力的人类将用于该方法的方法有很大不同。

[0121] 对系统的描述

[0122] 现在参考附图,图1A示出了实现根据本发明的示例性实施例的方法的系统的框图100。图100中的每个框代表一个软件模块,如存储在被编程以实现该模块的计算机中,或者存储在存储或提供对数据的访问的服务器中,例如以数字形式。报告数据库102存储或提供对系统将处理的医学报告的访问。拆分器模块104将每个报告分成多个输入文本106,每个输入文本被单独处理以找到它包含的概念。通常,报告的每个句子是单独的输入文本。可选地,长句、尤其是复合句被分成多个输入文本,例如由解析器使用语法规则来尝试确定复合句的每个组成的边界。在发明人进行的测试中,长句被Python中的Natural Language Tool Kit提供的拆分器分割,但是替代地,也可以使用本领域已知的任何其他句子拆分器。可选地,句子在单独的逗号处拆分,例如,在15个字符内没有其他逗号的逗号,但是在15个字符内具有其他逗号的逗号处不拆分,因为这些逗号可能是桥接单个概念的列表的一部分,诸如在“肺癌、肝癌或喉癌”中。

[0123] 概念数据库108具有医学概念的列表,例如在SNOMED数据库中列出的医学概念、或者在SNOMED数据库中列出的概念的选择。可选地,概念数据库108并非包括在SNOMED或类似数据库中列出的所有概念,而是限于描述疾病和形态异常的概念,并且限于不超过最大字数的概念。例如,发明人通常使用概念数据库,该概念数据块仅限于在预处理之后具有6个字或更少字的概念,因为该方法往往对字较多的概念具有较低成功率,并且这样的概念在医学报告中较少被找到。在概念数据库的当前实现中,仅包括描述疾病和形态异常的概念。

共有107453个概念,在预处理之后包括4606个单字概念、24399个两字概念、25838个三字概念、22815个四字概念、18329个五字概念、11466个六字概念。

[0124] 能够访问输入文本106和医学概念列表108两者的处理器模块110使用算法来标识在每个输入文本中找到的医学概念,如将在图2中描述的,虽然一般来说这个过程会产生一些错误。该过程的结果(即,在每个输入文本中找到的医学概念列表)被可选地发送给输出模块112,输出模块112以用户可以阅读的形式输出它们。可选地,输出模块在输出中不单独列出在每个输入文本中找到的医学概念,而只是集中列出在输入文本集合中(例如,在整个医学报告中)找到的医学概念。

[0125] 方法的概述

[0126] 通过图2的流程图200更详细地描述图1中的处理器模块110的功能。图1B的流程图114中示出了流程图200的概述,在高层示出了算法的不同部分。

[0127] 在图1B的116处,概念和输入文本的预处理。概念的预处理在处理很多输入文本之前只需要进行一次,并且可选地仅在概念集合被修改时再进行一次。输入文本的预处理可选地针对正在处理的每个新输入文本进行,但可以在处理任何输入文本之前针对输入文本集合进行,例如针对医学报告的所有输入文本进行。概念和输入文本的预处理包括例如词干提取、去除停用词、扩展首字母缩略词和将字替换为同义词。概念的预处理还可以包括例如为两个或更多个字的概念生成查询,为每个概念找到强制字,以及去除某些概念,例如具有超过最大字数的概念。

[0128] 在118处,考虑第一输入文本。

[0129] 在120处,对于所考虑的输入文本,执行概念的循环。循环120、以及图1B所示的每个其他循环可以串行或并行执行,或者使用串行和并行处理的组合来执行。对于图2所示的详细流程图,这些循环都示出为正在串行处理。对于每个概念,循环120包括:1) 在122处检查输入文本中的强制字;2) 在124处执行匹配测试并且找到匹配得分;以及3) 在126处执行match_phrase测试并且找到match_phrase得分。这三个测试可以串行或并行执行,并且如果串行执行,它们可以按照任意顺序执行。对于图2所示的详细流程图,这些测试示出为连续执行,首先检查强制字,然后是匹配测试,然后是match_phrase测试。按照这个顺序执行测试有一个潜在好处,即,检查强制字通常比匹配测试快,匹配测试通常比match_phrase测试快,并且如果概念没有通过较快测试之一,则可以跳过较慢测试,以节省计算机时间。执行match_phrase测试并且找到match_phrase得分包括该概念的查询的循环128。在循环120结束时,在130处,例如通过组合匹配得分和match_phrase得分来找到该概念的总分。可选地,为了被认为是输入文本的可能匹配,一个概念必须通过所有三个测试:122处的强制字检查、124处的匹配测试和126处的match_phrase测试。这些测试可以是按照任何顺序或并行进行,但如果测试是连续进行的,则如果一个概念未能通过测试之一,则循环可选地针对该概念结束并且继续考虑下一概念。对于图2所示的详细流程图,强制字检查被示出为首先进行,然后是match_test,match_phrase测试被示出为最后进行。记录哪些概念通过了所有测试并且因此是该输入文本的可能匹配、以及作为可能匹配的每个概念的总分。

[0130] 在循环120找到作为该输入文本的可能匹配的所有概念之后,在132处进行后处理。后处理包括可能的匹配概念的循环134。在循环134中,如果可能的匹配概念与具有更高得分的另一可能的匹配概念显著重叠,例如因为在输入文本中找到另一概念的更多字,和/

或输入文本中字的顺序更接近于另一概念中字的顺序,则从考虑中排除可能的匹配概念。

[0131] 在对所有可能的匹配概念完成后处理132之后,在136处输出尚未消除的剩余匹配概念。这些是算法已经标识为在输入文本中找到的概念。可选地,不是单独列出在每个输入文本中标识的概念,而是在若干输入文本中(例如,在报告中的所有输入文本中)标识的所有概念的列表在这些输入文本被处理之后被提供作为输出。

[0132] 在138处,如果有任何剩余输入文本未被处理,则考虑下一输入文本。

[0133] 输入文本和概念的预处理

[0134] 在202处,对输入文本进行预处理。预处理可选地包括以下中的一项或多项:

[0135] 1) 去除停用词,包括冠词、连词、介词、和/或不太可能提供有关是否在输入文本中找到给定概念的信息的其他常用词。

[0136] 2) 去除具有多于一个含义可能导致错误的歧义词。例如,“exposure”这个字有时可能是指患者暴露在寒冷中,这是一个医学概念,但几乎所有放射学报告中也使用它来指代图像的曝光参数。发明人已经发现,将其省略会导致算法的F1得分更好。在算法的当前实现中没有删除但有时会导致错误的歧义词的另一示例是“sinus”,它既可以指代身体部位,也可以指代某种情况。

[0137] 3) 将字替换为词干,这可以称为“词干提取”。

[0138] 4) 将具有相同含义的字集合中的所有字替换为单个优选形式。例如,如“肝脏的”等形容词被替换为对应名词“肝脏”。

[0139] 5) 将首字母缩略词扩展为它们所代表的字。软件可选地包括在医学报告中使用的首字母缩略词列表及其扩展形式。例如,这样的列表可以手动创建。不时地,可选地,修改首字母缩略词列表,例如通过添加已经开始出现在医学报告中的新首字母缩略词,或者通过去除已经被发现有歧义和/或不再常用的首字母缩略词。

[0140] 6) 对于椎骨的字母数字名称,添加字“椎骨”和“脊柱”,并且将字母C替换为“颈椎”,将字母T替换为“胸椎”,将字母L替换为“腰椎”。

[0141] 在204处,对概念进行预处理,可选地使用与针对输入文本描述的相同的过程。可选地,在从概念列表中去除预处理之后具有超过最大字数的概念。在该算法的当前实现中,排除了具有超过六个字的概念。替代地,具有超过五个字、或超过七个字、或超过八个字的概念被排除。发明人的测试表明,确定输入文本中是否存在概念越困难,该概念的字越多,并且字数非常多的概念可能会缩小范围而不是经常发生,因此在实践中考虑它们并不则重要。在206处,为每个概念生成查询。可选地,每个查询由在(预处理后的)概念中找到的一对字组成,顺序与在概念中找到的字相同,并且为每个这样的字对生成查询。ElasticSearch的“Percolator”特征提供了一种生成查询的便捷方式。应当理解,查询中字的顺序很重要,不是因为要求输入文本的字顺序相同,而是因为输入文本中字乱序的概念在字与概念中的顺序相同时通常得分较低。可选地,还生成具有两个以上字的查询,例如最多具有三个字的查询。然而,发明人发现,使用多于两个字的查询几乎没有什么收获。

[0142] 在208处,检查每个概念以找到它包含的强制字。概念的强制字是必须出现在输入文本中的字,以便将该概念认为是输入文本的匹配。强制字可选地包括以下中的任何一项:

[0143] 1) 在该概念的所有字中,该字是概念集合中的所有概念中最不常见的字。

[0144] 2) 表示疾病的任何字。

[0145] 3) 指示身体的方向或一侧的任何字,该字描述身体部位,例如“左肺”概念中的字“左”。必须在输入文本中它所描述的身体部位的一个字内找到指示方向的字,以便将概念视为输入文本的匹配。

[0146] 此外,可选地为某些概念手动指定其他强制字,例如,如果发现它们有用。

[0147] 应当理解,过程204、206和208可以在处理任何医学报告之前进行一次,并且它们成为正在使用的算法版本的一部分。无需为所处理的每个新报告再次执行这些操作。当算法的新版本被建立时,概念列表可能会有一些变化,例如如果要从概念中去除的字有不同选择,或者使用扩展数据库来产生概念列表,并且在这种情况下,过程204、206和208可以被再次执行,至少对于新的或已经改变的概念。

[0148] 输入文本的循环和概念的循环

[0149] 在210处,考虑第一输入文本。在212处,考虑第一概念,以决定它是否是该输入文本的匹配。在214处,确定在输入文本中以及在可能具有所要求的位置的任何强制字(诸如描述身体部位的方向词)的情况下在所要求的位置处是否找到该概念的强制字。

[0150] 如果存在所有强制字,则在216处确定是否通过匹配测试。匹配测试要求输入文本中存在概念中所有字中的某个最小数目的字,但它们的顺序无关紧要,因为匹配测试是“词袋”测试。应理解,输入文本中存在的字数和概念中的字总数可选地均指代在概念和输入文本被预处理之后,例如去除停用词和词干搜索。当输入文本中必须存在以通过匹配测试的概念中的最小字数大于该概念中所有字的 $\frac{2}{3}$ 但不大于该概念中所有字的 $\frac{3}{4}$ 时,发明人找到最佳结果。但是通过将最小值设置为大于概念中所有字的 $\frac{1}{2}$ 并且不大于 $\frac{5}{6}$ 的任何值,已经获取了相当好的结果。例如,对于两字概念,可选地必须在输入文本中找到这两个字。对于三字概念,可选地必须找到两个字,或者替代地,必须在输入文本中找到所有三个字。对于四字概念,必须在输入文本中找到的最少字数可选地是两个、三个或四个。对于五字概念,最少字数可选地是三个或四个。对于六字概念,最少字数可选地是三个、四个或五个。可选地,最小字数是概念中字总数的50%、60%、70%或80%。

[0151] 如果概念通过匹配测试,则在218处找到并且记录匹配得分。匹配得分取决于在输入文本中找到的概念中的字数。如果在输入文本中找到概念中的更多字,则通常更有可能在输入文本中找到该概念,即使在输入文本中找到的概念的字的百分比是不更大。可选地,匹配得分仅取决于在输入文本中找到的概念的字数,而不考虑概念中有多少字。替代地,在计算匹配得分时也考虑概念中的字总数。在该方法的示例性实现中,匹配得分是在输入文本中找到的概念的字数的平方根。

[0152] 查询match_phrase测试的循环

[0153] 在220处,在将考虑该概念的所有查询的循环开始时考虑对该概念的第一查询,以确定该概念是否通过match_phrase测试并且计算match_phrase得分。应当理解,在单字概念的情况下,没有查询,因为每个查询都包含来自概念的两个字,所以无法进行match_phrase测试,也无法计算match_phrase得分。在这种情况下,虽然这在流程图200中没有明确示出,但是控制进行到238,在238处查看下一概念。

[0154] 在222处,确定是否在输入文本中找到所考虑的查询。也就是说,确定是否在输入文本中找到查询的两个字,不一定与它们在查询中找到的顺序相同(即,与概念中的顺序相同)。如果使用超过两个字的查询,则确定是否在输入文本中找到查询的所有字。但是,如果

这些字在输入文本中的位置与它们在查询中的位置相差太大,则认为在输入文本中找不到查询。

[0155] 字的位置有多不同可选地通过位置的一步变化次数来定义,以使查询中字的位置与输入文本中这些字的位置相匹配。例如,如果查询有两个字,并且输入文本中的两个字彼此相邻,顺序与查询中的顺序相同,则位置差异 s 为零。如果两个字在输入文本中彼此相邻,但在查询中字的顺序相反,则 $s=2$,因为查询中的每个字都必须移动一个位置才能处于与它们在输入文本中的位置相同的位置。如果输入文本中的字与查询中的字顺序相同,但它们之间还有 n 个其他字,则 $s=n$ 。并且,如果输入文本中的字顺序相反,并且它们之间有 n 个其他字,则 $s=n+2$ 。

[0156] 为了考虑要在输入文本中找到的查询,指定允许的最大 s 的参数称为“slop”。slop值可选地取决于查询的字在概念中相距多远,当查询的字在概念中相距较远时,slop值更高。例如,如果 d 是查询中两个字在概念中的距离,对于在概念中相邻的字, $d=0$,则“slop”可选地等于 d 加上一个常数,例如 $d+15$ 。发明人已经发现,当“slop”大约在 $d+10$ 到 $d+25$ 之间时,算法的性能最好。实际上,医学报告中的大多数句子都少于15个字,并且对于那些输入文本,如果位置差异 s 定义为 $d+15$,则位置差异 s 永远不会超过“slop”,并且如果在输入文本中找到查询的字,则所有查询都被认为在输入文本中找到。但是,医学报告中有一些异常长的句子,其中经常会找到若干概念,并且使用在 $d+10$ 到 $d+25$ 之间的“slop”值可以防止算法从一个概念中取一个字并且从另一概念中取一个字,并且不恰当地将它们与单个查询匹配。正是由于这些相对较少的异常长的句子,算法的性能取决于为“slop”而选择的值。

[0157] 在224处,如果在输入文本中找到查询,则计算和记录查询子得分。查询子得分可选地基于查询的字在输入得分中的位置与其在查询中的位置的不同程度,如果字的位置不同较大,则查询子得分较低。可选地,位置差异 s 的定义如上所述。例如,查询子得分与 $1/(S+1)$ 成比例,其中 s 的定义如上所述。例如,查询子得分为 $\sqrt{2/(s+1)}$ 。

[0158] 在226处,记录是否在输入文本中找到查询。在228处,确定是否存在对该概念的任何另外的查询。如果还有另外的查询,则在230处检查下一查询,并且控制再次进行到222。当在该概念中没有更多查询要检查时,则控制进行到232,在232处,确定该概念是否通过match_phrase测试。如果在输入文本中找到该概念中至少最小数目的查询,则该概念通过match_phrase测试。例如,在输入文本中找到对该概念的至少最小百分比的查询,例如在输入文本中找到该概念的至少35%、或至少40%、或至少45%、或至少50%、或至少55%、或至少60%、或至少65%。

[0159] Match_phrase得分

[0160] 如果概念通过match_phrase测试,则在234处计算该概念的match_phrase得分。可选地,match_phrase得分是在输入文本中找到的该概念的查询的所有查询子得分的加权和再除以该概念的查询数 N 。可选地,较低权重用于其字在概念中相距更远的查询。例如,如果查询由概念中的第 i 和第 j 字组成,则用于该查询的权重为 $1.2-0.15(i-j)$ 。替代地,该表达式中的常数1.2被替换为另一常数,例如在0.9到1.5之间,和/或常数0.15被替换为在0.10到0.20之间的另一常数。应当注意,如果概念只被包括在概念集合中,如果它们有6个或更少的字,则 $i-j$ 总是在1到5之间,并且权重 $1.2-0.15(i-j)$ 可以在0.45到1.05之间。可选地,选择权重的表达式,以便在给定概念中的最大字数的情况下,权重永远不会为负。

[0161] 概念的总分

[0162] 如果概念通过match_phrase测试,则在236处将该概念记录为输入文本的可能匹配,并且找到和记录该概念的总分。概念的总分取决于匹配得分和match_phrase得分,例如总分是匹配得分和match_phrase得分的加权和。例如,匹配得分的权重为match_phrase得分的权重的1.5倍,匹配得分定义为在输入文本中找到的概念中的字数的平方根,match_phrase得分定义为

$$[0163] \quad \frac{1}{N} \sum_{i \leq j} [1.2 - 0.15(j - 1)] MP(i, j)$$

[0164] 其中MP(i, j)是由概念中的第i字和第j字组成的查询的查询子得分,定义为 $\sqrt{2/(S+1)}$,match_phrase得分的表达式中的总和超过该概念中的所有查询,N是该概念中的查询数。应当理解,如果匹配得分重新定义为相差因子x,match_phrase得分定义为相差因子y,并且匹配得分和match_phrase得分的相对权重重新定义为相差因子y/x,则总分的定义将不变。还应当理解,如果总分是匹配得分和match_phrase得分的加权和,则只有匹配得分和match_phrase的相对权重才重要,因为得分只是用来比较给定输入文本的不同匹配概念的得分,并且如果所有得分都以相同因子改变,则没有区别。

[0165] 一旦该概念被记录为输入文本的可能匹配,并且其总分被记录,控制就进行到238,在238处,确定对于该输入文本是否还有要考虑的更多概念。如果在214处在输入文本中在所需要的位置没有找到该概念的所有强制字,或者如果在216处该概念未通过匹配测试,或者如果在232处该概念未通过match_phrase测试,则该概念不被记录为该输入文本的可能匹配项,并且控制直接进行到238。

[0166] 单字概念

[0167] 如果概念只有一个字,则没有与之相关联的查询,因为每个查询都有来自概念的两个字。在这种情况下,无法如上所述执行match_phrase测试或计算match_phrase得分。可选地,在这种情况下,只要匹配测试通过,概念就被认为是匹配的,这表示在输入文本中找到该概念的一个字。可选地,对于单字概念,总分仅取决于匹配得分,如果匹配得分仅取决于在输入文本中找到的概念的字数,则总分对于单字概念始终相同。如果匹配得分被定义为在输入文本中找到的概念的字数的平方根,则对于单字概念,匹配得分将始终为1。如果得分一般定义为匹配得分和match_phrase得分的加权和,匹配得分的权重为1.5,并且如果相同的定义用于单字概念,match_phrase得分设置为零,则对于单字概念,如果该概念是该输入文本的匹配,则总分将始终为1.5。发明人已经发现,在某些情况下,以这种方式处理单字概念会产生算法的良好性能,从而找到更多真正存在的概念,并且减少误报。替代地,对于在输入文本中找到一个字的所有单字概念的总分使用不同常数值,例如1或2、或中间值、或小于1或大于1的值2。

[0168] 如果在238处确定在概念集合中有更多的概念要考虑,则在240处考虑下一概念,并且控制再次进行到214。当没有更多的概念要考虑时,则在242处,确定是否为该输入文本记录了一个以上的可能的匹配概念。如果有一个以上的可能的匹配概念,则在244中,按照得分对可能的匹配概念进行排序。

[0169] 可能的匹配概念的后处理

[0170] 在246处,考虑可能的匹配概念的第二高得分。在248处,将所考虑的概念的得分与比它具有更高得分的每个概念的得分进行比较。第一次这样做时,对于得分第二高的概念,只有一个得分较高的概念必须与之进行比较。如果所考虑的概念与具有更高得分的概念具有显著重叠,则从该输入文本的可能的匹配概念列表中去掉所考虑的概念。可选地,“显著重叠”被定义为以下三个条件中的任何一个为真:

[0171] 1) 在得分较高的概念中找到所考虑的概念中的所有字。

[0172] 2) 在所考虑的概念中找到得分较高的概念中的所有字。

[0173] 3) 在两个概念之一中找到但在另一概念中没有找到的字数相对于所考虑的概念中的字数足够小。例如,在一个概念中找到但在另一概念中没有找到的字数除以1.7小于最大整数,该最大整数小于或等于所考虑的概念中的字数的0.5倍。

[0174] 发明人已经发现,“显著重叠”的这种定义给出了算法的良好性能。替代地,使用不同定义,例如使用不同常数代替1.7和/或0.5,例如使用在1.3到2.1之间的常数代替1.7并且使用在0.35到0.65之间的常数代替0.5。

[0175] 在250处,确定是否有更多可能的匹配概念要考虑。如果有,则在252处考虑得分下一高的概念,并且控制返回到248。如果没有更多可能的匹配概念要考虑,则在254处在输出中列出算法的结果,即,该输入文本的剩余匹配概念。这些剩余匹配概念是算法找到的存在于输入文本中的概念。

[0176] 可选地,该概念集合包括被认为具有相同含义的两个或更多个概念的一些组,并且每组中的一个概念被指定为更优选形式,而每组中的其他概念被指定为较少优选形式。在整个算法过程中,概念被视为单独概念,直到输入文本中的概念被标识。然后,当标识出的概念在输出中被列出时,任何较少优选形式都会被替换为对应更优选形式。

[0177] 输入文本的循环的结束

[0178] 在256处,确定是否有更多的输入文本要处理。如果有,则在258处考虑下一输入文本,并且控制返回到212,以找到下一输入文本中存在的概念。如果没有更多的输入文本要处理,则算法在260处结束。

[0179] 仅使用一些特征

[0180] 如上所述,发明人找到了三个特征,这些特征改进了现有技术的结果,与现有技术相比,允许计算机系统找到更多存在的概念,或具有更少的误报,或者这两者。这些特征是: 1) 不需要概念的所有字都存在于输入文本中,即使在字顺序很重要的match_phrase测试中也是如此;

[0181] 2) 要求输入文本中必须存在概念的某些强制字; 3) 为每个可能的匹配概念计算得分,并且当两个可能的匹配概念具有明显重叠时,消除得分较低的概念。尽管发明人已经发现,当使用所有这三个特征时获取最佳结果,但也可以仅使用这些特征中的一个或两个。图3、4和5分别针对仅使用第一特征、仅使用第二特征和仅使用第三特征的情况示出了类似于图1B的流程图。在所有三种情况下,流程图以预处理116开始,采用第一输入文本118,并且执行概念120的循环,如上文针对图1B描述的。并且,在所有三种情况下,流程图都以输出该输入文本的结果136结束,然后考虑下一输入文本138,如上文针对图1B描述的。流程图的不同之处在于,概念循环中包括的内容以及它们是否包括后处理。

[0182] 在图3中,流程图300示出了标识医学报告中的医学概念的方法,其中在每个输入

文本中,在match_phrase测试中,不必为了概念被认为是输入文本的可能匹配而在输入文本中找到概念的所有字。概念循环120仅示出了match_phrase测试128,尽管可选地也可以进行匹配测试,并且可选地,检查强制字。在流程图300中没有示出得分的计算,也没有示出后处理。match_phrase测试可选地包括查询循环,每个查询由来自概念的一对词组成,并且不需要在输入文本中找到所有查询。例如,match_phrase只需要在输入文本中找到一定数目的查询,例如50%的查询,以便概念通过match_phrase测试。替代地,match_phrase测试不包括查询循环,但是,例如,match_phrase测试要求在输入文本中找到概念的一定数目的字,少于100%的字,以便通过match_phrase测试。即使只有这个特征,与概念中的所有字都必须存在于输入文本中以便通过match_phrase测试的现有技术相比,计算机系统也可以找到更多存在于输入文本中的概念。

[0183] 在图4中,流程图400示出了标识医学报告中的医学概念的方法,其中必须在输入文本中找到每个概念的某些强制字,以便将概念视为输入文本的可能匹配,即使不必在输入文本中找到概念中的所有字。概念的循环120包括关于强制字是否存在的检查122。循环120还包括测试402,例如匹配测试和/或match_phrase测试,测试402并不总是要求输入文本中存在概念的所有字以通过测试,但确实要求概念的一定数目的字存在以通过测试。在流程图400中,如同在流程图300中一样,没有示出得分的计算,并且没有示出后处理步骤。

[0184] 在图5中,流程图500示出了标识医学报告中的医学概念的方法,其中向每个可能的匹配概念给予得分,该得分指示与输入文本的匹配程度,并且当存在与较低得分概念具有显著重叠的较高得分概念时,排除较低得分概念作为可能匹配。概念的循环120示出了匹配测试124和match_phrase测试126两者,如果概念通过测试并且被找到为可能的匹配概念,则为每个测试计算得分。match_phrase测试包括查询的循环128。匹配得分和match_phrase得分在130处被组合以获取总分,例如通过取匹配得分和match_phrase得分的总和或加权。替代地,在概念循环中仅进行匹配测试或仅进行match_phrase测试,并且得分仅为匹配得分或仅为match_phrase得分。通过执行可能的匹配概念的循环134,在132处进行后处理。对于每个可能的匹配概念,如果存在与它显著重叠的更高得分的可能的匹配概念,则该概念被排除作为可能的匹配概念。如果只找到较窄概念,则该后处理可以防止在输入文本中标识较宽概念。例如,在仅提及左肺的输入文本中,将标识“左肺”概念,但不会标识较宽概念“肺”。

[0185] 例如,计算机程序产品可以包括一个或多个存储介质;磁存储介质,诸如磁盘(诸如软盘)或磁带;光存储介质,诸如光盘、光带或机器可读条形码;固态电子存储设备,诸如随机存取存储器(RAM)或只读存储器(ROM);或者用于存储具有用于控制一个或多个计算机实践根据本发明的方法的指令的计算机程序的任何其他物理设备或介质。

[0186] 如本文中使用的,术语“约”是指 $\pm 10\%$ 。

[0187] 术语“包括(comprises)”、“包括(comprising)”、“包括(includes)”、“包括(including)”、“具有(having)”及其词形变化表示“包括但不限于”。

[0188] 术语“由……组成”表示“包括并且限于”。

[0189] 术语“基本上由……组成”表示组合物、方法或结构可以包括附加成分、步骤和/或部分,但前提是附加成分、步骤和/或部分不会实质上改变所要求保护的组成、方法或结构的基本和新颖特征。

[0190] 如本文中使用的,单数形式“一个(a)”、“一个(an)”和“该(the)”包括复数参考,除非上下文另有明确指定。例如,术语“一种化合物”或“至少一种化合物”可以包括多种化合物,包括其混合物。

[0191] 在本申请中,本发明的各种实施例可以以范围格式呈现。应当理解,范围格式的描述仅仅是为了方便和简洁,而不应当理解为对本发明的范围的不灵活限制。因此,对范围的描述应当被认为具体公开了所有可能的子范围以及该范围内的个体数值。例如,对诸如从1到6等范围的描述应当被认为具体公开了诸如从1到3、从1到4、从1到5、从2到4、从2到6、从3到6等子范围、以及该范围内的个体数字,例如1、2、3、4、5和6。无论范围的广度如何,这都适用。

[0192] 无论何时在本文中指示数字范围,都表示包括在指示范围内的任何引用的数字(得分或整数)。短语“在第一指示数字和第二指示数字之间的范围(ranging)/范围(ranges between)”和“从第一指示数字到第二指示数字的范围(ranging)/范围(ranges from)”在本文中可互换使用并且意在包括第一指示数字和第二指示数字以及它们之间的所有小数和整数。

[0193] 应当理解,为了清楚起见,在单独实施例的上下文中描述的本发明的某些特征也可以在单个实施例中组合提供。相反,为了简洁起见在单个实施例的上下文中描述的本发明的各种特征也可以在本发明的任何其他描述的实施例中单独地或以任何合适的子组合或合适地提供。在各种实施例的上下文中描述的某些特征不被认为是那些实施例的基本特征,除非实施例在没有这些元素的情况下是不可操作的。

[0194] 上面描述的和下面的权利要求部分要求保护的本发明的各种实施例和方面在以下示例中得到实验支持。

[0195] 示例

[0196] 现在参考以下示例,这些示例与以上描述一起以非限制性方式说明了本发明的一些实施例。

[0197] 为了测试该算法,并且将其性能与在医学报告中找到概念的现有技术算法进行比较,从107份放射学报告中准备测试数据。每个报告的每个句子都经过人工检查以确定概念集合中的哪些概念取自SNOMED CT医学词典,仅限于不超过6个字的概念(经过如上所述的预处理,删除停用词,等等),并且仅限于描述疾病和医学异常的概念。以这种方式,确定在每个句子中真正找到了哪些概念。在107份报告的句子中共找到860个概念。

[0198] 该算法用于标识集合中的哪些概念出现在报告的每个句子中。对标识出的真实概念的总数(真阳性)、标识出的在该句子中没有真正找到的概念的总数(假阳性)、算法未能标识出的在每个句子中真正存在的概念的总数(假阴性)、以及算法未能正确标识出的在每个句子中不存在的概念总数(真阴性)进行计数。然后计算两个量。精度是真阳性与总阳性的比率。这是算法标识出的所有概念中真正存在的部分。召回率是真阳性与真阳性加上假阴性的比率。这是算法标识出的所有概念中真正存在的部分。量F1是精度和召回率的调和平均值,可以定义为 $2/(1/\text{精度}+1/\text{召回率})$ 。

[0199] 还用五种其他方法进行相同测试,其中一些在现有技术中是已知的。在所有这些测试中,当得分较低的概念之间存在显著重叠时,它们会被消除,取而代之的是得分较高的概念,并且使用相同的概念集合(SNOMED中的概念在预处理之后具有6个或更少的字)。

[0200] 1) 仅匹配测试和匹配得分 (词袋模型), 要求输入文本中存在概念中的至少75%的字。

[0201] 2) 类似于cTAKES算法的算法, 使用零“slop”的match_phrase, 要求概念中的所有字都存在, 与概念中的顺序相同, 并且中间没有其他字。

[0202] 3) 具有某个“slop”的Match_phrase, 要求概念中的所有字都存在, 不一定按顺序, 但如果概念中的字在输入文本中相距太远或太乱, 则不将概念视为匹配。“slop”的值在10到15之间, 并且结果不太取决于该范围内slop的确切值。

[0203] 4) 包括匹配得分和match_phrase得分 (具有某个“slop”) 两者的混合算法, match_phrase测试要求输入文本中存在概念中的所有字。

[0204] 5) 包括匹配得分和match_phrase得分两者的混合算法, 其中match_phrase得分和match_phrase测试基于概念中所有两字查询的集合, 而不是要求与概念中的所有字的匹配, 但没有使用强制字。

[0205] 测试结果如表1所示。使用频率为3.40GHz并且具有4个内核和8个逻辑处理器的Intel®Core™ i7-6700CPU, 并且使用全算法, 计算机系统每秒能够处理超过100份放射报告。

[0206] 表1. 不同算法的测试结果

[0207]		仅匹配	仅匹配短 语、slop 为零	具有某个 slop 的匹 配短语	混合, 匹 配短语中 slop 为零	混合, 匹配 短语中是 2 字查询	全算法
	精度	61.6%	75.7%	78.3%	86.5%	75.1%	88.3%
	召回率	80.8%	70.2%	78.7%	85.9%	86.7%	89.7%
	F1	69.9%	72.9%	78.5%	86.2%	80.4%	89.0%

[0208] 一些现有技术方法 (例如, 仅使用“匹配”) 在精度 (拒绝不存在的概念) 方面相当擅长, 但在召回率 (找到所有存在的概念) 方面不太擅长。而其他现有技术方法 (例如, 使用“match_phrase”而不是“匹配”的cTAKES) 在召回率方面相当不错, 但在精度方面却不太好。使用“match_phrase”允许一些“slop”允许算法找到几乎与仅使用“匹配”一样多的存在的概念, 同时更好地拒绝不存在的概念。使用结合匹配得分和match_phrase得分的混合算法, 并且对match_phrase得分和测试使用两字查询, 使得概念中的所有字都不需要存在, 进一步提高了召回率, 同时使用match_phrase测试中的slop为零的混合算法, 要求概念中的所有字以相同顺序出现, 进一步提高了精度。针对match_phrase得分和测试使用查询并且使用强制字的全算法具有最高精度和最高召回率, 均接近89%。对于全算法, 精度和召回率几乎相同的事实反映了这样以下事实: 当算法遗漏了存在的概念时, 它几乎总是会找到并不真正存在的另一概念, 反之亦然。

[0209] 应当注意, 对于使用全算法的这个测试数据, 大约27%的错误 (误报和漏报) 是由于算法找到了与手动找到的概念密切相关的概念, 但是概念是可以接受的, 并且算法找到的概念完全不能被认为是错误的。一个典型的示例是找到“桡骨和/或尺骨骨折”, 而不是找

到“桡骨骨折”和“尺骨骨折”两个概念。如果这些错误不计入错误,则F1会更高,大约92%。

[0210] 尽管通过对算法进行微调仍可以实现性能的微小改进,但发明人认为,如果不对每个概念进行类似AI的深入理解,则可能无法显著提高算法的性能。

[0211] 已经详细描述了本发明,并且可能已经特别参考合适的或当前优选的实施例进行了描述,但是应当理解,可以在本发明的精神和范围内实现变化和修改。因此,目前公开的实施例在所有方面都被认为是说明性的而非限制性的。本发明的范围由所附权利要求表示,在其等同含义和范围内的所有变化均旨在包含在其中。

[0212] 尽管已经结合其特定实施例描述了本发明,但很明显,对于本领域技术人员而言,很多替代、修改和变化将是很清楚的。因此,其意图包括落入所附权利要求的精神和广泛范围内的所有这样的替代、修改和变化。

[0213] 本说明书中提及的所有出版物、专利和专利申请均通过引用整体并入本说明书中,其程度就如同每个个体出版物、专利或专利申请被具体地和个体地指示为通过引用并入本文中一样。此外,本申请中对任何参考文献的引用或标识不应当被解释为承认这样的参考文献可以作为本发明的现有技术。就使用章节标题而言,它们不应当被解释为必然限制。此外,本申请的任何优先权文件通过引用整体并入本文。

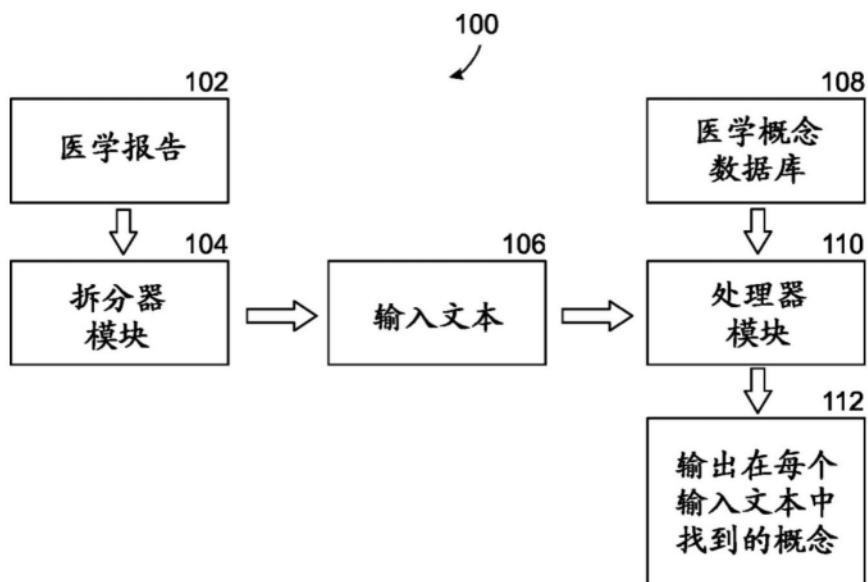


图1A

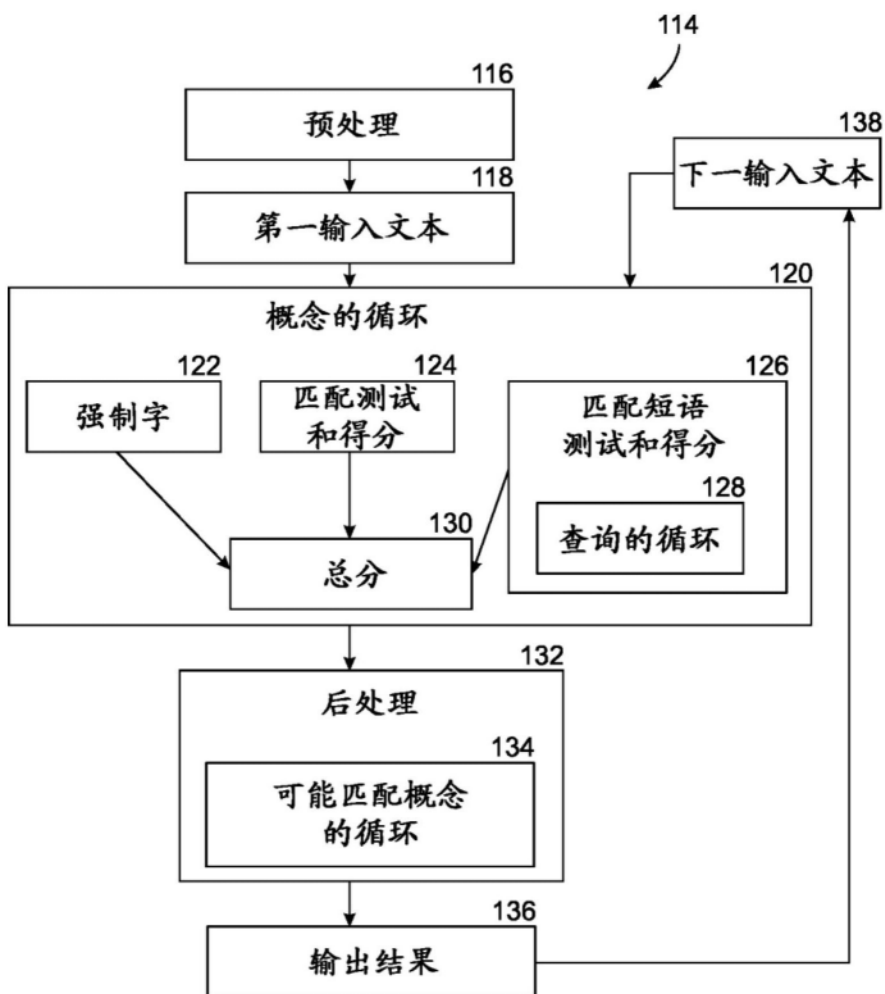


图1B

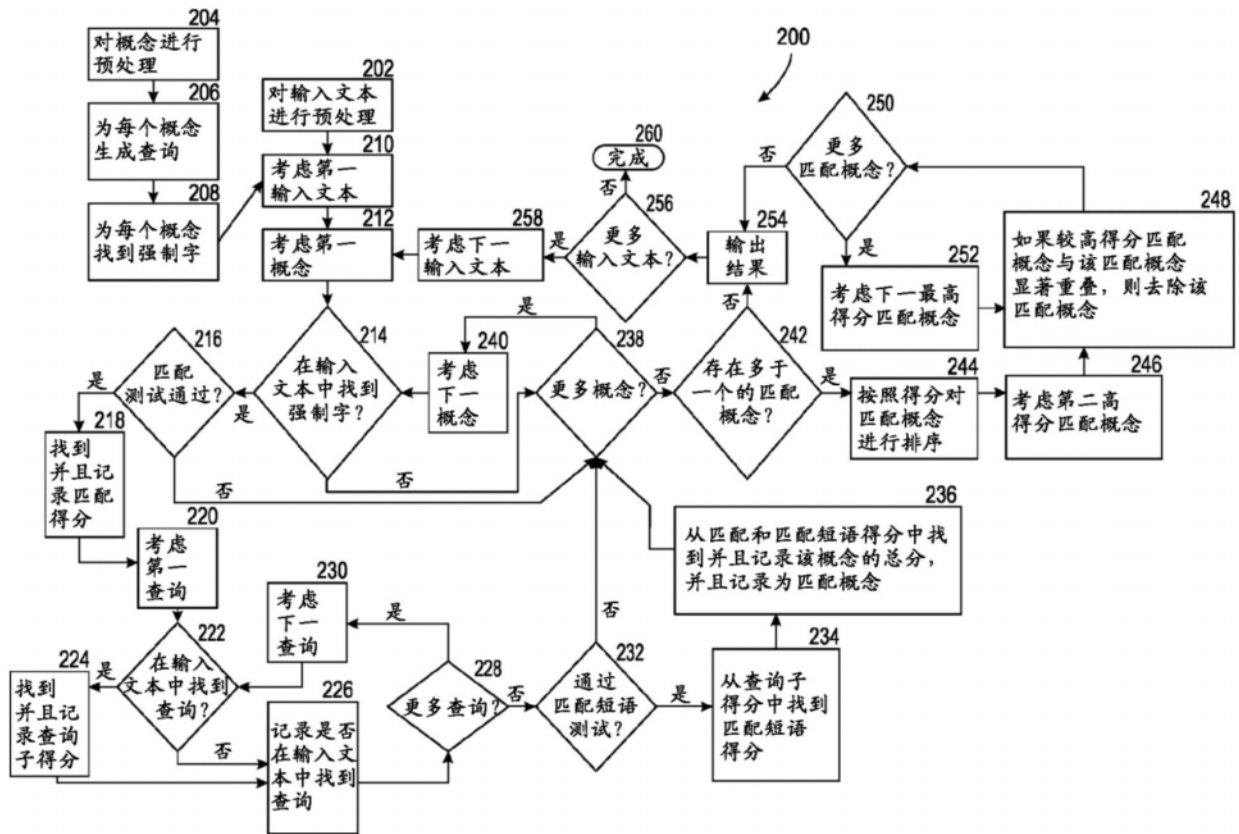


图2

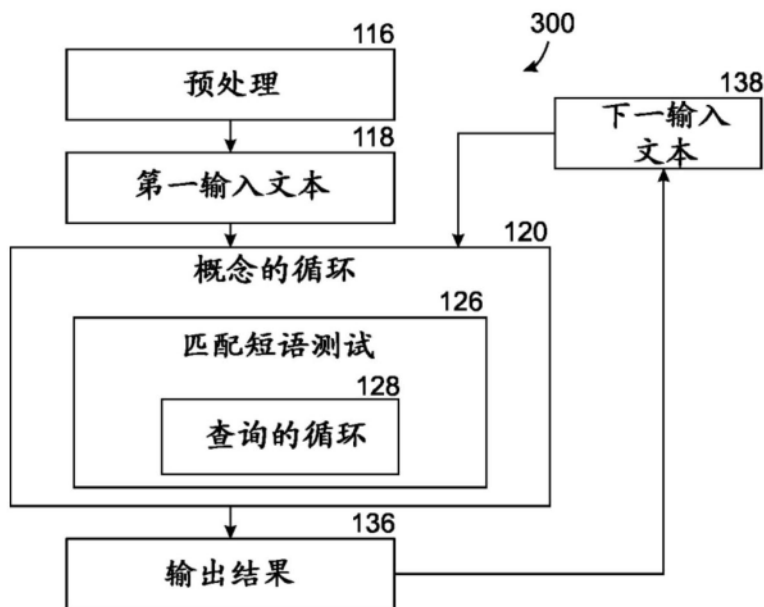


图3

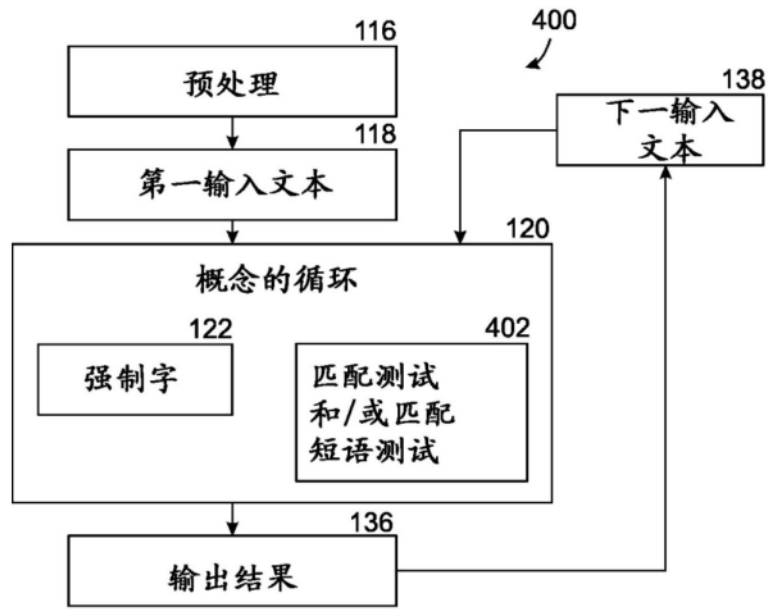


图4

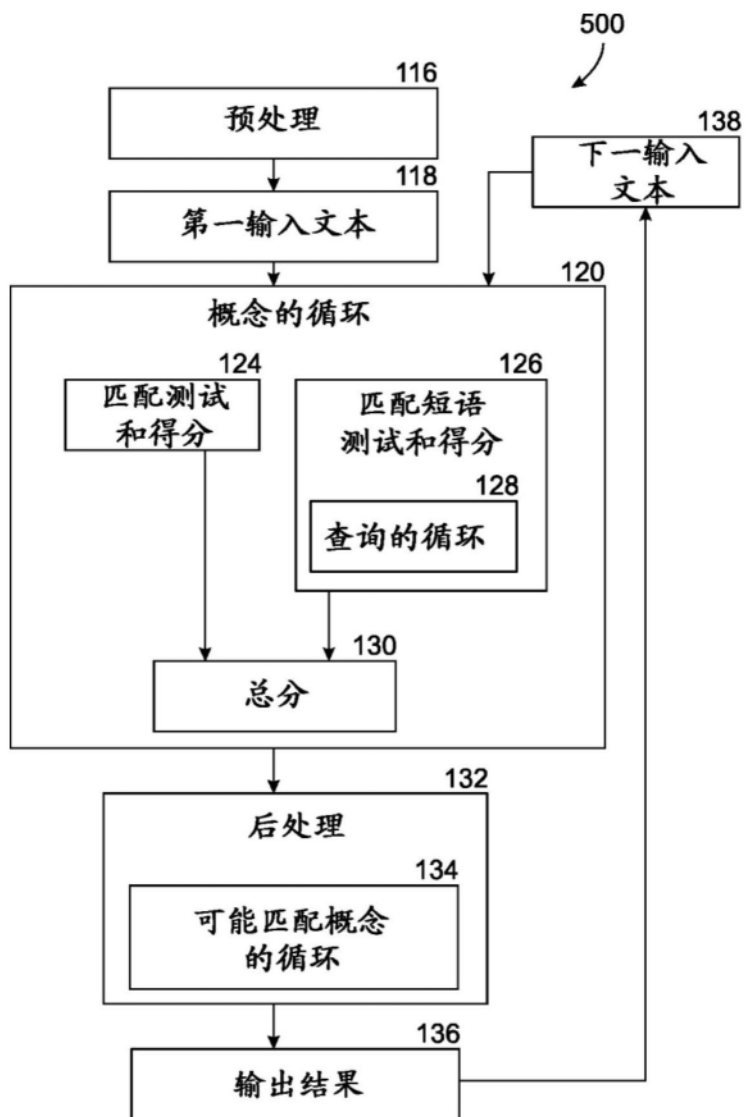


图5