(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106776544 B
(45)授权公告日 2020.05.05

(21)申请号 201611049141.8

(22)申请日 2016.11.24

(65)同一申请的已公布的文献号
申请公布号 CN 106776544 A

(43)申请公布日 2017.05.31

(73)专利权人 四川无声信息技术有限公司
地址 610041 四川省成都市高新区交子大道365号中海国际中心F座4楼

(72)发明人 黄勇 程芃森 欧晓聪 张磊 许春阳

(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371
代理人 唐维虎

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06F 40/284(2020.01)
G06F 40/253(2020.01)
G06F 40/247(2020.01)

审查员 王青

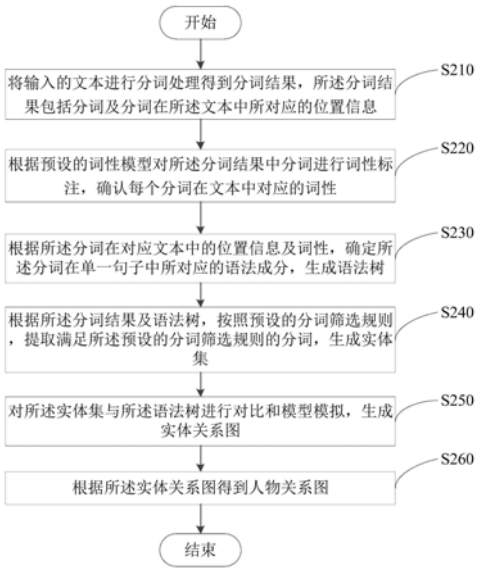
权利要求书3页 说明书11页 附图9页

(54)发明名称

人物关系识别方法及装置和分词方法

(57)摘要

本发明实施例提供一种人物关系识别方法及装置和分词方法,涉及互联网信息处理技术领域。所述方法包括:将输入的文本进行分词处理得到分词结果;对分词结果中分词进行词性标注;确定分词在单一句子中所对应的语法成分,生成语法树;提取满足预设的分词筛选规则的分词,生成实体集;对实体集与语法树进行对比和模型模拟,生成实体关系图;根据实体关系图得到人物关系图。所述方法相对于现有技术中人物关系图的构建过程具有运算量小、运算效率高、耗时少且实现难度较低的特点。



CN 106776544 B

1.一种人物关系识别方法,其特征在于,所述方法包括:

将输入的文本进行分词处理得到分词结果,所述分词结果包括分词及分词在所述文本中所对应的位置信息;

根据预设的词性模型对所述分词结果中分词进行词性标注,确认每个分词在文本中对应的词性;

根据所述分词在对应文本中的位置信息及词性,确定所述分词在单一句子中所对应的语法成分,生成语法树;

根据所述分词结果及语法树,按照预设的分词筛选规则,提取满足所述预设的分词筛选规则的分词,生成实体集,其中所述预设的分词筛选规则为针对生成人物关系交互图所需的相应信息进行挑选的规则,所述生成人物关系交互图所需的相应信息包括与人物相关的真实社会属性信息、虚拟身份属性信息及社会关系信息;

对所述实体集与所述语法树进行对比和模型模拟,生成实体关系图,其中通过利用有向图的模式对所述实体关系图进行信息的梳理总结,基于人物与人物之间的信息联系、人物与物品之间的信息联系、物品与物品之间的信息联系生成所述人物关系图;

根据所述实体关系图得到人物关系图。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括,对所述预设的词性模型进行训练的步骤,所述步骤包括:

读取已标注词性的词性语料集,对所述词性语料集中每个词在不同的词序列中的词性进行统计分析,得到词性集;

对所述词性集进行分析计算,得到每个词在不同的词序列中对应的词性出现概率,以生成所述词性模型。

3.根据权利要求1所述的方法,其特征在于,所述根据所述分词在对应文本中的位置信息及词性信息,确定所述分词在单一句子中所对应的语法成分,生成语法树的步骤包括:

使用预设的语法信息模型得到单一句子中各分词所对应的语法信息概率取值,选择各分词对应语法信息概率取值中的最大概率取值作为各分词对应的输出的语法信息。

4.根据权利要求3所述的方法,其特征在于,所述方法还包括,对所述预设的语法信息模型进行训练的步骤,所述步骤包括:

读取已标注语法的语法语料集,对所述语法语料集中每个词在不同的词序列中的语法信息进行统计分析,得到语法信息集;

对所述语法信息集进行分析计算,得到每个词在不同的词序列中对应出现的语法信息概率,以生成所述语法信息模型。

5.根据权利要求1所述的方法,其特征在于,所述对所述实体集与所述语法树进行对比和模型模拟,生成实体关系图的步骤包括:

将所述实体集内的实体与所述实体在语法树中对应的语法信息进行对比,判断所述实体在语法树中的语法结构成分,判断所述实体是否存在对应的代词指代关系;

当所述实体存在对应的代词指代关系时,将所述代词与其被指代的实体从文本中抽取出来,通过遍历文本中实体的方式,将同一实体所对应的不同代词进行还原,生成实体关系图。

6.一种人物关系识别装置,其特征在于,所述装置包括:

分词模块，用于将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及分词在所述文本中所对应的位置信息；

词性标注模块，用于根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性；

语义解析模块，用于根据所述分词在对应文本中的位置信息及词性信息，确定所述分词在单一句子中所对应的语法成分，生成语法树；

实体识别模块，用于根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集，其中所述预设的分词筛选规则为针对生成人物关系交互图所需的相应信息进行挑选的规则，所述生成人物关系交互图所需的相应信息包括与人物相关的真实社会属性信息、虚拟身份属性信息及社会关系信息；

指代消解模块，用于对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图；

人物关系图生成模块，用于根据所述实体关系图得到人物关系图，其中通过利用有向图的模式对所述实体关系图进行信息的梳理总结，基于人物与人物之间的信息联系、人物与物品之间的信息联系、物品与物品之间的信息联系生成所述人物关系图。

7.根据权利要求6所述的装置，其特征在于，所述语义解析模块通过使用预设的语法信息模型得到单一句子中各分词所对应的语法信息概率取值，选择各分词对应语法信息概率取值中的最大概率取值作为各分词对应的输出的语法信息。

8.一种分词方法，应用于权利要求6-7中任意一项所述的装置，其特征在于，所述方法包括：

读取字典文件，根据所述字典文件生成前缀树；

获取待处理的文本，将所述待处理的文本切分为多个语句，生成句子集；

对所述句子集中的语句进行分词处理；

判断所述句子集中是否还有未分词完成的语句；

当所述句子集中还有未分词完成的语句时，采用所述前缀树对所述语句进行循环遍历查找分词；

当所述语句中存在未被分词的部分时，提取所述语句中未被分词的部分，通过隐含马尔可夫模型对所述未被分词的部分进行分词处理，得到分词结果，并返回所述判断所述句子集中是否还有未分词完成的语句的步骤继续执行，直到所述句子集中的所有语句分词完成。

9.根据权利要求8所述的方法，其特征在于，所述采用所述前缀树对所述语句进行循环遍历查找分词的步骤，包括：

以所述语句开头第一个文字为开始在所述前缀树中查找以所述第一个文字开始的词；

如果查找成功，在所述前缀树中查找以所述查找成功的文字为起点，以所述语句中距离查找成功的文字对应词的长度的文字为开始的词；

如果查找失败，从所述语句中查找失败的文字的后一个文字开始在所述前缀树中查找以所述文字开始的词。

10.根据权利要求8所述的方法，其特征在于，所述通过隐含马尔可夫模型对所述语句部分进行分词处理，得到分词结果的步骤，包括：

根据所述隐含马尔可夫模型结合维特比算法从提取出的所述语句部分中找到一个隐藏状态序列；

根据模式匹配算法匹配出所述隐藏状态序列中以起始字开始，结束字结束的词语，或者单独成词的词语；

将匹配出的所述词语组成一个集合，得到所述分词结果。

人物关系识别方法及装置和分词方法

技术领域

[0001] 本发明涉及互联网信息处理技术领域，具体而言，涉及一种人物关系识别方法及装置和分词方法。

背景技术

[0002] 随着互联网技术的发展，尤其是基于互联网的各种社交媒体的发展，目前针对社会关系网络的研究已逐渐成为了当下的热点。人们每天通过各种互联网社交媒体发布各种各样的信息，这些信息在有意或无意中可能含有信息提供者或其他人的相关信息。上述相关信息不仅可以包括如：人名、家庭地址、工作地址、电话等真实社会属性信息；也可以包括如：电子邮件、微信号、QQ号等虚拟身份属性信息；同时也可能包括人物间的相互称谓等社会关系信息。

[0003] 因此，可以以互联网上信息为输入，对上述输入信息进行处理后生成一张以人物为节点的社会关系交互图，该社会关系交互图中可以含有人物真实社会身份属性和虚拟身份属性，关系含有称谓信息。

[0004] 但就现有技术而言，现有的人物关系交互图的实现具有运算量大、运算效率低、耗费时间长及实现难度高等缺点。

发明内容

[0005] 为了克服现有技术中的上述不足，本发明实施例的目的在于提供一种用于构造运算量小、运算效率高、耗时少且实现难度较低的人物关系交互图的人物关系识别方法及装置和分词方法，以改善现有技术中人物关系交互图实现时暴露出的问题，向用户提供丰富而准确的人物关系信息。

[0006] 就人物关系识别方法而言，本发明较佳的实施例提供了一种人物关系识别方法。所述方法包括：

[0007] 将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及分词在所述文本中所对应的位置信息；

[0008] 根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性；

[0009] 根据所述分词在对应文本中的位置信息及词性，确定所述分词在单一句子中所对应的语法成分，生成语法树；

[0010] 根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集；

[0011] 对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图；

[0012] 根据所述实体关系图得到人物关系图。

[0013] 就人物关系识别装置而言，本发明较佳的实施例提供了一种人物关系识别装置。所述装置包括：

[0014] 分词模块，用于将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及分词在所述文本中所对应的位置信息；

[0015] 词性标注模块，用于根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性；

[0016] 语义解析模块，用于根据所述分词在对应文本中的位置信息及词性信息，确定所述分词在单一句子中所对应的语法成分，生成语法树；

[0017] 实体识别模块，用于根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集；

[0018] 指代消解模块，用于对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图；

[0019] 人物关系图生成模块，用于根据所述实体关系图得到人物关系图。

[0020] 就分词方法而言，本发明较佳的实施例提供了一种分词方法，应用于上述的人物关系识别装置。所述方法包括：

[0021] 读取字典文件，根据所述字典文件生成前缀树；

[0022] 获取待处理的文本，将所述待处理的文本切分为多个语句，生成句子集；

[0023] 对所述句子集中的语句进行分词处理；

[0024] 判断所述句子集中是否还有未分词完成的语句；

[0025] 当所述句子集中还有未分词完成的语句时，采用所述前缀树对所述语句进行循环遍历查找分词；

[0026] 当所述语句中存在未被分词的部分时，提取所述语句中未被分词的部分，通过隐含马尔可夫模型对所述未被分词的部分进行分词处理，得到分词结果，并返回所述判断所述句子集中是否还有未分词完成的语句的步骤继续执行，直到所述句子集中的所有语句分词完成。

[0027] 相对于现有技术而言，本发明实施例提供的人物关系识别方法及装置和分词方法具有以下有益效果：所述方法通过对输入文本进行分词处理，并对分词处理后得到的分词进行词性和语法信息的标注，提取所述分词中符合预设的分词筛选规则的分词，让所述分词与语法树进行对比和模型模拟，指代生成实体关系图，得到人物关系图。所述方法相对于现有技术中人物关系图的构建过程具有运算量小、运算效率高、耗时少且实现难度较低的特点。

[0028] 为使本发明的上述目的、特征和优点能更明显易懂，下文特举本发明较佳实施例，并配合所附附图，作详细说明如下。

附图说明

[0029] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，应当理解，以下附图仅示出了本发明的某些实施例，因此不应被看作是对范围的限定，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相关的附图。

[0030] 图1为本发明较佳实施例提供的服务器与至少一用户终端进行通信的交互示意图。

[0031] 图2为图1所示的服务器的方框示意图。

[0032] 图3为本发明较佳实施例提供的图2中所示的人物关系识别装置的一种功能模块框图。

[0033] 图4为本发明较佳实施例的一种实施方式提供的图2中所示的人物关系识别装置的一种功能模块框图。

[0034] 图5为本发明较佳实施例的另一种实施方式提供的图2中所示的人物关系识别装置的一种功能模块框图。

[0035] 图6为本发明较佳实施例提供的人物关系识别方法的一种流程示意图。

[0036] 图7为本发明较佳实施例的一种实施方式提供的人物关系识别方法的一种流程示意图。

[0037] 图8为本发明较佳实施例的另一种实施方式提供的人物关系识别方法的一种流程示意图。

[0038] 图9为本发明较佳实施例提供的一种分词方法的一种流程示意图。

[0039] 图10为图9中步骤S350的子步骤的流程示意图。

[0040] 图11为图9中步骤S360的子步骤的流程示意图。

[0041] 图12为本发明较佳实施例提供的一种分词方法的另一种流程示意图。

[0042] 图标：10-服务器；20-用户终端；30-网络；11-存储器；12-处理器；13-通信单元；100-人物关系识别装置；110-分词模块；120-词性标注模块；130-语义解析模块；140-实体识别模块；150-指代消除模块；160-人物关系图生成模块；170-词性模型训练模块；180-语法模型训练模块。

具体实施方式

[0043] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例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。通常在此处附图中描述和示出的本发明实施例的组件可以以各种不同的配置来布置和设计。

[0044] 因此，以下对在附图中提供的本发明的实施例的详细描述并非旨在限制要求保护的本发明的范围，而是仅仅表示本发明的选定实施例。基于本发明中的实施例，本领域普通技术人员在没有作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。

[0045] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步定义和解释。

[0046] 在本发明的描述中，需要说明的是，术语“下”、“后”等指示的方位或位置关系为基于附图所示的方位或位置关系，或者是该发明产品使用时惯常摆放的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”仅用于区分描述，而不能理解为指示或暗示相对重要性。

[0047] 对于本领域的普通技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。

[0048] 下面结合附图，对本发明的一些实施方式作详细说明。在不冲突的情况下，下述的实施例及实施例中的特征可以相互组合。

[0049] 请参照图1，是本发明较佳实施例提供的服务器10与至少一用户终端20进行通信的交互示意图。所述用户终端20可通过网络30访问服务器10，以使用服务器10提供的人物关系识别服务。本实施例中，所述服务器10可以是，但不限于，web(网站)服务器。所述用户终端20可以是，但不限于，智能手机、个人电脑(personal computer，PC)、平板电脑、个人数字助理(personal digital assistant，PDA)、移动上网设备(mobile Internet device，MID)等。所述网络30可以是，但不限于，有线网络或无线网络。

[0050] 请参照图2，是图1所示的服务器10的方框示意图。所述服务器10可以包括人物关系识别装置100、存储器11、处理器12、及通信单元13。

[0051] 所述存储器11、处理器12以及通信单元13各个元件相互之间直接或间接地电性连接，以实现数据的传输或交互。例如，这些元件相互之间可通过一条或多条通讯总线或信号线实现电性连接。

[0052] 其中，所述存储器11可以是，但不限于，随机存取存储器(Random Access Memory，RAM)，只读存储器(Read Only Memory，ROM)，可编程只读存储器(Programmable Read-Only Memory，PROM)，可擦除只读存储器(Erasable Programmable Read-Only Memory，EPROM)，电可擦除只读存储器(Electric Erasable Programmable Read-Only Memory，EEPROM)等。其中，存储器11用于存储程序，所述处理器12在接收到执行指令后，执行所述程序。所述通信单元13用于通过所述网络30建立所述服务器10与用户终端20之间的通信连接，并用于通过所述网络30收发数据。

[0053] 所述人物关系识别装置100包括至少一个可以软件或固件(firmware)的形式存储于所述存储器11中或固化在所述服务器10的操作系统(operating system，OS)中的软件功能模块。所述处理器12用于执行所述存储器11中存储的可执行模块，例如所述人物关系识别装置100所包括的软件功能模块及计算机程序等。本实施例中，所述人物关系识别装置100为服务器10提供针对不同用户的人物关系识别服务，具体的方法在后续进行详细说明。本实施例中，所述人物关系识别装置100可以是服务器10提供的关系识别引擎或者属于该关系识别引擎的一部分(如该关系识别引擎的其中一个功能模块)。所述关系识别引擎是指根据一定的策略、运用计算机程序从网络30(如互联网)上搜集人物之间的相关信息，在对相关信息进行组织和处理后，得到人物关系交互图，为用户提供人物关系识别服务，将用户需要查找的人物的相关信息(例如人名、家庭地址、工作地址、电子邮件、微信号、QQ号等)展示给用户的系统。

[0054] 可以理解的是，图2所示的结构仅为服务器10的结构示意图，所述服务器10还可包括比图2中所示更多或者更少的组件，或者具有与图2所示不同的配置。图2中所示的各组件可以采用硬件、软件或其组合实现。

[0055] 请参照图3，是本发明较佳实施例提供的图2中所示的人物关系识别装置100的一种功能模块框图。所述人物关系识别装置100包括分词模块110、词性标注模块120、语义解析模块130、实体识别模块140、指代消除模块150及人物关系图生成模块160。

[0056] 所述分词模块110用于将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及分词在所述文本中所对应的位置信息。

[0057] 具体地，所述分词处理为根据预设分词策略将所述输入的文本分隔为多个词的过程，其中，预设分词策略可以采用现有成熟的分词策略，也可以根据实际需求对其进行相应的调整。所述分词结果包括经过上述分词处理后得到的各个分词及各个分词在所述输入的文本中所述对应的位置信息。

[0058] 在本实施例中，所述分词处理可以采用后续描述的分词方法进行分词。

[0059] 所述词性标注模块120用于根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性。

[0060] 在自然语言处理领域中，若要对一个自然语句进行深入研究，一般都需要对出现在文本中的各词语词性进行分析，确认词语在文本中相应的词性。具体地，在本实施例中，对所述分词结果中的分词进行词性标注，确认每个分词在文本中对应的词性，可方便对人物关系识别的后续工作。所述词性可以是，但不限于，名词、代词、动词、形容词、数词、副词等。

[0061] 所述语义解析模块130用于根据所述分词在对应文本中的位置信息及词性，确定所述分词在单一句子中所对应的语法成分，生成语法树。

[0062] 在自然语言处理领域中，对一个自然语句进行深入研究，除了需要对出现在文本中的各词语词性进行分析外，还需对各词语在文本中相对应的语法信息进行了解分析。所述语法成分可以是，但不限于，主语、状语、谓语、宾语等。

[0063] 具体地，在本实施例中，使用预设的语法信息模型得到单一句子中各分词所对应的语法信息概率取值，选择各分词对应语法信息概率取值中的最大概率取值作为各分词对应的输出的语法信息。

[0064] 所述实体识别模块140用于根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集。

[0065] 具体地，所述预设的分词筛选规则为用户根据实际需求设定的挑选满足用户需求的分词的规则。在本实施例中，所述预设的分词筛选规则为针对生成人物关系交互图所需的相应信息进行挑选的规则。所述相应信息不仅可以包括如：人物姓名、家庭地址、工作地址、电话等真实社会属性信息；也可以包括如：电子邮件、微信号、QQ号等虚拟身份属性信息；同时也可能包括人物间的相互称谓等社会关系信息。

[0066] 所述指代消除模块150用于对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图。

[0067] 具体地，所述对比和模型模拟为对实体集内的实体与该实体在语法树中可能对应的语法信息进行对比选择和将实体带有相应的语法信息置入语句之中进行模拟，形成代词指代关系模型的过程，进而判断实体集中实体对应的合适的语法信息，判断实体是否存在相应的代词指代关系。

[0068] 在本实施例中，所述指代消除模块150对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图的方式包括：

[0069] 将所述实体集内的实体与所述实体在语法树中对应的语法信息进行对比，判断所述实体在语法树中的语法结构成分，判断所述实体是否存在对应的代词指代关系；

[0070] 当所述实体存在对应的代词指代关系时，将所述代词与其被指代的实体从文本中抽取出来，通过遍历文本中实体的方式，将同一实体所对应的不同代词进行还原，生成实体

关系图。

[0071] 具体地，所述语法树是以计算机组织方法由自然语言中的语法结构表述而成一种树形结构，该结构严格遵守主谓宾、主系表等语法规则。依照自然语言结构规则，判断实体在语法树中所担任的语法结构成分，判断其是否存在对应的代词指代关系。将代词的原始表述形式，如：你、我、他、她等，同其被指代的实体词从原始文档中抽取出来，然后再在此基础上遍历实体词，将同一实体词所对应的不同代词形式一一还原，即可生成实体关系。再根据生成的实体关系，得到相应的实体关系图。所述实体包括上述的相关信息。

[0072] 所述人物关系图生成模块160用于根据所述实体关系图得到人物关系图。

[0073] 具体地，利用有向图的模式对所述实体关系图进行信息的梳理总结，基于人物与人物之间、人物与物品之间、物品与物品之间的信息联系生成人物关系图。所述人物关系图表述了人物与人物之间、人物与物品之间、物品与物品之间的关系。

[0074] 请参照图4，在本实施例的一种较佳实施方式中，所述人物关系识别装置100还可以包括：词性模型训练模块170。

[0075] 所述词性模型训练模块170用于对预设的词性模型进行训练。在本实施方式中，所述词性模型训练模块170对预设的词性模型进行训练的方式可以包括：

[0076] 读取已标注词性的词性语料集，对所述词性语料集中每个词在不同的词序列中的词性进行统计分析，得到词性集；

[0077] 对所述词性集进行分析计算，得到每个词在不同的词序列中对应的词性出现概率，以生成所述词性模型。

[0078] 具体地，所述词性语料集为一种存放着已经标注了相应词性的可在语言的实际使用中真实出现过的语言材料的以电子计算机为载体承载语言知识的基础资源。训练词性模型便是对该词性语料集中每个词在不同的词序列中对应的词性出现概率的总结，可方便人物关系图的建立。

[0079] 请参照图5，在本实施例的另一种较佳实施方式中，所述人物关系识别装置100还可以包括：语法模型训练模块180。

[0080] 所述语法模型训练模块180用于对预设的语法信息模型进行训练。在本实施方式中，所述语法模型训练模块180对预设的语法信息模型进行训练的方式可以包括：

[0081] 读取已标注语法的语法语料集，对所述语法语料集中每个词在不同的词序列中的语法信息进行统计分析，得到语法信息集；

[0082] 对所述语法信息集进行分析计算，得到每个词在不同的词序列中对应出现的语法信息概率，以生成所述语法信息模型。

[0083] 具体地，所述语法语料集与所述词性语料集类似，所述语法语料集为一种存放着已经标注了相应语法信息的可在语言的实际使用中真实出现过的语言材料的以电子计算机为载体承载语言知识的基础资源。训练语法信息模型便是对每个词在不同的词序列中对应出现的语法信息概率的总结，可以提高所述人物关系图的构建效率。

[0084] 请参照图6，本发明较佳的实施例提供一种人物关系识别方法的具体流程示意图，以下对上述方法的具体流程进行描述。

[0085] 在本实施例中，所述人物关系识别方法可以包括以下步骤：

[0086] 步骤S210，将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及

分词在所述文本中所对应的位置信息。

[0087] 在本实施例中，所述步骤S210由图3中所示的分词模块110执行，所述步骤S210详细描述可以参照对所述分词模块110的具体描述。

[0088] 步骤S220，根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性。

[0089] 在本实施例中，所述步骤S220由图3中所示的词性标注模块120执行，所述步骤S220详细描述可以参照对所述词性标注模块120的具体描述。

[0090] 步骤S230，根据所述分词在对应文本中的位置信息及词性，确定所述分词在单一句子中所对应的语法成分，生成语法树。

[0091] 在本实施例中，所述步骤S230由图3中所示的语义解析模块130执行，所述步骤S230详细描述可以参照对所述语义解析模块130的具体描述。

[0092] 步骤S240，根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集。

[0093] 在本实施例中，所述步骤S240由图3中所示的实体识别模块140执行，所述步骤S240详细描述可以参照对所述实体识别模块140的具体描述。

[0094] 步骤S250，对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图。

[0095] 在本实施例中，所述步骤S250由图3中所示的指代消除模块150执行，所述步骤S250详细描述可以参照对所述指代消除模块150的具体描述。

[0096] 具体地，所述步骤S250可以包括：

[0097] 将实体集内的实体与所述实体在语法树中对应的语法信息进行对比，判断所述实体在语法树中的语法结构成分，判断所述实体是否存在对应的代词指代关系；

[0098] 当所述实体存在对应的代词指代关系时，将所述代词与其被指代的实体从文本中抽取出来，通过遍历文本中实体的方式，将同一实体所对应的不同代词进行还原，生成实体关系图。

[0099] 步骤S260，根据所述实体关系图得到人物关系图。

[0100] 在本实施例中，所述步骤S260由图3中所示的人物关系图生成模块160执行，所述步骤S260详细描述可以参照对所述人物关系图生成模块160的具体描述。

[0101] 请参照图7，所述人物关系识别方法还可以包括：

[0102] 步骤S208，对预设的词性模型进行训练。

[0103] 在本实施例中，所述步骤S208由图4中所示的词性模型训练模块170执行，所述步骤S208详细描述可以参照对所述词性模型训练模块170的具体描述。

[0104] 请参照图8，所述人物关系识别方法还可以包括：

[0105] 步骤S209，对预设的语法信息模型进行训练。

[0106] 在本实施例中，所述步骤S209由图5中所示的语法模型训练模块180执行，所述步骤S209详细描述可以参照对所述语法模型训练模块180的具体描述。

[0107] 请参照图9，本发明较佳的实施例提供一种分词方法的一种流程示意图，所述分词方法应用于上述的人物关系识别装置，以下对上述方法的具体流程进行描述。

[0108] 在本实施例中，所述分词方法可以包括以下步骤：

[0109] 步骤S310，读取字典文件，根据所述字典文件生成前缀树。

[0110] 具体地，字典文件是计算机中预先定义好的常用词语数据库，而生成前缀树的过程就是把计算机中预先定义好的字典文件表述为内存空间占用少、数据查询速度快的成树状的计算机内存结构。在本实施例中，从硬盘中读取所述字典文件。

[0111] 步骤S320，获取待处理的文本，将所述待处理的文本切分为多个语句，生成句子集。

[0112] 具体地，以中文的句子分隔符和文本文件换行符为标识将所述待处理的文本分成多个语句，所述句子分隔符可以是，但不限于，句号、叹号、问号、分号等。

[0113] 步骤S330，对所述句子集中的语句进行分词处理。

[0114] 具体地，所述分词处理的步骤可以包括后续的步骤S350及步骤360。

[0115] 步骤S340，判断所述句子集中是否还有未分词完成的语句。

[0116] 具体地，当所述句子集中没有未分词完成的语句时，结束整个分词过程，完成了对待处理文本的分词处理。

[0117] 当所述句子集中有未分词完成的语句时，继续执行后续的步骤S350及步骤S360，直到整个句子集中再无未分词完成的语句为止。

[0118] 步骤S350，采用所述前缀树对所述语句进行循环遍历查找分词。

[0119] 具体地，请参照图10，所述步骤S350可以包括：

[0120] 子步骤S351，以语句开头第一个文字为开始在所述前缀树中查找以所述第一个文字开始的词。

[0121] 子步骤S352，如果查找成功，在所述前缀树中查找以所述查找成功的文字为起点，以所述语句中距离查找成功的文字对应词的长度的文字为开始的词。

[0122] 子步骤S353，如果查找失败，从所述语句中查找失败的文字的后一个文字开始在所述前缀树中查找以所述文字开始的词。

[0123] 具体地，假设一个句子集用{Si}表示，其中Si(i＝1、2、3、……、n)分别表示着句子集中相应的一个句子。所述循环遍历查找分词便是从1到n依次从句子集中取出一个句子，然后对该句子进行查找，完成分词。

[0124] 在本实施例中，假设一个句子的长度为L，则查找分词的步骤可具体为：

[0125] 从该长度为L的句子开头的第一个文字开始，在前缀树中查找以该文字开始的词。

[0126] 如果该词能在所述句子上找到相应的位置，即所述句子上相应的位置上的文字组成的词与所述的词相同的话，即为查找成功，然后在所述前缀树中查找以所述查找成功的文字为起点，以所述语句中距离查找成功的文字对应词的长度的文字为开始的词。

[0127] 具体地，假设查找到的词的长度为l，该词开头第一文字在所述句子中对应的位置为a点，则在长度为L的句子中，从a点所对应的位置开始向后移动长度为l的距离，到达b点位置，找到该句子中b点位置对应的文字，然后在前缀树中继续查找以b点位置上的文字为开始的词。

[0128] 如果该词不能在所述句子上找到相应的位置，即所述句子上相应的位置上的文字组成的词与所述的词不同的话，即为查找失败，然后从所述语句中查找失败的文字的后一个文字开始在所述前缀树中查找以所述文字开始的词，所述文字为查找失败的文字后面的那一个文字。

[0129] 具体地，如果查找失败对应在长度为L的句子上的位置为c点的话，那么就是从c点

开始向后移动一个文字的位置，即c点后面一个文字的位置开始在前缀树中查找以所述文字为开始的词，所述文字为c点后面的一个文字。

[0130] 步骤S360，当所述语句中存在未被分词的部分时，提取所述语句中未被分词的部分，通过隐含马尔可夫模型对所述未被分词的部分进行分词处理，得到分词结果。

[0131] 具体地，所述语句表示的是句子集中经历了步骤S350后的语句。

[0132] 请参照图11，所述步骤S360可以包括：

[0133] 根据隐含马尔可夫模型结合维特比算法从提取出的所述语句部分中找到一个隐藏状态序列；

[0134] 根据模式匹配算法匹配出所述隐藏状态序列中以起始字开始，结束字结束的词语，或者单独成词的词语；

[0135] 将匹配出的所述词语组成一个集合，得到所述分词结果。

[0136] 具体地，维特比算法是一种用于寻找最有可能产生目标观察序列维特比路径(隐含状态序列)的动态规划算法。

[0137] 给定的隐含马尔可夫模型的各个参数如下：

[0138] 状态空间S＝{B(词语的起始字)，E(词语的结束字)，M(词语的中间字)，S(单独成词)}；

[0139] 大小为4的初始概率数组p，其中pi(i＝B,E,M,S)是状态初始值为B,E,M,S的概率；

[0140] 观察值空间O{O1,O2,……,On}(n为模型中不重复的字的个数)；

[0141] A为4×4的转移矩阵，把从状态Si到状态Sj(i,j＝B,E,M,S)的转移概率记为aij；

[0142] B为4×N的混淆矩阵，其中bij表示在状态Si的前提下观察到Oj的概率，令观察到的输出值为Y＝{y1,y2,……,yt}，称X＝{x1,x2,……,xt}为生成观察值Y＝{y1,y2,……,yt}的状态序列。

[0143] 其核心计算如下：

[0144] V1,k＝P(y1|k)×pk

[0145] Vt,k＝P(yt|k)×max(axk×Vt-1,x)

[0146] 其中Vt,k是前t个最终状态为k的观察结果最有可能对应的状态序列的概率。通过记录第二个等式中所用到的状态x的转化轨迹便可获得维特比路径，也就得到了隐藏状态序列。

[0147] 模式匹配算法是单独的一种常用计算机算法。从目标序列的第一个状态起与模式序列的第一个状态比较，若相等，则继续对序列进行后续的比较，否则目标序列从第二个状态起与模式序列的第一个状态重新比较，直至模式序列中的每个状态依次和目标序列中的一个连续的状态序列相等为止，此时称为匹配成功，否则匹配失败。

[0148] 通过运用维特比算法、模式匹配算法及隐含马尔可夫模型对所述语句中还未进行分词处理的部分进行分词，可对句子集进行更细致的分词，分词效率更高，便于解决现有技术中人物关系图的构造过程中的技术问题。

[0149] 在本实施例中，所述分词方法还包括，在步骤S360完成之后返回步骤S340，判断句子集中是否还有未分词完成的语句，然后依次执行下去，直到所述句子集中的所有语句分词完成。

[0150] 请参照图12，本发明较佳的实施例提供的分词方法的另一种流程示意图。所述方

法还可以包括：

[0151] 步骤S307，获取训练样本，对所述训练样本的第一个文字进行统计分析，得到初始状态，其中，所述初始状态为所述第一个文字在句中作为起始字的概率、中间字的概率、结束字的概率或单字成词的概率。

[0152] 具体地，通过对训练样本中的句子的第一个文字属于词语的起始文字、词语的结束字、词语的中间字或单字成词等四个状态进行统计，如文本开头的第一个文字只可能为词语的首字(B)或者单字成词(S)的状态，得到文本的初始状态，所述初始状态即为句子的第一个文字是属于{B,E,M,S}这四种状态的概率。

[0153] 步骤S308，对所述训练样本中各个状态下的文字所对应的下一状态进行统计，计算所述下一状态的出现概率，得到转移矩阵，并根据各状态下不同文字的出现概率，生成混淆矩阵。

[0154] 所述下一状态即为与文字相连的下一文字所对应的状态。具体地，步骤S307中生成的初始状态的集合中只含有B,E,M,S四种状态，所以对训练样本中各个状态所对应的下一状态进行统计，计算其出现概率，得到转移矩阵值。而从B转移到B的概率为0，即不存在具有两个连续起始字状态的词语，符合状态集的设定含义，B状态的下一状态只能为M或E。因此，转移矩阵即为一个4×4的二维矩阵，其中部分转移概率为0。

[0155] 而以字为单位遍历整个训练样本，然后统计所有的文在组成的词中所属于的B,E,M,S四种状态的概率，进而生成一个4*m的矩阵，该矩阵便为混淆矩阵。其中，m表示不重复的文字的个数。具体的相关参数可参照步骤S360详细描述中的隐含马尔可夫模型的相关参数。

[0156] 步骤S309，根据所述转移矩阵和混淆矩阵生成隐含马尔可夫模型。

[0157] 具体地，通过对文字的现有状态和下一状态的概率进行统计，找到训练文件中文字与文字之间关于概率的联系，从而生成隐含马尔可模型。

[0158] 综上所述，本发明实施例提供的人物关系识别方法及装置和分词方法。所述方法通过对输入文本进行分词处理，并对分词处理后得到的分词进行词性和语法信息的标注，提取所述分词中符合预设的分词筛选规则的分词，让所述分词与语法树进行对比和模型模拟，指代生成实体关系图，得到人物关系图。所述方法相对于现有技术中人物关系图的构建过程具有运算量小、运算效率高、耗时少且实现难度较低的特点。

[0159] 在本发明实施例所提供的几个实施例中，应该理解到，所揭露的装置和方法，也可以通过其它的方式实现。以上所描述的装置和方法实施例仅仅是示意性的，例如，附图中的流程图和框图显示了根据本发明的多个实施例的装置、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现方式中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的装置来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0160] 另外，在本发明各个实施例中的各功能模块可以集成在一起形成一个独立的部

分，也可以是各个模块单独存在，也可以两个或两个以上模块集成形成一个独立的部分。

[0161] 所述功能如果以软件功能模块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，电子设备，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。需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0162] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

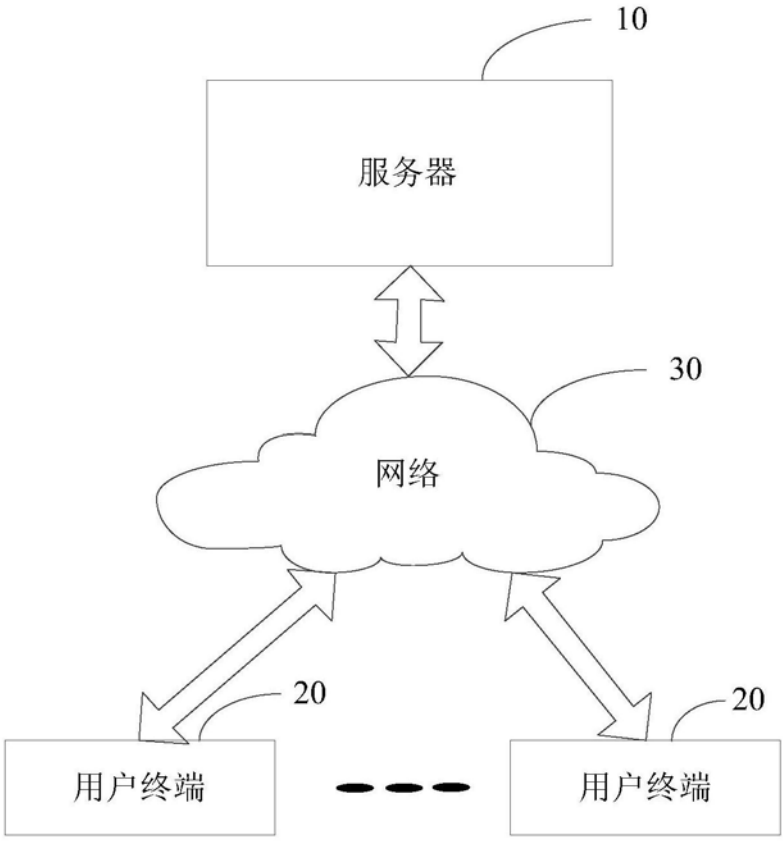

10
服务器
30
网络
20
20
用户终端
用户终端

图1

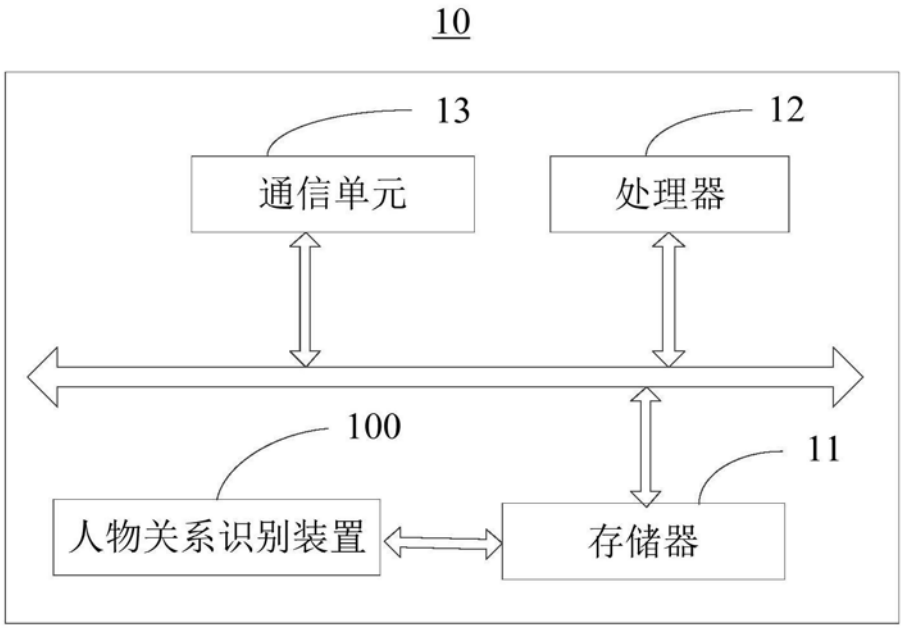

10
13
12
通信单元
处理器
100
11
人物关系识别装置
存储器

图2


100
分词模块
110
词性标注模块
120
语义解析模块
130
实体识别模块
140
指代消除模块
150
人物关系图生成模块
160

图3


100
词性模型训练模块
170
分词模块
110
词性标注模块
120
语义解析模块
130
实体识别模块
140
指代消除模块
150
人物关系图生成模块
160

图4


100
词性模型训练模块
170
语法模型训练模块
180
分词模块
110
词性标注模块
120
语义解析模块
130
实体识别模块
140
指代消除模块
150
人物关系图生成模块
160

图5

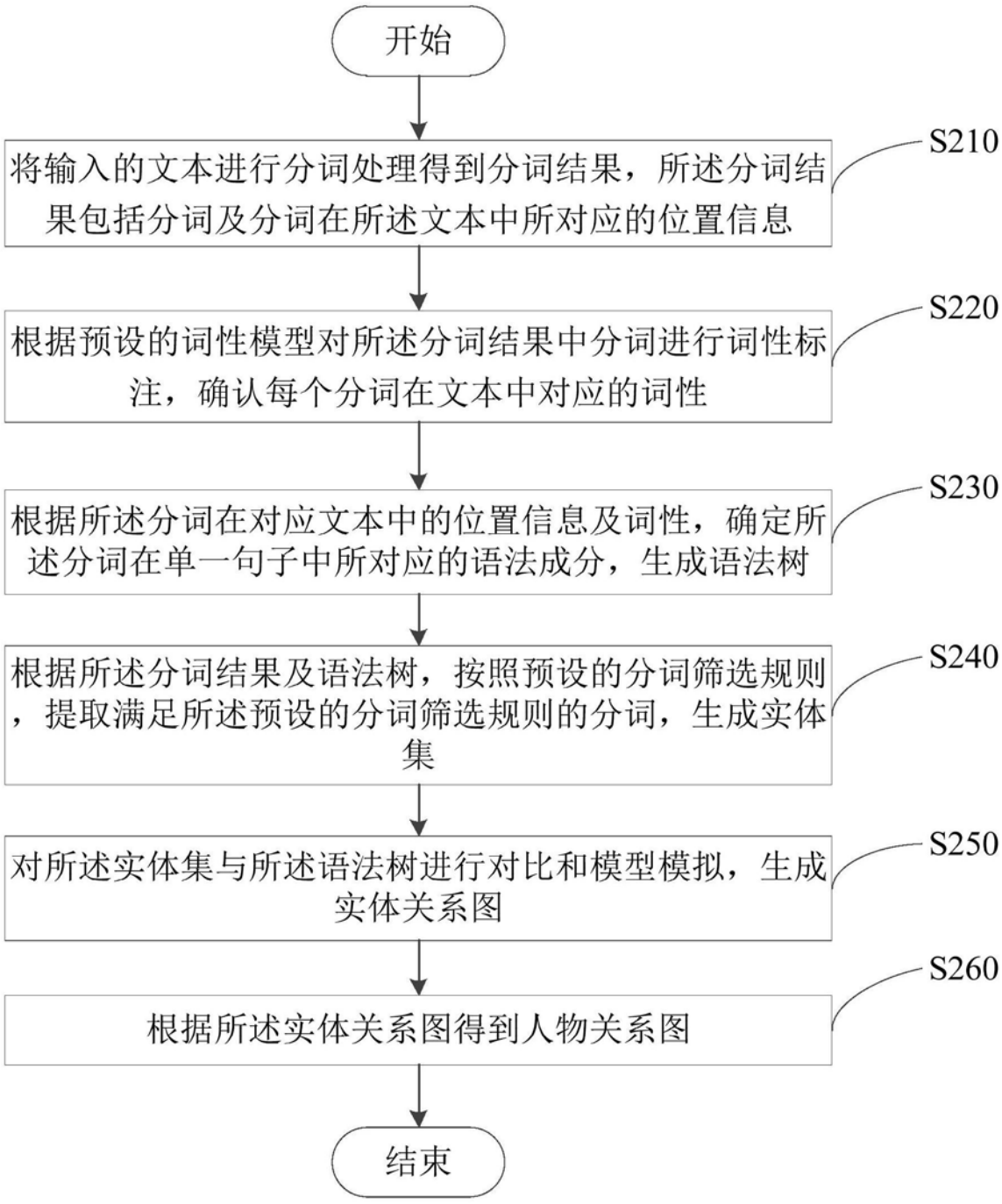

开始
将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及分词在所述文本中所对应的位置信息
S210
根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性
S220
根据所述分词在对应文本中的位置信息及词性，确定所述分词在单一句子中所对应的语法成分，生成语法树
S230
根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集
S240
对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图
S250
根据所述实体关系图得到人物关系图
S260
结束

图6

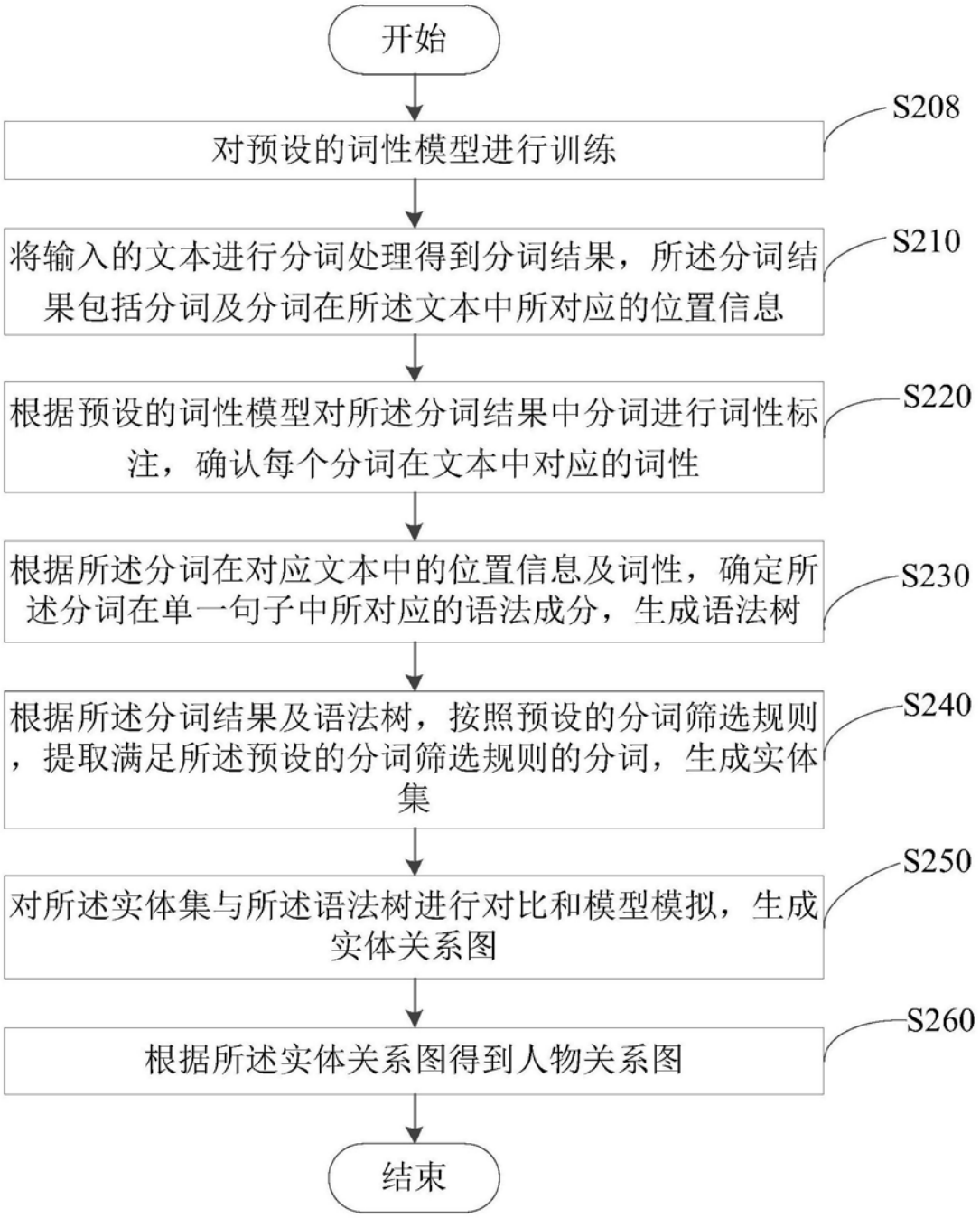

开始
S208
对预设的词性模型进行训练
S210
将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及分词在所述文本中所对应的位置信息
S220
根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性
S230
根据所述分词在对应文本中的位置信息及词性，确定所述分词在单一句子中所对应的语法成分，生成语法树
S240
根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集
S250
对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图
S260
根据所述实体关系图得到人物关系图
结束

图7

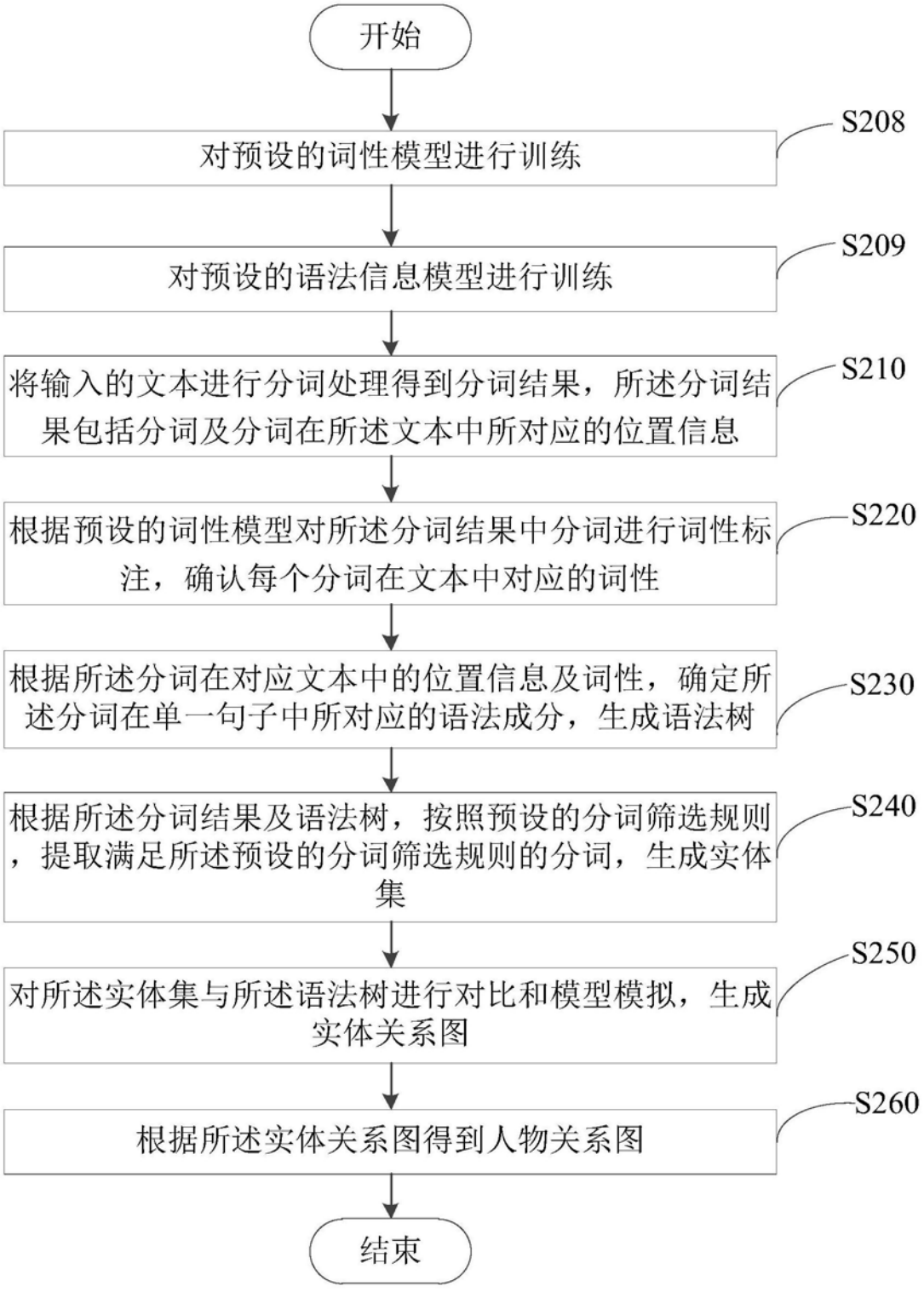

开始
对预设的词性模型进行训练
S208
对预设的语法信息模型进行训练
S209
将输入的文本进行分词处理得到分词结果，所述分词结果包括分词及分词在所述文本中所对应的位置信息
S210
根据预设的词性模型对所述分词结果中分词进行词性标注，确认每个分词在文本中对应的词性
S220
根据所述分词在对应文本中的位置信息及词性，确定所述分词在单一句子中所对应的语法成分，生成语法树
S230
根据所述分词结果及语法树，按照预设的分词筛选规则，提取满足所述预设的分词筛选规则的分词，生成实体集
S240
对所述实体集与所述语法树进行对比和模型模拟，生成实体关系图
S250
根据所述实体关系图得到人物关系图
S260
结束

图8

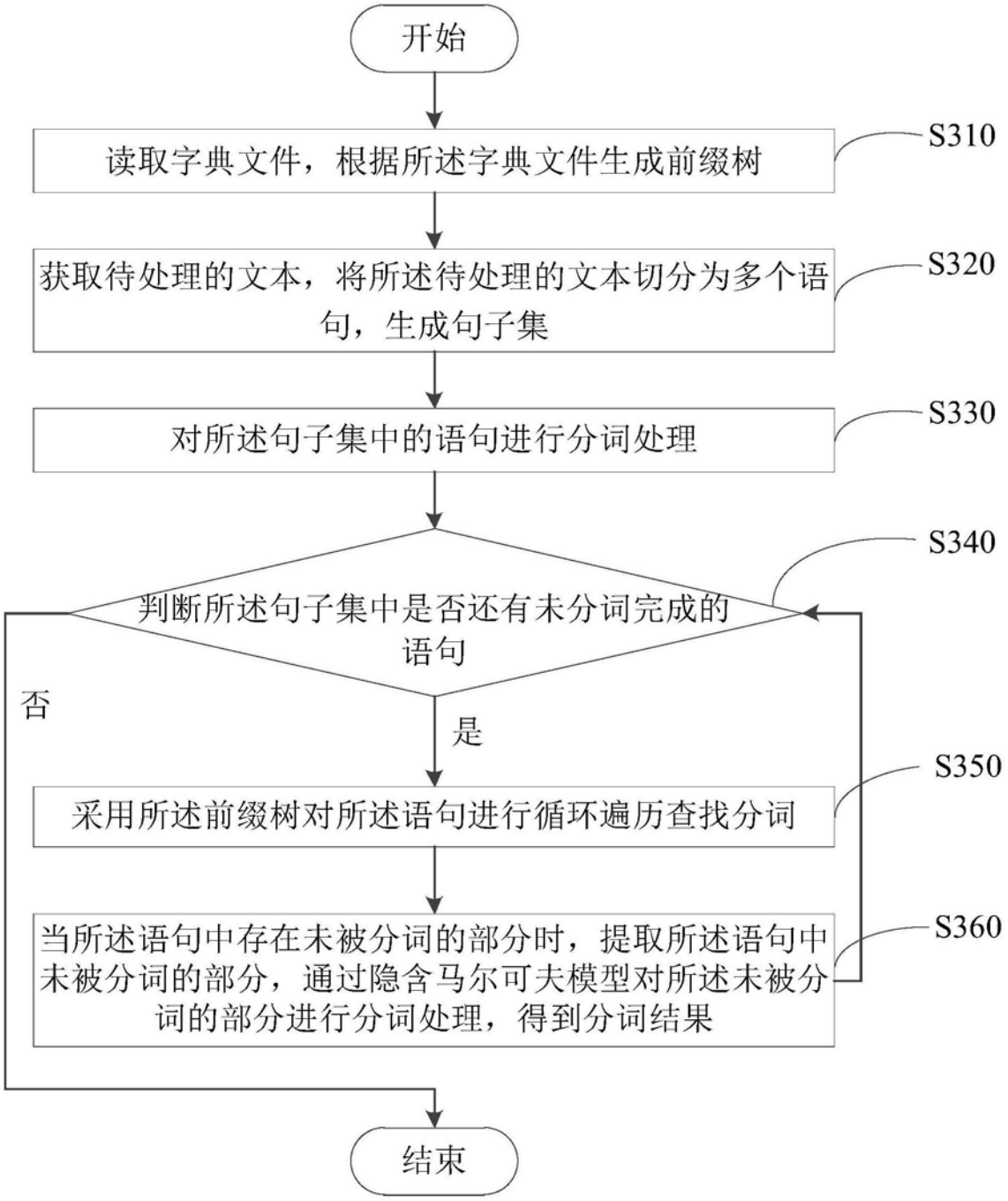

开始
读取字典文件，根据所述字典文件生成前缀树
S310
获取待处理的文本，将所述待处理的文本切分为多个语句，生成句子集
S320
对所述句子集中的语句进行分词处理
S330
判断所述句子集中是否还有未分词完成的语句
S340
否
是
采用所述前缀树对所述语句进行循环遍历查找分词
S350
当所述语句中存在未被分词的部分时，提取所述语句中未被分词的部分，通过隐含马尔可夫模型对所述未被分词的部分进行分词处理，得到分词结果
S360
结束

图9

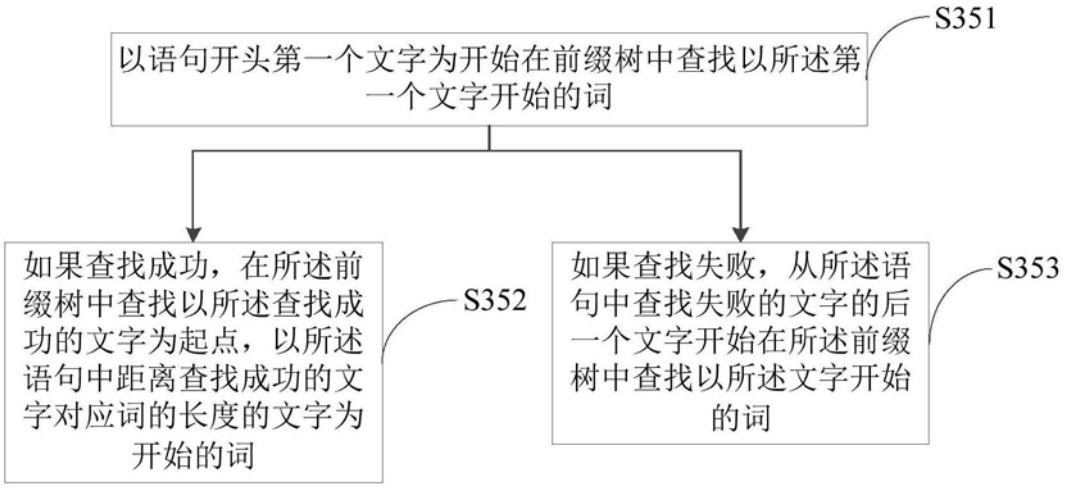

S351
以语句开头第一个文字为开始在前缀树中查找以所述第一个文字开始的词
如果查找成功，在所述前缀树中查找以所述查找成功的文字为起点，以所述语句中距离查找成功的文字对应词的长度的文字为开始的词
S352
如果查找失败，从所述语句中查找失败的文字的后一个文字开始在所述前缀树中查找以所述文字开始的词
S353

图10

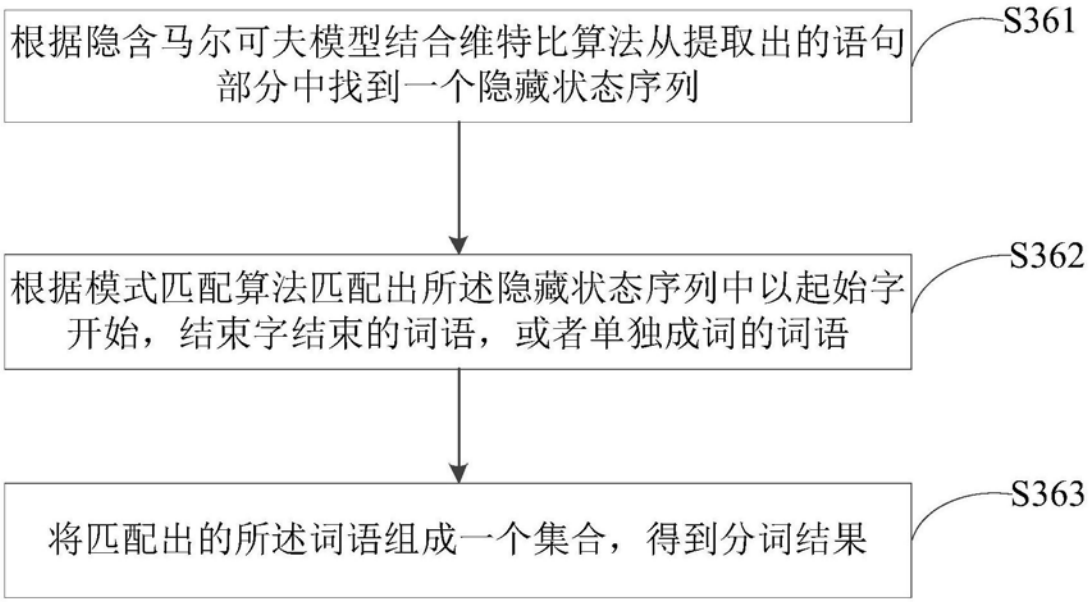

根据隐含马尔可夫模型结合维特比算法从提取出的语句部分中找到一个隐藏状态序列
S361
根据模式匹配算法匹配出所述隐藏状态序列中以起始字开始，结束字结束的词语，或者单独成词的词语
S362
将匹配出的所述词语组成一个集合，得到分词结果
S363

图11

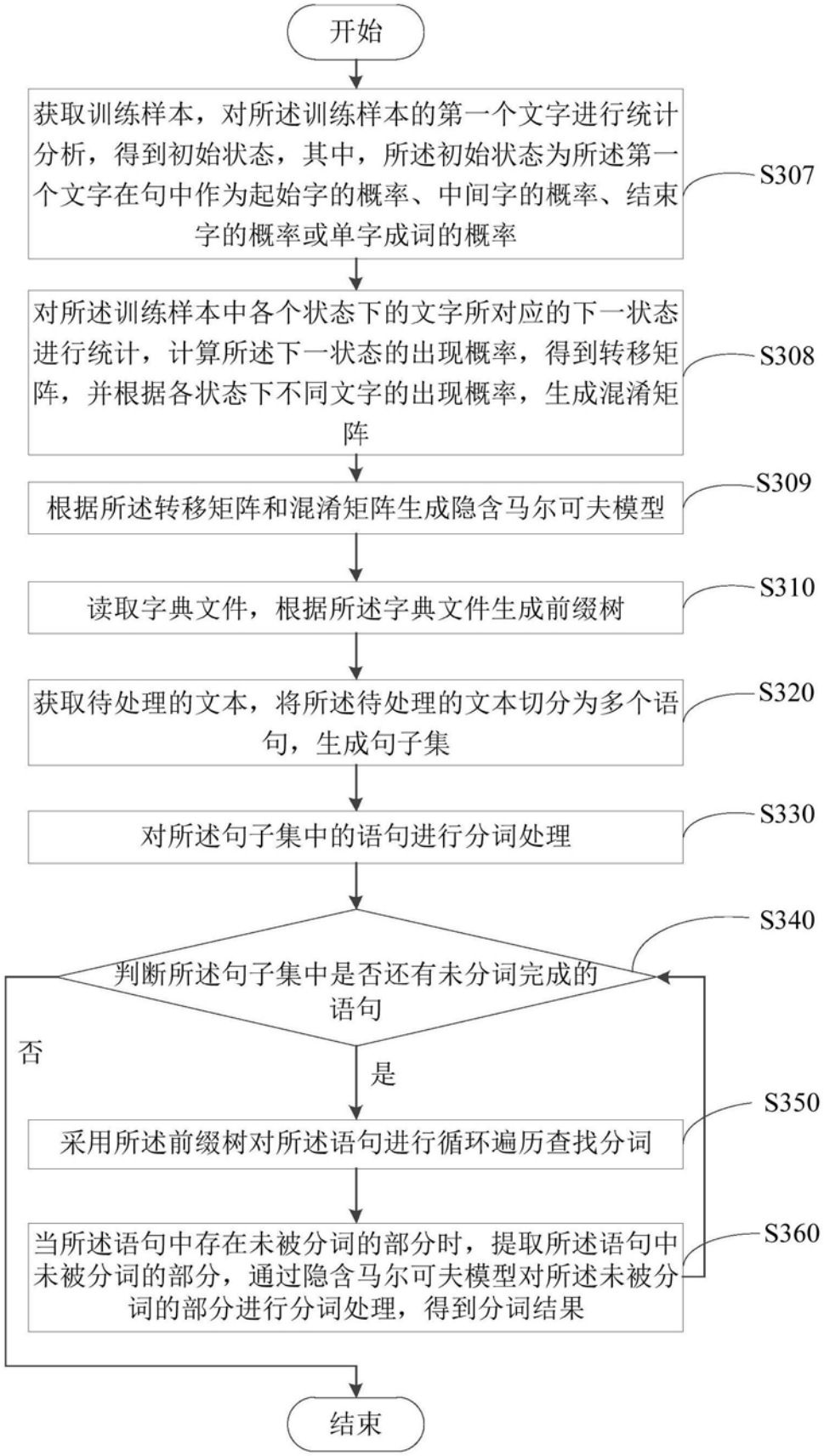

开始
获取训练样本，对所述训练样本的第一个文字进行统计分析，得到初始状态，其中，所述初始状态为所述第一个文字在句中作为起始字的概率、中间字的概率、结束字的概率或单字成词的概率
S307
对所述训练样本中各个状态下的文字所对应的下一状态进行统计，计算所述下一状态的出现概率，得到转移矩阵，并根据各状态下不同文字的出现概率，生成混淆矩阵
S308
根据所述转移矩阵和混淆矩阵生成隐含马尔可夫模型
S309
读取字典文件，根据所述字典文件生成前缀树
S310
获取待处理的文本，将所述待处理的文本切分为多个语句，生成句子集
S320
对所述句子集中的语句进行分词处理
S330
判断所述句子集中是否还有未分词完成的语句
S340
否
是
采用所述前缀树对所述语句进行循环遍历查找分词
S350
当所述语句中存在未被分词的部分时，提取所述语句中未被分词的部分，通过隐含马尔可夫模型对所述未被分词的部分进行分词处理，得到分词结果
S360
结束

图12