(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113157946 A
(43)申请公布日 2021.07.23

(21)申请号 202110529673.6

(22)申请日 2021.05.14

(71)申请人 咪咕文化科技有限公司
地址 100032 北京市西城区华远街11号
申请人 中国移动通信集团有限公司

(72)发明人 周效军 李东晓

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
代理人 王宇杨

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/28(2019.01)
G06F 16/33(2019.01)
G06F 40/295(2020.01)

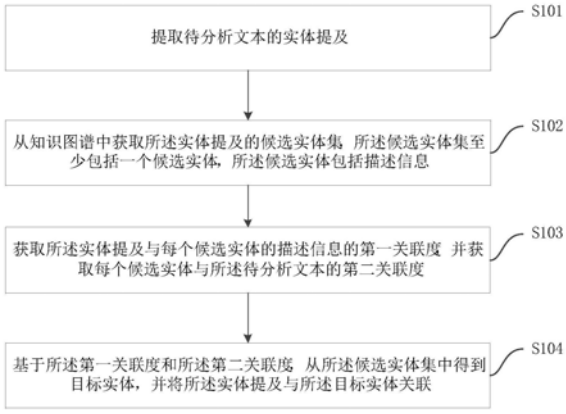
权利要求书2页 说明书9页 附图2页

(54)发明名称

实体链接方法、装置、电子设备及存储介质

(57)摘要

本发明提供一种实体链接方法、装置、电子设备及存储介质。其中，实体链接方法，包括：提取待分析文本的实体提及；从知识图谱中获取实体提及的候选实体集，候选实体集至少包括一个候选实体，候选实体包括描述信息；获取实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与待分析文本的第二关联度；基于第一关联度和第二关联度，从候选实体集中得到目标实体，并将实体提及与目标实体关联。本发明的实体链接方法，利用词汇间的共现关系，可准确地捕捉到词汇间语义的关联度，进而，保证了实体提及能够准确地连接到知识图谱中对应的实体上，提升了实体链接的准确性和可靠性，有效地扩充了知识图谱的规模。



CN 113157946 A

1.一种实体链接方法,其特征在于,包括:

提取待分析文本的实体提及;

从知识图谱中获取所述实体提及的候选实体集,所述候选实体集至少包括一个候选实体,所述候选实体包括描述信息;

获取所述实体提及与每个候选实体的描述信息的第一关联度,并获取每个候选实体与所述待分析文本的第二关联度;

基于所述第一关联度和所述第二关联度,从所述候选实体集中得到目标实体,并将所述实体提及与所述目标实体关联。

2.根据权利要求1所述的实体链接方法,其特征在于,所述获取所述实体提及与每个候选实体的描述信息的第一关联度,并获取每个候选实体与所述待分析文本的第二关联度,包括:

获取所述实体提及与每个候选实体的描述信息的共现概率,并获取每个候选实体与所述待分析文本的共现概率;

根据所述实体提及与每个候选实体的描述信息的共现概率,得到所述实体提及与每个候选实体的描述信息的第一关联度,并根据所述每个候选实体与所述待分析文本的共现概率,得到所述每个候选实体与所述待分析文本的第二关联度。

3.根据权利要求2所述的实体链接方法,其特征在于,所述获取所述实体提及与每个候选实体的描述信息的共现概率,并获取每个候选实体与所述待分析文本的共现概率,包括:

获取所述实体提及和每个候选实体的描述信息中各词汇组合得到的二元词汇组的频率,并基于所述实体提及和每个候选实体的描述信息中各词汇组合得到的二元词汇组的频率,得到所述实体提及与每个候选实体的描述信息的共现概率;

获取所述每个候选实体与所述待分析文本中各词汇组合得到的二元词汇组的频率,并基于所述每个候选实体与所述待分析文本中各词汇组合得到的二元词汇组的频率,得到所述每个候选实体与所述待分析文本的共现概率。

4.根据权利要求3所述的实体链接方法,其特征在于,所述二元词汇组的频率是基于预设的基础语料库中的文本统计得到。

5.根据权利要求1所述的实体链接方法,其特征在于,所述从所述知识图谱中获取所述实体提及的候选实体集,包括:

获取所述实体提及的别称;

基于所述实体提及和所述别称,从所述知识图谱中匹配得到所述实体提及的候选实体集。

6.根据权利要求1-5任一项所述的实体链接方法,其特征在于,所述基于所述第一关联度和所述第二关联度,从所述候选实体集中得到目标实体,包括:

获取所述第一关联度和所述第二关联度的加权平均值,并将所述加权平均值作为综合关联度量值;

基于所述综合关联度量值和所述实体提及与每个候选实体的描述信息的第一关联度,得到所述目标实体。

7.根据权利要求6所述的实体链接方法,其特征在于,所述基于所述综合关联度量值和所述实体提及与每个候选实体的描述信息的第一关联度,得到所述目标实体,包括:

获取大于所述综合关联度量值的第一关联度；

将大于所述综合关联度量值的第一关联度中的最大第一关联度对应的候选实体作为所述目标实体。

8.一种实体链接装置，其特征在于，包括：

提取模块，用于提取待分析文本的实体提及；

候选实体获取模块，用于从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息；

关联度计算模块，用于获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度；

实体链接模块，用于基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联。

9.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现根据权利要求1至7任一项所述实体链接方法的步骤。

10.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现根据权利要求1至7任一项所述实体链接方法的步骤。

实体链接方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种实体链接方法、装置、电子设备及存储介质。

背景技术

[0002] 在知识图谱构建过程中，需要进行实体链接，实体链接是将外部实体提及关联到已有知识图谱中对应的实体上。目前，实体链接技术包括根据实体提及的上下文语境与知识图谱中候选实体描述语境进行语义或者文本结构比较，计算相似度，进而判断是否链接。然而，实体提及上下文语境与知识图谱中候选实体描述语境在很多情况下不存在相似语义的词汇集，例如："孙悟空大闹天宫"一文中"孙悟空"是实体提及，在知识图谱中存在一个候选实体"猴哥"，其描述为"猴哥西天取经"此时，基于语义计算和基于文本结构的词汇比较的实体链接技术均不能够将实体提及"孙悟空"与知识图谱中的候选实体"猴哥"链接。

发明内容

[0003] 本发明提供一种实体链接方法、装置、电子设备及存储介质，可以提升了实体链接的准确性和可靠性，有效地扩充了知识图谱的规模。

[0004] 本发明提供一种实体链接方法，包括：

[0005] 提取待分析文本的实体提及；

[0006] 从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息；

[0007] 获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度；

[0008] 基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联。

[0009] 根据本发明提供的一种实体链接方法，所述获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度，包括：

[0010] 获取所述实体提及与每个候选实体的描述信息的共现概率，并获取每个候选实体与所述待分析文本的共现概率；

[0011] 根据所述实体提及与每个候选实体的描述信息的共现概率，得到所述实体提及与每个候选实体的描述信息的第一关联度，并根据所述每个候选实体与所述待分析文本的共现概率，得到所述每个候选实体与所述待分析文本的第二关联度。

[0012] 根据本发明提供的一种实体链接方法，所述获取所述实体提及与每个候选实体的描述信息的共现概率，并获取每个候选实体与所述待分析文本的共现概率，包括：

[0013] 获取所述实体提及和每个候选实体的描述信息中各词汇组合得到的二元词汇组的频率，并基于所述实体提及和每个候选实体的描述信息中各词汇组合得到的二元词汇组的频率，得到所述实体提及与每个候选实体的描述信息的共现概率；

[0014] 获取所述每个候选实体与所述待分析文本中各词汇组合得到的二元词汇组的频率，并基于所述每个候选实体与所述待分析文本中各词汇组合得到的二元词汇组的频率，得到所述每个候选实体与所述待分析文本的共现概率。

[0015] 根据本发明提供的一种实体链接方法，所述二元词汇组的频率是基于预设的基础语料库中的文本统计得到。

[0016] 根据本发明提供的一种实体链接方法，所述从所述知识图谱中获取所述实体提及的候选实体集，包括：

[0017] 获取所述实体提及的别称；

[0018] 基于所述实体提及和所述别称，从所述知识图谱中匹配得到所述实体提及的候选实体集。

[0019] 根据本发明提供的一种实体链接方法，所述基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，包括：

[0020] 获取所述第一关联度和所述第二关联度的加权平均值，并将所述加权平均值作为综合关联度量值；

[0021] 基于所述综合关联度量值和所述实体提及与每个候选实体的描述信息的第一关联度，得到所述目标实体。

[0022] 根据本发明提供的一种实体链接方法，所述基于所述综合关联度量值和所述实体提及与每个候选实体的描述信息的第一关联度，得到所述目标实体，包括：

[0023] 获取大于所述综合关联度量值的第一关联度；

[0024] 将大于所述综合关联度量值的第一关联度中的最大第一关联度对应的候选实体作为所述目标实体。

[0025] 本发明还提供一种实体链接装置，包括：

[0026] 提取模块，用于提取待分析文本的实体提及；

[0027] 候选实体获取模块，用于从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息；

[0028] 关联度计算模块，用于获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度；

[0029] 实体链接模块，用于基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联。

[0030] 本发明还提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如上述任一种所述实体链接方法的步骤。

[0031] 本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如上述任一种所述实体链接方法的步骤。

[0032] 本发明提供的实体链接方法、装置、电子设备及存储介质，首先从知识图谱中获取如与实体提及的名称相关的多个候选实体，然后，可以基于如大数据等确定出实体提及与每个候选实体的描述信息中出现的各词汇共同出现的情况，以得到第一关联度、以及每个候选实体与实体提及的待分析文本中各词汇的共同出现的情况，以得到第二关联度，进而，可以第一关联度和第二关联度的关联的密切程度准确地捕捉到实体提及与多个候选实体

之间的关联性，并根据实体提及与多个候选实体之间的关联性确定出链接的实体，进而，保证了实体提及能够准确地链接到知识图谱中对应的实体上，提升了实体链接的准确性和可靠性，有效地扩充了知识图谱的规模。

附图说明

[0033] 为了更清楚地说明本发明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一简单地介绍，显而易见地，下面描述中的附图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获得其他的附图。

[0034] 图1是本发明提供的实体链接方法的流程示意图；

[0035] 图2是本发明提供的实体链接方法的示意图；

[0036] 图3是本发明提供的实体链接装置的结构示意图；

[0037] 图4是本发明提供的电子设备的结构示意图。

具体实施方式

[0038] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0039] 以下结合附图描述本发明实施例的实体链接方法、装置、电子设备及存储介质。

[0040] 其中，实体链接是将外部的实体提及关联到已有的知识图谱中对应的实体上，这样，一方面可以扩展实体提及的语义信息，另一方面能够更准确地扩充知识图谱的规模。

[0041] 在以上描述中，实体提及是从待分析文本中提取的实体信息，如待分析文本中的主语，该实体提及作为待链接对象，即：待关联到已有的知识图谱中对应的实体上。

[0042] 图1是根据本发明一个实施例的实体链接方法的流程图。如图1所示，根据本发明实施例的实体链接方法，包括如下步骤：

[0043] S101：提取待分析文本的实体提及。

[0044] 如图2所示，可以通过实体识别的方式进行待分析文本中的实体提及的提取。例如：依据预先准备好的命名实体识别模型(简称NER模型，Named Entity Recognition模型)实现实体提及的自动提取，其中，命名实体识别模型中采用命名实体识别算法进行待分析文本中的实体提及的提取，由此，通过对待分析文本进行字符级别的预处理并得到命名实体识别模型输入数据后，命名实体识别模型会自动完成命名实体序列标注，进而，根据预测的标注序列完成待分析文本中实体提及的提取。

[0045] 待分析文本例如为“张三毕业于中央戏剧学院表演系，中国辽宁籍男演员”，则通过命名实体识别模型提取的实体提及可以包括“张三”、“中国”、“演员”等。

[0046] S102：从知识图谱中获取实体提及的候选实体集，候选实体集至少包括一个候选实体，候选实体包括描述信息。

[0047] 知识图谱中包括多个实体，候选实体集是从知识图谱中的多个实体中选择出的一个或多个实体组成的集合，其中，实体提及需要关联的实体是候选实体集中的实体，因此，

本发明的实施例中，候选实体集中的实体称为候选实体。

[0048] 如图2所示，候选实体的生成，即：从知识图谱中获取实体提及的候选实体集，包括：获取实体提及的别称；基于实体提及和别称，从知识图谱中匹配得到实体提及的候选实体集。候选实体集中的候选实体包括描述信息，描述信息是候选实体的实体属性，通常是一段文本，例如：对于一段已经链接到知识图谱中的文本“孙悟空大闹天宫的剧情很精彩”，对于该文本，在知识图谱中实体可以是“孙悟空”，其对应的描述信息为“大闹天宫的剧情很精彩”，当然，也可以将每个候选实体的实体属性单独获取，则候选实体集对应一个候选实体属性集。

[0049] 在本发明的一个实施例中，候选实体的生成可根据实体提及以及别称，通过正则表达式从知识图谱中匹配得到的名称集合。即：这些名称集合与知识图谱中的实体的名称存在对应关系，也就是说，名称及合中的名称对应知识图谱中的实体。具体过程如下：

[0050] 根据实体提及查询预存的别称库，该别称库中存在实体提及以及其别称，通过别称库可以扩展实体提及的可能的名称，由于实体提及的别称有可能与实体提及的本名使用的汉字无交集，导致不能用正则表达式的方式进行匹配查询，因此，首先获取到实体提及的别称，进而，可以根据别称和实体提及分别进行匹配查询。例如：电影实体名“天堂电影院”的别称为“星光伴我心”。则别称库中记录的是各实体对应的别称，格式为但不限于二元组形式，如：<天堂电影院，星光伴我心>。

[0051] 在得到实体提及的别称后，根据实体提及和别称，应用正则表达式或者编辑距离等方式，从知识图谱中得到候选实体集，其中，正则表达式匹配的方式包括但不限于如下形式：

[0052] 候选实体集中的候选实体的名称是包含实体提及或者其别称的实体，或者，实体提及或者别称中包含知识图谱中候选实体集中的候选实体的名称。

[0053] 编辑距离描述的是针对字符级的变化次数统计，使其中一个字符序列转变为另一个字符序列。通过计算两个实体的名称的编辑距离，并与设定好的阈值比较，进而，从知识图谱中筛选出可能的实体集作为候选实体集。

[0054] S103：获取实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与待分析文本的第二关联度。

[0055] 在本发明的一个实施例中，获取实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与待分析文本的第二关联度，包括：获取实体提及与每个候选实体的描述信息的共现概率，并获取每个候选实体与待分析文本的共现概率；根据实体提及与每个候选实体的描述信息的共现概率，得到实体提及与每个候选实体的描述信息的第一关联度，并根据每个候选实体与待分析文本的共现概率，得到每个候选实体与待分析文本的第二关联度。

[0056] 该示例中，获取实体提及与每个候选实体的描述信息的共现概率，并获取每个候选实体与待分析文本的共现概率，包括：获取实体提及和每个候选实体的描述信息中各词汇组合得到的二元词汇组的频率，并基于实体提及和每个候选实体的描述信息中各词汇组合得到的二元词汇组的频率，得到实体提及与每个候选实体的描述信息的共现概率；获取每个候选实体与待分析文本中各词汇组合得到的二元词汇组的频率，并基于每个候选实体与待分析文本中各词汇组合得到的二元词汇组的频率，得到每个候选实体与待分析文本的

共现概率。

[0057] 在上述示例中，二元词汇组的频率例如是基于预设的基础语料库中的文本统计得到。其中，基础语料库中包括多个文本，即：记录有多个文本，记录的文本可以预先收集得到，例如：通过人工进行收集、通过网络上进行收集等。

[0058] 在以上描述中，第一关联度表示的是实体提及与每个候选实体的描述信息中各词汇之间关联的密切程度，例如：关联的密切程度可以通过在大数据中共同出现的概率确定，可以是共同出现的概率越大，密切程度越大，则说明实体提及与该候选实体之间的关联性越强；同样地，第二关联度表示的是每个候选实体与待分析文本中各词汇之间关联的密切程度。

[0059] 第一关联度和第二关联度是与共现概率相关的，因此，首先需要确定共现概率，其中，共现概率是共同出现的概率。具体来说，如图2所示，共现概率生成则是基于基础语料库(即：大数据)的。也就是说，共现概率的生成主要依据基础语料库中的文本记录，统计不同二元词汇组合的频率作为共现概率，并进行存储。生成过程如下：

[0060] 首先对文本进行分词，统计每个词汇出现的次数，然后依次以每个词汇作为头词汇，统计头词汇与其他任意词汇组成的二元词汇组出现的次数，进而得到文本记录中的每个词汇对应的统计信息，例如：<词汇i，词汇j>：<词汇i出现次数，词汇i和词汇j共同出现次数>。其中，共现描述的是文本中同时出现的情况。可以先对文本进行分词和实体识别，并针对分词结果去除停用词，停用词包括但不限于：助词、副词和介词等虚词，得到去重后的词汇集和实体集。针对文本得到统计结果通常为各实体出现一次，各实体与各词汇共现一次。

[0061] 例如：针对文本S1：“张三毕业于中央戏剧学院表演系，中国辽宁籍男演员。”操作步骤分文以下步骤1和步骤2，其中，步骤1为：

[0062] 分词，则S1的分词结果为“张三”，“毕业”，“中央”，“戏剧”，“学院”，“表演系”，“中国”，“辽宁”，“籍”，“男演员”。

[0063] 实体识别，S1的实体识别结果为：张三。

[0064] 统计频次，首先针对每个实体得到：

[0065] 张三：1

[0066] 再对实体与各词汇的二元共现组合得到：

[0067] <张三，毕业>：<1，1>；

[0068] <张三，中央>：<1，1>；

[0069] <张三，戏剧>：<1，1>；

[0070] <张三，学院>：<1，1>；

[0071] <张三，表演系>：<1，1>；

[0072] <张三，中国>：<1，1>；

[0073] <张三，辽宁>：<1，1>；

[0074] <张三，籍>：<1，1>；

[0075] <张三，男演员>：<1，1>。

[0076] 步骤2：遍历基础语料库中的每一条文本记录。针对每一条文本记录重复上述的步骤1，得到统计结果后进行合并操作，合并过程中对应词汇及二元词汇组出现的次数进行相加，得到统计结果。如果不存在新的实体统计信息或者二元词汇组共现的统计信息，则新增

该信息，否则在原记录统计结果中进行相加的更新操作。

[0077] 针对统计结果中的二元词汇组合频次矩阵转化为概率形式，其中，共现概率p为对应的频率，即：

[0078] $$p(<\text{实体}，\ \text{词汇}>)=\frac{count(<\text{实体}，\ \text{词汇}>)}{count(\text{实体})}$$，

[0079] 其中，count()表示统计的次数。

[0080] 关联度的计算分为第一关联度和第二关联度的计算，机：计算实体提及与各候选实体属性集的关联度，以及计算各候选实体与实体提及的上下文(即：待分析文本中除实体提及的剩余文本)的关联度，其中，关联度r计算方法如下：

[0081] $$r=\frac{1}{n}\sum p(<\text{实体}，\ \text{词汇}>)$$，

[0082] 其中，n表示文本分词后词汇个数。最后得到第一关联度和第二关联度，再将第一关联度与第二关联度加权平均后作为各候选实体的综合关联度量值。其中，可以通过查询得到各个二元词汇组的共现概率。

[0083] 关联度的具体计算过程例如：

[0084] 实体提及：张三。待分析文本S2为：2011年张三在《新水浒传》中饰演小李广花荣。

[0085] 对S2进行分词得到：张三、新水浒传、饰演、小李广、花荣。

[0086] 根据实体提及和分词结果得到二元词汇组：<张三，新水浒传>，<张三，饰演>，<张三，小李广>，<张三，花荣>。

[0087] 根据得到的二元词汇组，查询得到共现概率p1(张三，新水浒传)，共现概率p2(张三，饰演)，共现概率p3(张三，小李广)以及共现概率p4(张三，花荣)，进而，得到关联度为：

[0088] $$r=\frac{1}{4}(p1+p2+p3+p4)$$。

[0089] 以上是计算一个实体和一段文本之间的关联度，本发明的实施例中，综合关联度R计算方法为实体提及与各候选实体属性集(通常拼接成一段文本形式)之间的关联度r1...m，和各候选实体与实体提及上下文之间的关联度R1...m的加权平均值，其中，m表示候选实体个数。即

[0090] $$R=\frac{1}{m}\left(w1*\sum_{i=1}^{m}ri+w2*\sum_{j=1}^{m}Rj\right)$$，

[0091] 其中，w1和w2为权重，w1和w2为0-1之间的数值，可以根据需要进行设定。

[0092] S104：基于第一关联度和第二关联度，从候选实体集中得到目标实体，并将实体提及与目标实体关联。

[0093] 在本发明的一个实施例中，基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，包括：获取所述第一关联度和所述第二关联度的加权平均值，并将所述加权平均值作为综合关联度量值；基于所述综合关联度量值和所述实体提及与每个候选实体的描述信息的第一关联度，得到所述目标实体。

[0094] 该示例中，基于综合关联度量值和实体提及与每个候选实体的描述信息的第一关

联度，得到所述目标实体，包括：获取大于所述综合关联度量值的第一关联度；将大于综合关联度量值的第一关联度中的最大第一关联度对应的候选实体作为目标实体。也就是说，得到实体提及对应候选实体集的综合关联度量值后，取最大的大于所述综合关联度量值的第一关联度与综合关联度量值比较，超过该综合关联度量值，表明可以链接，即：将实体提及与对应的实体进行关联，从而，准确地扩充了知识图谱的规模。

[0095] 根据本发明实施例的实体链接方法，首先从知识图谱中获取如与实体提及的名称相关的多个候选实体，然后，可以基于如大数据等确定出实体提及与每个候选实体的描述信息中出现的各词汇共同出现的情况，以得到第一关联度、以及每个候选实体与实体提及的待分析文本中各词汇的共同出现的情况，以得到第二关联度，进而，可以第一关联度和第二关联度的关联的密切程度准确地捕捉到实体提及与多个候选实体之间的关联性，并根据实体提及与多个候选实体之间的关联性确定出链接的实体，进而，保证了实体提及能够准确地链接到知识图谱中对应的实体上，提升了实体链接的准确性和可靠性，有效地扩充了知识图谱的规模。

[0096] 与现有的实体链接技术相比，如：“孙悟空大闹天宫”一文中“孙悟空”是实体提及，在知识图谱中存在一个候选实体“猴哥”，其描述为“猴哥西天取经”此时，现有技术中，基于语义计算和基于文本结构的词汇比较的实体链接技术均不能够将实体提及“孙悟空”与知识图谱中的候选实体“猴哥”链接，然而，实际上两者之间是可以链接的。而通过本发明实施例的实体链接方法，可以分析出“孙悟空”与“西天取经”之间的关联性，同样地，可以确定出“猴哥”与“大闹天宫”之间的关联性，进而，能够保证实体提及准确地链接到知识图谱中对应的实体上。

[0097] 下面对本发明提供的实体链接装置进行描述，下文描述的实体链接装置与上文描述的实体链接方法可相互对应参照。

[0098] 如图3所示，根据本发明一个实施例的实体链接装置，包括：提取模块310、候选实体获取模块320、关联度计算模块330和实体链接模块340，其中：

[0099] 提取模块310，用于提取待分析文本的实体提及；

[0100] 候选实体获取模块320，用于从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息；

[0101] 关联度计算模块330，用于获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度；

[0102] 实体链接模块340，用于基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联。

[0103] 根据本发明实施例的实体链接装置，首先从知识图谱中获取如与实体提及的名称相关的多个候选实体，然后，可以基于如大数据等确定出实体提及与每个候选实体的描述信息中出现的各词汇共同出现的情况，以得到第一关联度、以及每个候选实体与实体提及的待分析文本中各词汇的共同出现的情况，以得到第二关联度，进而，可以第一关联度和第二关联度的关联的密切程度准确地捕捉到实体提及与多个候选实体之间的关联性，并根据实体提及与多个候选实体之间的关联性确定出链接的实体，进而，保证了实体提及能够准确地链接到知识图谱中对应的实体上，提升了实体链接的准确性和可靠性，有效地扩充了知识图谱的规模。

[0104] 图4示例了一种电子设备的实体结构示意图，如图4所示，该电子设备可以包括：处理器(processor)410、通信接口(CommunicationsInterface)420、存储器(memory)430和通信总线440，其中，处理器410，通信接口420，存储器430通过通信总线440完成相互间的通信。处理器410可以调用存储器430中的逻辑指令，以执行实体链接方法，该方法包括：提取待分析文本的实体提及；从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息；获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度；基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联。

[0105] 此外，上述的存储器430中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。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或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-OnlyMemory)、随机存取存储器(RAM，RandomAccessMemory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0106] 另一方面，本发明还提供一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，计算机能够执行上述各方法所提供的实体链接方法，该方法包括：提取待分析文本的实体提及；从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息；获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度；基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联。

[0107] 又一方面，本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现以执行上述各提供的实体链接方法，该方法包括：提取待分析文本的实体提及；从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息；获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度；基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联。

[0108] 以上所描述的装置实施例仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例的方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0109] 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各实施方式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。基于这样的理解，上

述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，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0110] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

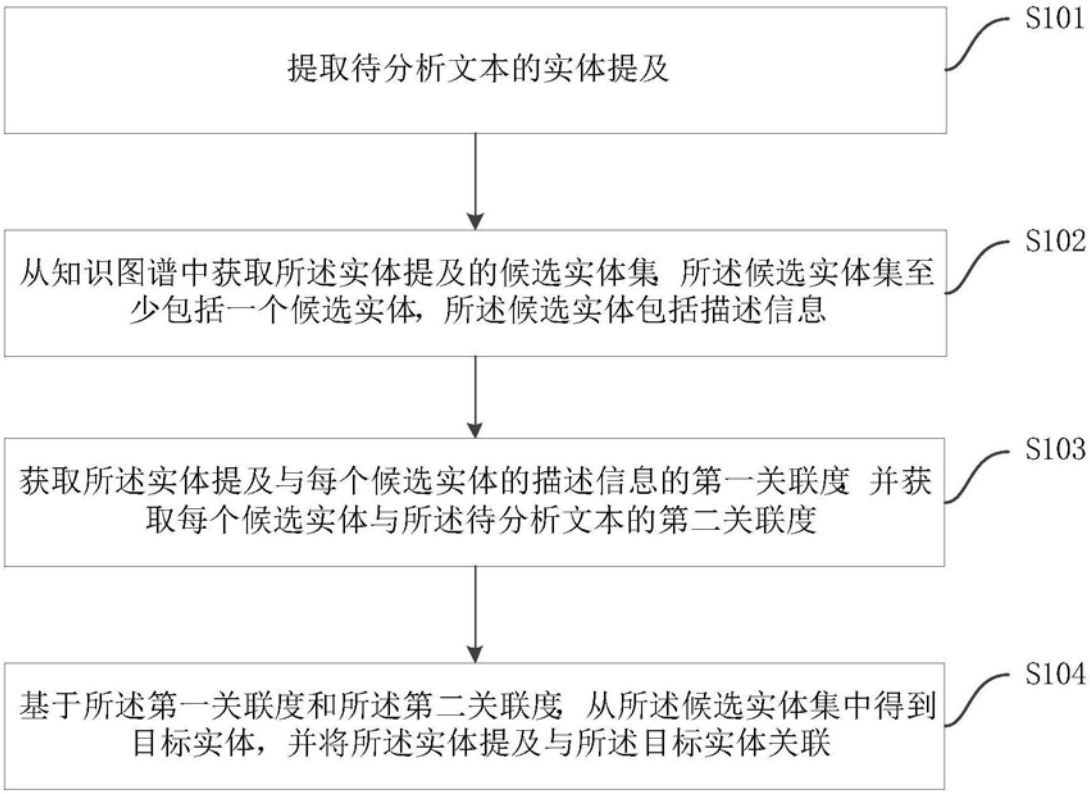

提取待分析文本的实体提及
S101
从知识图谱中获取所述实体提及的候选实体集，所述候选实体集至少包括一个候选实体，所述候选实体包括描述信息
S102
获取所述实体提及与每个候选实体的描述信息的第一关联度，并获取每个候选实体与所述待分析文本的第二关联度
S103
基于所述第一关联度和所述第二关联度，从所述候选实体集中得到目标实体，并将所述实体提及与所述目标实体关联
S104

图1

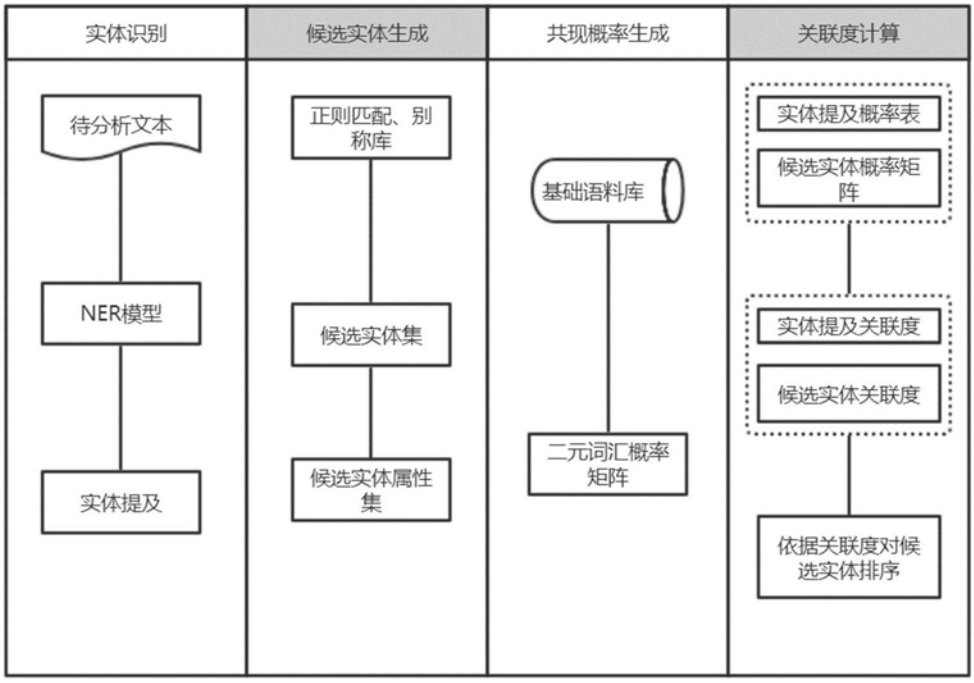

实体识别
候选实体生成
共现概率生成
关联度计算
待分析文本
NER模型
实体提及
正则匹配、别称库
候选实体集
候选实体属性集
基础语料库
二元词汇概率矩阵
实体提及概率表
候选实体概率矩阵
实体提及关联度
候选实体关联度
依据关联度对候选实体排序

图2

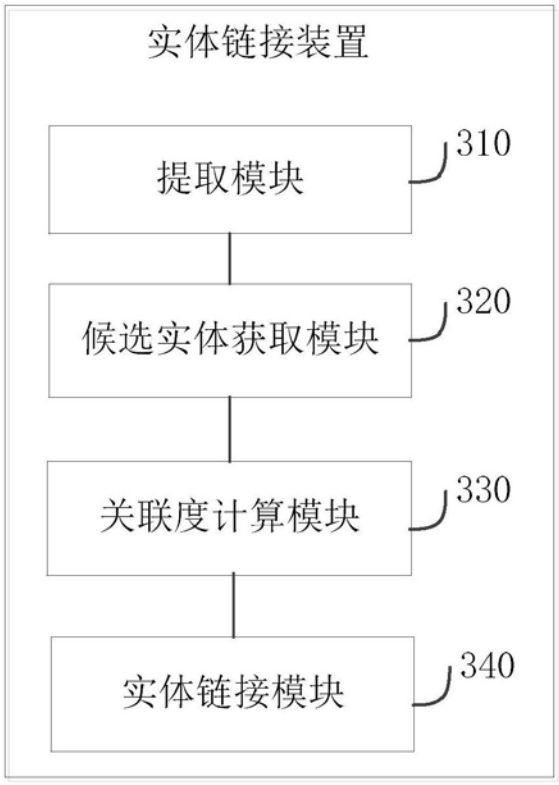

实体链接装置
提取模块
310
候选实体获取模块
320
关联度计算模块
330
实体链接模块
340

图3

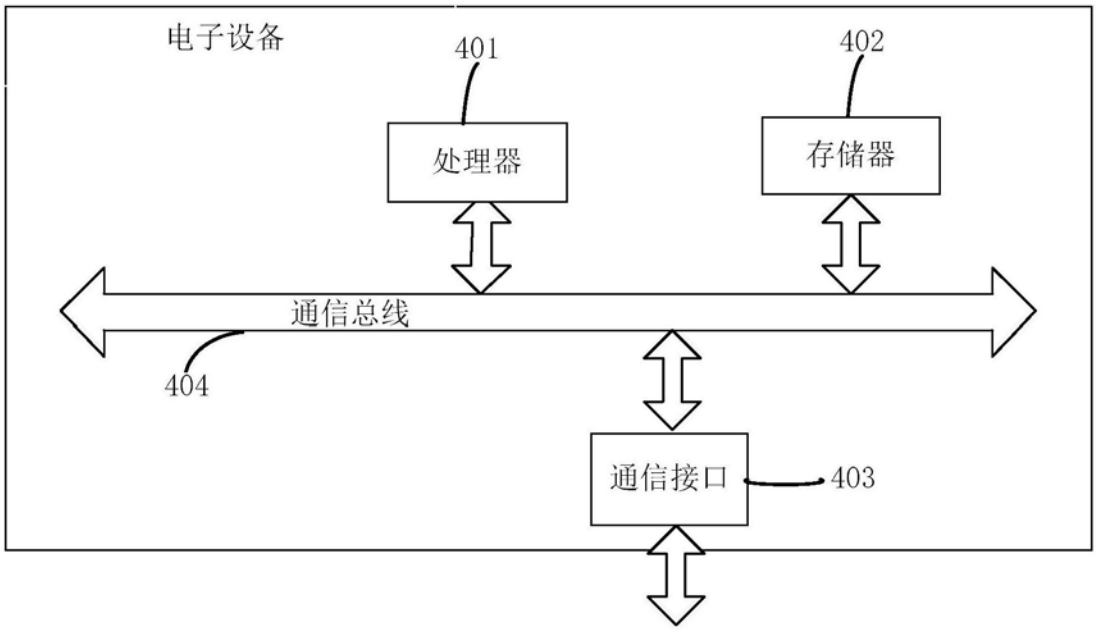

电子设备
401
处理器
402
存储器
通信总线
404
通信接口
403

图4