(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114625850 A
(43)申请公布日 2022.06.14

(21)申请号 202210226528.5

(22)申请日 2022.03.09

(71)申请人 上海弘玑信息技术有限公司
地址 201240 上海市闵行区紫星路588号2幢13层055室

(72)发明人 李维 秦海龙 林天兵 彭滢 穆啸天 刘郑勇

(74)专利代理机构 北京超凡宏宇专利代理事务所(特殊普通合伙) 11463
专利代理师 何明伦

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 40/289(2020.01)

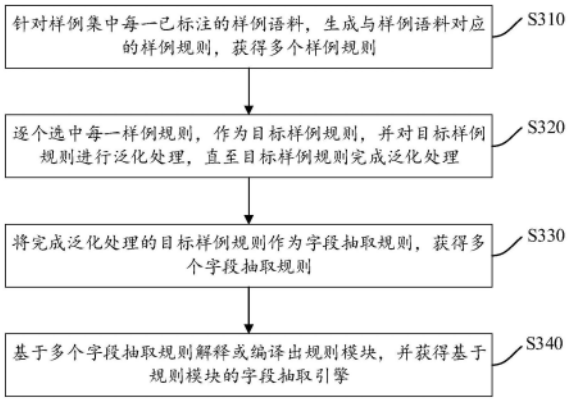
权利要求书3页 说明书17页 附图6页

(54)发明名称

字段抽取引擎的生成方法及装置、电子设备、存储介质

(57)摘要

本申请提供一种字段抽取引擎的生成方法及装置、电子设备、存储介质，方法包括：针对样例集中每一已标注的样例语料，生成与所述样例语料对应的样例规则，获得多个样例规则；逐个选中每一样例规则，作为目标样例规则，并对所述目标样例规则进行泛化处理，直至所述目标样例规则完成泛化处理；当每一样例规则均完成泛化处理，获得多个字段抽取规则；基于所述多个字段抽取规则解释或编译出规则模块，并获得基于所述规则模块的字段抽取引擎。本申请方案，为NLP信息抽取应用在各种缺乏标注数据的业务场景提供了一种快捷的解决方案，克服了主流监督学习方案依赖大规模标注数据的知识瓶颈，也大大减轻了纯手工高代码开发的时间成本。



CN 114625850 A

1.一种字段抽取引擎的生成方法,其特征在于,包括:

针对样例集中每一样例语料,生成与所述样例语料对应的样例规则,获得多个样例规则;

逐个选中每一样例规则,作为目标样例规则,并对所述目标样例规则进行泛化处理,直至所述目标样例规则完成泛化处理;

将完成泛化处理的目标样例规则作为字段抽取规则,获得多个字段抽取规则;

基于所述多个字段抽取规则解释或编译出规则模块,并获得基于所述规则模块的字段抽取引擎。

2.根据权利要求1所述的方法,其特征在于,所述生成与所述样例语料对应的样例规则,包括:

对所述样例语料进行分词处理,得到多个词节点;

基于所述样例语料中若干标注字段确定所述多个词节点中的字段左右边界,以及每一标注字段的字段标签,构造所述样例规则;

其中,所述标注字段为携带字段标签的字段,所述标注字段包括若干词节点。

3.根据权利要求1所述的方法,其特征在于,所述对所述目标样例规则进行泛化处理,直至所述目标样例规则完成泛化处理,包括:

对所述目标样例规则进行一轮泛化处理;

在一轮泛化处理后,根据泛化处理后的目标样例规则和当前规则集,分别对开发集的语料和召回测试集的语料进行字段抽取,获得抽取结果;

根据所述抽取结果判断经过一轮泛化处理的目标样例规则是否通过质量测试,并根据判断结果进行下一轮泛化处理;

当所述目标样例规则达到终止泛化条件时,确定所述目标样例规则完成所有轮次泛化处理,并将完成所有轮次泛化处理的目标样例规则放入所述当前规则集。

4.根据权利要求3所述的方法,其特征在于,在所述针对样例集中每一样例语料,生成与所述样例语料对应的样例规则之前,所述方法还包括:

从目标业务领域的原始数据源获取第一指定数量的多个字符串,作为语料构建所述开发集;

从所述原始数据源获取第二指定数量的多个字符串,作为语料构建所述召回测试集;

从所述开发集中选取第三指定数量的多个语料,并对选择的多个语料进行标注,获得样例集。

5.根据权利要求3所述的方法,其特征在于,所述抽取结果包括所述开发集对应的第一抽取结果,以及所述召回测试集对应的第二抽取结果;

根据所述抽取结果判断经过一轮泛化处理的目标样例规则是否通过质量测试,所述方法还包括:

响应于比对指令,确定所述第一抽取结果与所述开发集的基准抽取结果之间的差异槽位信息点;

当所述差异槽位信息点的精确率满足预设精确率阈值,判断所述第二抽取结果中槽位信息点数量,超出所述召回测试集的基准槽位信息点数量的比例是否超过比例阈值;

若是,确定经过一轮泛化处理的目标样例规则通过质量测试,否则,确定经过一轮泛化

处理的目标样例规则未通过质量测试。

6.根据权利要求5所述的方法，其特征在于，在所述确定经过一轮泛化处理的目标样例规则通过质量测试之后，所述方法还包括：

将所述第一抽取结果，作为所述开发集新的基准抽取结果；

将所述第二抽取结果中槽位信息点数量，作为所述召回测试集新的基准槽位信息点数量。

7.根据权利要求3所述的方法，其特征在于，所述根据判断结果进行下一轮泛化处理，包括：

当经过一轮泛化处理的目标样例规则通过质量测试时，对该轮泛化处理后的目标样例规则，进行下一轮泛化处理；

当经过一轮泛化处理的目标样例规则未通过质量测试时，对该轮泛化处理前的目标样例规则，进行下一轮泛化处理。

8.根据权利要求3所述的方法，其特征在于，所述对所述目标样例规则进行泛化处理，包括：

交替选择上下文泛化和词节点泛化作为每一轮泛化处理的路径，对所述目标样例规则进行多轮泛化处理。

9.根据权利要求8所述的方法，其特征在于，所述对所述目标样例规则进行一轮泛化处理，包括：

当通过上下文泛化作为泛化处理的路径，从所述目标样例规则的开头和/或结尾删除与标注字段不相交的若干词节点。

10.根据权利要求8所述的方法，其特征在于，所述对所述目标样例规则进行一轮泛化处理，包括：

当通过上下文泛化作为泛化处理的路径，将所述目标样例规则的多个词节点中若干非关键词节点，替换为最大可选项，并更新所述最大可选项；其中，所述非关键词节点未携带字段标签，所述最大可选项表示被替换的非关键词节点位置可允许的词节点的最大数量。

11.根据权利要求8所述的方法，其特征在于，所述对所述目标样例规则进行一轮泛化处理，包括：

当通过上下文泛化作为泛化处理的路径，将所述目标样例规则拆分为若干子样例规则。

12.根据权利要求8所述的方法，其特征在于，所述对所述目标样例规则进行一轮泛化处理，包括：

当通过词节点泛化作为泛化处理的路径，对所述目标样例规则的多个词节点中若干关键词节点，进行同义词拓展；其中，所述关键词节点携带字段标签。

13.根据权利要求8所述的方法，其特征在于，所述对所述目标样例规则进行一轮泛化处理，包括：

当通过词节点泛化作为泛化处理的路径，将所述词节点中若干关键词节点替换为其对应的本体入口特征，并依据本体上下位链条指示的泛化顺序，对所述本体入口特征进行更新。

14.根据权利要求3所述的方法，其特征在于，所述根据泛化处理后的目标样例规则和

当前规则集，分别对开发集的语料和召回测试集的语料进行字段抽取，包括：

对泛化处理后的目标样例规则和当前规则集，解释或编译出规则模块，并获得基于所述规则模块的字段抽取引擎；

通过所述字段抽取引擎分别对开发集的语料和召回测试集的语料进行字段抽取。

15.一种电子设备，其特征在于，所述电子设备包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为执行权利要求1-14任意一项所述的字段抽取引擎的生成方法。

字段抽取引擎的生成方法及装置、电子设备、存储介质

技术领域

[0001] 本申请涉及自然语言处理技术领域，特别涉及一种字段抽取引擎的生成方法及装置、电子设备、计算机可读存储介质。

背景技术

[0002] 自然语言处理(Natural Language Processing，NLP)系统包含两大类：一类是机器学习系统，另一类是传统的规则系统。自然语言处理系统可以应用于在多个领域的信息抽取任务。例如，智能助理对话系统的一项关键任务是问句理解，包括识别问句的意图(intent)以及抽取问句中的相关角色槽位(role slots)。从信息抽取的角度来看，抽取角色槽位就是字段抽取，识别意图就是问句分类。示例性的，对于订票方面的问句，智能助理可从问句中抽取如下槽位信息点(也就是技能开发中所称的“角色槽位”)：时间(time)、出发地(FromLocation)、目的地(ToLocation)、交通方式(VehicleType)等；可识别的意图包括：订票(Booking)、退票(Canceling)等。对于音乐方面的问句，智能助理可从问句中抽取如下槽位信息点：歌唱家(Singer)、歌名(Song)、音乐类型(MusicType)等；可识别的意图包括：播放(PlayMusic)、停止(Stop)等。

[0003] 字段抽取包括两个子任务，一是确定字段的左右边界，二是为字段赋予一个标签(例如角色槽位)。为应对字段抽取任务，通常可组织或外包足量的标注任务，然后用监督学习的算法训练机器学习模型。标注工作在定义标注规范、培训标注人员、手工标注、标注质量控制等环节需耗费大量人力成本和时间成本。而实际应用场景中，涉及字段抽取的场景非常多，单就智能助理来看，问句包括问天气、问音乐、问股票、问时间、问地点等成千上万场景，针对每一场景必须单独标注数据，需要投入大量资源。

[0004] 如果通过规则系统处理字段抽取任务，对于每一个需要抽取的字段，也需要手工编写大量规则代码才能实现，同样需要耗费大量人力成本和时间成本。

发明内容

[0005] 本申请实施例的目的在于提供一种字段抽取引擎的生成方法及装置、电子设备、计算机可读存储介质，用于在较低人力成本和时间成本的前提下，生成执行字段抽取任务的字段抽取引擎。

[0006] 一方面，本申请方案提供了一种字段抽取引擎的生成方法，包括：

[0007] 针对样例集中每一样例语料，生成与所述样例语料对应的样例规则，获得多个样例规则；

[0008] 逐个选中每一样例规则，作为目标样例规则，并对所述目标样例规则进行泛化处理，直至所述目标样例规则完成泛化处理；

[0009] 将完成泛化处理的目标样例规则作为字段抽取规则，获得多个字段抽取规则；

[0010] 基于所述多个字段抽取规则解释或编译出规则模块，并获得基于所述规则模块的字段抽取引擎。

[0011] 在一实施例中，所述生成与所述样例语料对应的样例规则，包括：

[0012] 对所述样例语料进行分词处理，得到多个词节点；

[0013] 基于所述样例语料中若干标注字段确定所述多个词节点中的字段左右边界，以及每一标注字段的字段标签，构造所述样例规则；

[0014] 其中，所述标注字段为携带字段标签的字段，所述标注字段包括若干词节点。

[0015] 在一实施例中，所述对所述目标样例规则进行泛化处理，直至所述目标样例规则完成泛化处理，包括：

[0016] 对所述目标样例规则进行一轮泛化处理；

[0017] 在一轮泛化处理后，根据泛化处理后的目标样例规则和当前规则集，分别对开发集的语料和召回测试集的语料进行字段抽取，获得抽取结果；

[0018] 根据所述抽取结果判断经过一轮泛化处理的目标样例规则是否通过质量测试，并根据判断结果进行下一轮泛化处理；

[0019] 当所述目标样例规则达到终止泛化条件时，确定所述目标样例规则完成所有轮次泛化处理，并将完成所有轮次泛化处理的目标样例规则放入所述当前规则集。

[0020] 在一实施例中，在所述针对样例集中每一样例语料，生成与所述样例语料对应的样例规则之前，所述方法还包括：

[0021] 从目标业务领域的原始数据源获取第一指定数量的多个字符串，作为语料构建所述开发集；

[0022] 从所述原始数据源获取第二指定数量的多个字符串，作为语料构建所述召回测试集；

[0023] 从所述开发集中选取第三指定数量的多个语料，并对选择的多个语料进行标注，获得样例集。

[0024] 在一实施例中，所述抽取结果包括所述开发集对应的第一抽取结果，以及所述召回测试集对应的第二抽取结果；

[0025] 根据所述抽取结果判断经过一轮泛化处理的目标样例规则是否通过质量测试，所述方法还包括：

[0026] 响应于比对指令，确定所述第一抽取结果与所述开发集的基准抽取结果之间的差异槽位信息点；

[0027] 当所述差异槽位信息点的精确率满足预设精确率阈值，判断所述第二抽取结果中槽位信息点数量，超出所述召回测试集的基准槽位信息点数量的比例是否超过比例阈值；

[0028] 若是，确定经过一轮泛化处理的目标样例规则通过质量测试，否则，确定经过一轮泛化处理的目标样例规则未通过质量测试。

[0029] 在一实施例中，在所述确定经过一轮泛化处理的目标样例规则通过质量测试之后，所述方法还包括：

[0030] 将所述第一抽取结果，作为所述开发集新的基准抽取结果；

[0031] 将所述第二抽取结果中槽位信息点数量，作为所述召回测试集新的基准槽位信息点数量。

[0032] 在一实施例中，所述根据判断结果进行下一轮泛化处理，包括：

[0033] 当经过一轮泛化处理的目标样例规则通过质量测试时，对该轮泛化处理后的目标

样例规则，进行下一轮泛化处理；

[0034] 当经过一轮泛化处理的目标样例规则未通过质量测试时，对该轮泛化处理前的目标样例规则，进行下一轮泛化处理。

[0035] 在一实施例中，所述对所述目标样例规则进行泛化处理，包括：

[0036] 交替选择上下文泛化和词节点泛化作为每一轮泛化处理的路径，对所述目标样例规则进行多轮泛化处理。

[0037] 在一实施例中，所述对所述目标样例规则进行一轮泛化处理，包括：

[0038] 当通过上下文泛化作为泛化处理的路径，从所述目标样例规则的开头和/或结尾删除与标注字段不相交的若干词节点。

[0039] 在一实施例中，所述对所述目标样例规则进行一轮泛化处理，包括：

[0040] 当通过上下文泛化作为泛化处理的路径，将所述目标样例规则的多个词节点中若干非关键词节点，替换为最大可选项，并更新所述最大可选项；其中，所述非关键词节点未携带字段标签，所述最大可选项表示被替换的非关键词节点位置可允许的词节点的最大数量。

[0041] 在一实施例中，所述对所述目标样例规则进行一轮泛化处理，包括：

[0042] 当通过上下文泛化作为泛化处理的路径，将所述目标样例规则拆分为若干子样例规则。

[0043] 在一实施例中，所述对所述目标样例规则进行一轮泛化处理，包括：

[0044] 当通过词节点泛化作为泛化处理的路径，对所述目标样例规则的多个词节点中若干关键词节点，进行同义词拓展；其中，所述关键词节点携带字段标签。

[0045] 在一实施例中，所述对所述目标样例规则进行一轮泛化处理，包括：

[0046] 当通过词节点泛化作为泛化处理的路径，将所述词节点中若干关键词节点替换为其对应的本体入口特征，并依据本体上下位链条指示的泛化顺序，对所述本体入口特征进行更新。

[0047] 在一实施例中，所述根据泛化处理后的目标样例规则和当前规则集，分别对开发集的语料和召回测试集的语料进行字段抽取，包括：

[0048] 对泛化处理后的目标样例规则和当前规则集，解释或编译出规则模块，并获得基于所述规则模块的字段抽取引擎；

[0049] 通过所述字段抽取引擎分别对开发集的语料和召回测试集的语料进行字段抽取。

[0050] 另一方面，本申请方案提供了一种字段抽取引擎的生成装置，包括：

[0051] 生成模块，用于针对样例集中每一已标注的样例语料，生成与所述样例语料对应的样例规则，获得多个样例规则；

[0052] 泛化模块，用于逐个选中每一样例规则，作为目标样例规则，并对所述目标样例规则进行泛化处理，直至所述目标样例规则完成泛化处理；

[0053] 获取模块，用于将完成泛化处理的目标样例规则作为字段抽取规则，获得多个字段抽取规则；

[0054] 编译模块，用于基于所述多个字段抽取规则编译或解释得到字段抽取引擎，并获得基于所述规则模块的字段抽取引擎。

[0055] 进一步的，本申请提供了一种电子设备，所述电子设备包括：

[0056] 处理器；

[0057] 用于存储处理器可执行指令的存储器；

[0058] 其中，所述处理器被配置为执行上述字段抽取引擎的生成方法。

[0059] 此外，本申请还提供了一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序可由处理器执行以完成上述字段抽取引擎的生成方法。

[0060] 本申请方案，在少量已标注的样例语料的基础上，自动生成样例规则，并逐个样例规则进行泛化处理，以得到由样例集而来的多个字段抽取规则，对字段抽取规则编译或解释后，可以得到用于字段抽取任务的字段抽取引擎；样例语料的数量较少，但泛化能力强，因此，标注的人力成本和时间成本较低；样例驱动的自动规则生成，有效避免手工代码可能产生的错误，而样例规则的泛化迭代，可以覆盖语言表层现象的很多变体，进一步降低传统规则代码开发需要众多规则的工作量；本方案为NLP信息抽取应用在各种缺乏标注数据的业务场景提供了一种快捷的解决方案，克服了主流监督学习方案依赖大规模标注数据的知识瓶颈，也大大减轻了纯手工高代码开发的时间成本。

附图说明

[0061] 为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例中所需要使用的附图作简单地介绍。

[0062] 图1为本申请一实施例提供的电子设备的结构示意图；

[0063] 图2为本申请一实施例提供的字段抽取引擎的生成方法的流程示意图；

[0064] 图3为本申请一实施例提供的NLP平台的架构示意图；

[0065] 图4为本申请一实施例提供的生成语料集的示意图；

[0066] 图5为图2对应实施例中步骤S320的细节流程示意图；

[0067] 图6为本申请一实施例提供的质量测试通过条件的判断方法的流程示意图；

[0068] 图7为本申请一实施例提供的基准参数的构建方式示意图；

[0069] 图8为本申请一实施例提供的样例规则泛化处理的整体流程示意图；

[0070] 图9为本申请一实施例提供的字段抽取引擎的生成装置的框图。

具体实施方式

[0071] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行描述。

[0072] 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。同时，在本申请的描述中，术语“第一”、“第二”等仅用于区分描述，而不能理解为指示或暗示相对重要性。

[0073] 图1为本申请一实施例提供的电子设备的结构示意图。该电子设备200可以用于执行本申请实施例提供的字段抽取引擎的生成方法。如图1所示，该电子设备200包括：一个或多个处理器202、一个或多个存储处理器可执行指令的存储器204。其中，所述处理器202被配置为执行本申请下述实施例提供的字段抽取引擎的生成方法。

[0074] 所述处理器202可以是包含中央处理单元(CPU)、图像处理单元(GPU)或者具有数据处理能力和/或指令执行能力的其它形式的处理单元的设备，可以对所述电子设备200中的其它组件的数据进行处理，还可以控制所述电子设备200中的其它组件以执行期望的功

能。

[0075] 所述存储器204可以包括一个或多个计算机程序产品，所述计算机程序产品可以包括各种形式的计算机可读存储介质，例如易失性存储器和/或非易失性存储器。所述易失性存储器例如可以包括随机存取存储器(RAM)和/或高速缓冲存储器(cache)等。所述非易失性存储器例如可以包括只读存储器(ROM)、硬盘、闪存等。在所述计算机可读存储介质上可以存储一个或多个计算机程序指令，处理器202可以运行所述程序指令，以实现下文所述的字段抽取引擎的生成方法。在所述计算机可读存储介质中还可以存储各种应用程序和各种数据，例如所述应用程序使用和/或产生的各种数据等。

[0076] 在一实施例中，图1所示电子设备200还可以包括输入装置206、输出装置208以及数据采集装置210，这些组件通过总线系统212和/或其它形式的连接机构(未示出)互连。应当注意，图1所示的电子设备200的组件和结构只是示例性的，而非限制性的，根据需要，所述电子设备200也可以具有其他组件和结构。

[0077] 所述输入装置206可以是用户用来输入指令的装置，并且可以包括键盘、鼠标、麦克风和触摸屏等中的一个或多个。所述输出装置208可以向外部(例如，用户)输出各种信息(例如，图像或声音)，并且可以包括显示器、扬声器等中的一个或多个。所述数据采集装置210可以采集对象的图像，并且将所采集的图像存储在所述存储器204中以供其它组件使用。示例性地，该数据采集装置210可以为摄像头。

[0078] 在一实施例中，用于实现本申请实施例的字段抽取引擎的生成方法的示例电子设备200中的各器件可以集成设置，也可以分散设置，诸如将处理器202、存储器204、输入装置206和输出装置208集成设置于一体，而将数据采集装置210分离设置。

[0079] 在一实施例中，用于实现本申请实施例的字段抽取引擎的生成方法的示例电子设备200可以被实现为诸如笔记本电脑、台式电脑、服务器等智能设备。

[0080] 参见图2，为本申请一实施例提供的字段抽取引擎的生成方法的流程示意图。该方法可以由上述电子设备200执行，如图2所示，该方法包括以下步骤S310-步骤S340。

[0081] 步骤S310：针对样例集中每一样例语料，生成与样例语料对应的样例规则，获得多个样例规则。

[0082] 其中，样例集包括多个样例语料，样例集中的语料可以从需要应用字段提取引擎的业务领域的原始文本数据中提取得到，且语料经过标注得到样例语料。这里，业务领域可以包括交通服务领域、音乐服务领域、法律领域、金融领域等。样例集中样例语料的数量通常较少，大约为100至500条。

[0083] 针对每一样例语料，通过直接量全匹配的方式，可以自动获得与样例语料对应的样例规则。样例规则指示从样例语料中抽取槽位信息点的方式，此时，样例规则不具备概括性，仅能涵盖其对应的样例语料。

[0084] 步骤S320：逐个选中每一样例规则，作为目标样例规则，并对目标样例规则进行泛化处理，直至目标样例规则完成泛化处理。

[0085] 其中，目标样例规则是当前被选中，进行泛化处理的样例规则。泛化处理是对样例规则进行调整，放宽规则模式的条件约束，使得经过调整的样例规则可以适用于更多语料的处理过程。

[0086] 在逐个选中样例规则进行泛化处理的过程中，当前进行泛化处理的目标样例规则

的泛化效果受已完成泛化的样例规则的影响。当目标样例规则的泛化效果无法优化时，例如进一步放宽规则模式约束条件无法通过质量测试，则可认为目标样例规则完成泛化，此时可将该目标样例规则作为泛化完成的最终字段抽取规则。

[0087] 当任一目标样例规则完成泛化处理后，则可以继续选择下一条样例规则作为目标样例规则，从而对新的目标样例规则进行泛化处理。

[0088] 步骤S330：将完成泛化处理的目标样例规则作为字段抽取规则，获得多个字段抽取规则。

[0089] 步骤S340：基于多个字段抽取规则编译或解释出规则模块，并获得基于规则模块的字段抽取引擎。

[0090] 在获得多个字段抽取规则之后，可以根据NLP专用的语言规范对多个字段抽取规则进行编译或解释。这里，编译(compile)是将源代码一次性转换成目标代码的过程；解释(interpret)是将源代码逐条转换成目标代码同时逐条运行的过程。其中，NLP专用的语言的核心为有限状态机制(Finite State formalism)。本申请可以通过NLP平台对字段抽取规则进行编译或解释，从而得到可执行的规则模块，此时，获得基于规则模块的字段抽取引擎。在字段抽取引擎中，多个字段抽取规则以逻辑或的形式存在，换而言之，字段抽取引擎可以依据任意一条字段抽取规则从语料中抽取槽位信息点。

[0091] 参见图3，为本申请一实施例提供的NLP平台的架构示意图，如图3所示，NLP平台可以包括规则编译器，该规则编译器把规则集编译成可执行的规则模块，这些规则模块是字段抽取引擎(NLP-Tagger)的核心。引擎中，通用词典和领域词典是NLP平台运行的必要资源，可使得抽取引擎能够对语料进行分词处理，并依据分词得到的若干词节点以及每一词节点的词典特征信息。NLP平台的赋能抽取模块确定若干字段(一个字段可以由若干词节点组成)，并依据字段内各词节点的词典特征信息为上述字段标注字段标签，该字段标签表示该字段为槽位信息点。

[0092] 通过上述措施，可以基于少量已标注的样例语料，生成样例规则，并泛化出多个字段抽取规则，并以字段抽取规则编译或解释得到字段抽取引擎；这种情况下，无需投入大量人力成本和时间成本执行标注工作，且可以基于抽取样例自动生成字段抽取规则，节省了纯人工编码所带来的人力成本和时间成本，避免了人工代码中可能出现的错误。

[0093] 在一实施例中，在执行上述步骤S310至步骤S340的方法之前，可以构建本申请方案所需的语料集。其中，语料集可以包括开发集、召回测试集和样例集。

[0094] 可以从目标业务领域的原始数据源获取第一指定数量的多个字符串，作为语料构建开发集。这里，目标业务领域为字段抽取引擎所应用的业务领域。第一指定数量可以是预配置的经验值，示例性的，可以为1000到5000之间的数值。

[0095] 从原始数据源获取第二指定数量的多个字符串，作为语料构建所述召回测试集。第二指定数量可以是预配置的经验值，示例性的，可以为10000到50000之间的数值。

[0096] 从开发集中选取第三指定数量的多个语料，并对选择的多个语料进行标注，获得样例集。第三指定数量为第一指定数量的十分之一左右。参见图4，为本申请一实施例提供的生成语料集的示意图，如图4所示，可以从业务领域的原始数据源随机获取与业务相关的字符串(句子)构成开发集DevCorpus，开发集中语料大约为1000至5000条；从原始数据源随机获取更多字符串构成召回测试集RecallCorpus，召回测试集中语料数量为开发集中语料

数量的十倍左右。开发集对于开发人员是开放的，而召回测试集则是不开放的。由于召回测试集中语料未经过标注，本方案以抽取规则击中召回测试集中数据的数量作为召回指标。该召回指标体现了相对召回的依据，一般而言，规则击中语料的数量与召回率成正比。

[0097] 原始数据源作为语料的原始数据未经人工标注。对于不同业务领域，原始数据源可以不同。一些业务领域的原始数据是开源的，比如，法律领域的法律判决书历史档案、问答系统领域的问句集；一些业务领域的原始数据是客户提供的，比如，客服服务领域的客户沟通记录；一些领域的原始数据可以是网络爬虫定向收集得到。在构建语料集时，可以通过关键词收集具有针对性的代表性语料。示例性的，对于交通服务领域，关键词可以是“订票”、“预订”、“飞机票”、“高铁票”、“直达”等。另外，在构建开发集时，为增强后续所生成的字段提取引擎的鲁棒性，可以选择一部分与业务领域无关的语料。

[0098] 在构建出开发集后，可以从开发集中选取语料，以构建初始样例集SampleCorpus。一种情况下，可以从开发集中随机选择语料；另一种情况下，可以以业务领域的关键词，选取针对性语料。

[0099] 在获得初始样例集之后，可以对初始样例集中每一语料进行人工标注，从而为语料的若干字段添加字段标签，从而得到经过标注的样例语料。添加字段标签的字段属于槽位信息点。比如，交通服务领域订票场景的字段标签可以包括：时间(Time)、出发地(FromLocation)、目的地(ToLocation)、交通方式(VehicleType)等。样例语料的标注结果可以示例如下：

[0100] 我想订一张[明天下午:Time]从[南京:FromLocation]到[北京:ToLocation]的[飞机票:VehicleType]

[0101] 初始样例集中每一语料进行标注后，获得多个样例语料，构成样例集TaggedSampleCorpus。

[0102] 在一实施例中，生成与样例语料对应的样例规则可以包括：根据样例语料的若干标注字段的字段标签，自动构造与样例语料对应的样例规则。其中，标注字段为携带字段标签的字段。

[0103] 针对每一样例语料，可以对该样例语料进行分词处理，得到多个词节点。基于标注字段确定多个词节点中的字段左右边界，以及标注字段的字段标签，可以通过直接量全匹配的方式，自动生成样例规则。

[0104] 示例性的，样例语料：我想订一张[明天下午:Time]从[南京:FromLocation]到[北京:ToLocation]的[飞机票:VehicleType]

[0105] 该样例语料分词后所对应的样例规则的伪码(pseudo-code)可以表示为：[我][想][订][一][张]<[明天][下午]:Time>[从]<[南京]:FromLocation>[到]<[北京]:ToLocation>[的]<[飞机票]:VehicleType>

[0106] 这里，样例语料是由13个词组成的句子，样例规则的伪码就以与13个词一一对应的直接量(即词节点)来表示，形式为方括号。而字段常常不是单个词节点，可能包含两个或更多词，为确定字段的左右边界，可以用尖括号的伪码形式来表示。任一标注字段对应的抽取模式，包括条件和结论两部分，表示为<条件:结论>，它们是规则的一部分。后续抽取字段时，当条件满足时，字段标签作为匹配成功的结论，可以放在冒号后面。样例规则的生成过程中，非标注字段的直接量是标注字段抽取模式的上下文条件，作为直接量全匹配规则的

一部分。

[0107] 本方案字段标注的设计与规则形式(formalism)同形,基本一一对应,这为从已标注的样例语料自动生成可编译或解释的规则提供了良好的条件,也增加了代码的可读性和可调试性。上述表示方法采用有限状态形式机制(finite state formalism),本方案形成的规则也可近似等价地表示为类似的有限状态的形式机制,例如施乐FST(Finite State Transducer)机制。

[0108] 在一实施例中,如图5所示,上述步骤S320具体包括以下步骤S321-步骤S324。

[0109] 步骤S321:对目标样例规则进行一轮泛化处理。

[0110] 在选中任一样例规则作为目标样例规则之后,对该样例规则进行一轮泛化处理。泛化处理可以包括多种处理策略,在执行一轮泛化处理前,可由开发人员选择一种处理策略,从而以该处理策略执行本轮对目标样例规则的泛化。

[0111] 步骤S322:在一轮泛化处理后,根据泛化处理后的目标样例规则和当前规则集,分别对开发集的语料和召回测试集的语料进行字段抽取,获得抽取结果。

[0112] 其中,当前规则集可以包括若干现存字段抽取规则,现存字段抽取规则为已经完成所有轮次泛化处理的样例规则。在初始情况下,当前规则集为空。

[0113] 在目标样例规则经过一轮泛化处理之后,可以依据经过泛化处理的目标样例规则和若干现存字段抽取规则,经过编译或解释,分别对开发集和召回测试集的语料进行字段抽取。对于任一一条语料而言,在匹配到泛化处理后的目标样例规则或任一现存字段抽取规则后,可以基于匹配到的规则提取出若干字段,作为抽取结果。

[0114] 对于第一条样例规则Rule(1),Rule(1)每一轮泛化处理后,以经过泛化处理Rule(1)进行字段抽取。对于第n条样例规则Rule(n),Rule(n)每一轮泛化处理后,以经过泛化处理的Rule(n),以及当前规则集中现存已完成泛化处理的Rule(1)、Rule(2)、Rule(3)……Rule(n-2)、Rule(n-1)作为统一的规则集进行编译或解释来执行字段抽取。

[0115] 上述抽取结果可以评估本轮对目标样例规则的泛化效果。

[0116] 步骤S323:根据抽取结果判断经过一轮泛化处理的目标样例规则是否通过质量测试,并根据判断结果进行下一轮泛化处理。

[0117] 步骤S324:当目标样例规则达到终止泛化条件时,确定目标样例规则完成所有轮次泛化处理,并将完成所有轮次泛化处理的目标样例规则放入当前规则集。

[0118] 其中,质量测试用于测试目标样例规则在这一轮泛化处理的泛化效果。

[0119] 下一轮泛化处理,可以针对经过泛化处理后的目标样例规则,或者,针对回滚到本轮泛化前的目标样例规则。

[0120] 终止泛化条件可以为接收到终止泛化目标样例规则的操作指令,或者,终止泛化条件可以为目标样例规则的泛化轮次数量达到预设数量上限,该数量上限可以是经验值(比如:20)。需要说明的是,目标样例规则经过一轮泛化处理后,无论后续是在该轮泛化后的目标样例规则基础上继续泛化,还是以回滚到本轮泛化前的目标样例规则基础上继续泛化,本轮泛化均计入泛化轮次数量。

[0121] 一种情况下,如果以操作指令作为终止泛化条件,可由开发者可以根据抽取结果确定目标样例规则是否已经完成泛化,从而下发相应的操作指令。在操作指令指示目标样例规则已经完成泛化的情况下,操作指令还可指示选择本轮经过泛化处理的目标样例规

则，或，上一轮经过泛化处理的目标样例规则，为完成所有轮次泛化处理的目标样例规则。

[0122] 另一种情况下，如果以泛化轮次数量作为终止泛化条件的依据，在对目标样例规则进行泛化过程中，每一轮泛化结束后，可以判断目标样例规则经过的泛化处理轮次是否达到数量上限。一方面，若未达到，可继续对目标样例规则进行泛化处理。另一方面，若达到，说明达到终止泛化条件，此时，可根据本轮泛化对应的抽取结果，选择本轮经过泛化处理的目标样例规则，或，上一轮经过泛化处理的目标样例规则，为完成所有轮次泛化处理的目标样例规则。

[0123] 完成泛化处理的目标样例规则可以作为现存字段抽取规则，放入当前规则集，从而更新当前规则集。

[0124] 通过上述措施，对目标样例规则进行泛化处理的过程中，在已完成所有轮次泛化处理的样例规则的基础上，对开发集和召回测试集的语料进行字段抽取，并以抽取结果评估每一轮的泛化效果，从而在泛化效果无法继续优化时，确定目标样例规则完成泛化处理。

[0125] 在一实施例中，目标样例规则经过一轮泛化处理，对开发集的语料和召回测试集的语料进行字段抽取时，可以对泛化处理后的目标样例规则和当前规则集，编译或解释出规则模块，从而得到基于该规则模块的字段抽取引擎。

[0126] 通过NLP平台的规则编译器，在通用词典、领域词典的基础上，对完成本轮泛化处理后的目标样例规则和当前规则集中的现存字段抽取规则统一进行编译或解释出规则模块，获得基于该规则模块的字段抽取引擎。在编译或解释得到的字段抽取引擎中，经过泛化处理的目标样例规则和各字段抽取规则，以逻辑或的形式存在。

[0127] 在获得字段抽取引擎之后，可以通过字段抽取引擎分别对开发集的语料和召回测试集的语料进行字段抽取，从而得到抽取结果。

[0128] 每一轮泛化处理后，由于目标样例规则已经发生变化，因此，需要重新编译或解释字段抽取引擎，以根据重新编译或解释的字段抽取引擎执行字段抽取任务。

[0129] 在一实施例中，如图6所示，为本申请一实施例提供的质量测试通过条件的判断方法的流程示意图，可以通过如下步骤S610-步骤S630判断抽取结果是否满足质量测试的通过条件，其中，抽取结果包括开发集对应的第一抽取结果，以及召回测试集对应的第二抽取结果。

[0130] 步骤S610：响应于比对指令，确定第一抽取结果与开发集的基准抽取结果之间的差异槽位信息点。

[0131] 其中，基准抽取结果为前一轮通过质量测试的泛化处理后，从开发集的语料中抽取出的字段。基准抽取结果用于与当前第一抽取结果作比对，在初始情况下，基准抽取结果为空。

[0132] 比对指令可以通过diff工具下发，用于确定并展示第一抽取结果与基准抽取结果之间的差异槽位信息点。对于第一抽取结果相对于基准抽取结果新增的差异槽位信息点，diff工具可以以高亮的形式突出展示，从而便于开发人员检查差异槽位信息点的精确率是否满足精确率阈值(比如：95%-100%)。

[0133] 精确率＝抽取出的正确槽位信息点数量/抽取出的槽位信息点数量，而由于开发集中语料并未经过标注，在抽取出多个槽位信息点后，无法自动计算出精确率。这种情况下，只要保证每一轮泛化处理后新增的差异槽位信息点的精确率满足精确率阈值，则可以

确定累积得到的基准抽取结果满足精确率的期望值。因此，在获得第一抽取结果之后，只需查验新增的差异槽位信息点是否满足精确率阈值即可。

[0134] 一方面，若新增的差异槽位信息点的精确率不满足精确率阈值，可以确定本轮的抽取结果不满足质量测试的通过条件。另一方面，若不存在新增的差异槽位信息点，则说明本轮泛化处理后，精确率不变，即没有新的抽取错误，可以认定差异槽位信息点的精确率满足精确率阈值，可以继续执行步骤S620。再一方面，若新增的差异槽位信息点的精确率满足精确率阈值，则可以继续执行步骤S620。

[0135] 步骤S620：当差异槽位信息点的精确率满足预设精确率阈值，判断第二抽取结果中槽位信息点数量，超出召回测试集的基准槽位信息点数量的比例是否超过比例阈值。

[0136] 其中，基准槽位信息点数量为前一轮通过质量测试的泛化处理后从召回测试集的语料中抽取出的作为槽位信息点的字段的数量。基准槽位信息点数量用于与当前第二抽取结果作比较，在初始情况下，基准槽位信息点数量为0。

[0137] 召回测试集对应的第二抽取结果用于评估反映规则覆盖面的召回率。召回率＝抽取出的正确槽位信息点数量/语料集中的槽位信息点总量，而由于召回测试集中语料并未经过标注，在抽取出多个槽位信息点之后，无法直接计算召回率：没有标注，分母无从知道；而且召回测试集测试属于大数据集的盲测，分子也不能靠人工查验来判断。这种情况下，可以通过与召回率存在正相关的“相对召回”的统计方法来反映槽位信息点的覆盖面，作为回归测试的近似手段。根据上述召回率的标准公式，可以知道，未标注的召回测试集虽然槽位信息点总量未知，但该数值是恒定不变的。此时，召回率决定于抽取出的正确槽位信息点的数量。在字段抽取引擎的精确率满足精确率阈值的情况下，可认为抽取出的槽位信息点数量正比于抽取出的正确槽位信息点数量，抽取出的槽位信息点数量越大，说明召回率越高。

[0138] 对于第二抽取结果，判断该第二抽取结果中槽位信息点数量，超出基准槽位信息点数量的比例，是否达到预定的比例阈值目标。这里，比例阈值可以是经验值，一般在0％到5％之间，比如，可以是3％。一方面，若达到比例阈值，说明第二抽取结果通过了回归测试的相对召回指标。另一方面，若未达到比例阈值，说明第二抽取结果未通过回归测试的相对召回指标。

[0139] 此外，初始情况下，基准槽位信息点数量为0。因此，在对第一个目标样例规则进行第一轮泛化处理后，由于基准槽位信息点数量为0，第二抽取结果超出基准槽位信息点数量的比例为100％，因此在对第一个目标样例规则进行第一轮泛化处理后，无需额外计算超出基准槽位信息点数量的比例。

[0140] 步骤S630：若是，确定经过一轮泛化处理的目标样例规则通过质量测试，否则，确定经过一轮泛化处理的目标样例规则未通过质量测试。

[0141] 当第一抽取结果与开发集的基准抽取结果之间的差异槽位信息点的精确率，满足精确率阈值，并且第二抽取结果超出召回测试集的基准槽位信息点数量的比例达到预定的比例阈值时，可以确定抽取结果在精确率和召回率上满足要求，即满足数据质量测试的通过条件。

[0142] 当第一抽取结果与开发集的基准抽取结果之间的差异槽位信息点的精确率，不满足精确率阈值，可以确定抽取结果在精确率上不满足要求。当第二抽取结果超出召回测试集的基准槽位信息点数量的比例未达到比例阈值，可以确定抽取结果在召回率上不满足要

求。在精确率和召回率上任一要求不满足时，可以确定不满足数据质量测试的通过条件。

[0143] 通过上述措施，可以在未标注的开发集和未标注的召回测试集的基础上，评估每一轮泛化处理后的泛化效果。在一实施例中，确定抽取结果满足质量测试的通过条件之后，可以将本轮的第一抽取结果，作为开发集新的基准抽取结果，以及，将第二抽取结果中槽位信息点数量，作为召回测试集槽位信息点新的基准槽位信息点数量。

[0144] 如图7所示，为本申请一实施例提供的基准参数的构建方式示意图，基准参数可以包括反映抽取结果精确率的基准抽取结果(图7中的Precision Base line)和反映相对召回率的基准槽位信息点数量(图7中的Recall Hits)。在初始情况下，基准抽取结果为空且基准信息点数量为0。对第一个样例规则进行第一轮泛化处理后，可以基于第一次得到的第一抽取结果和第二抽取结果，确定新的基准抽取结果和新的基准槽位信息点数量。在后续泛化过程中，可以根据新的第一抽取结果和第二抽取结果，不断迭代更新基准抽取结果和基准槽位信息点数量。

[0145] 在一实施例中，根据判断结果进行下一轮泛化处理时，一方面，当经过一轮泛化处理的目标样例规则通过质量测试时，可以对本轮泛化处理后的目标样例规则，继续进行下一轮的泛化处理。如前所述，质量测试的通过条件是与精确率与召回率相关的条件。

[0146] 开发集对应的抽取结果可用于精确率判断，召回测试集对应的抽取结果可用于召回率判断。

[0147] 在满足质量测试的测试通过条件时，说明经过本轮泛化的目标样例规则，相比泛化前更有助于字段抽取，通常是增加了规则的覆盖面(召回)而且保持了规则的精确率目标阈值，因此，将本轮泛化后的目标样例规则，作为下一轮泛化处理的基础。

[0148] 另一方面，当经过一轮泛化处理的目标样例规则未通过质量测试时，对该轮泛化处理前的目标样例规则，进行下一轮泛化处理。

[0149] 在不满足质量测试的测试通过条件时，说明经过本轮泛化的目标样例规则，相比泛化前并无足够大的质量改进，因此，可以回滚到本轮泛化前的目标样例规则，作为下一轮泛化处理的基础。

[0150] 在一实施例中，泛化处理的策略可以分为两类：一类为上下文泛化，另一类为词节点泛化。上下文泛化为对目标样例规则中标注字段的上下文约束条件进行逐轮放松的调整，词节点泛化为对目标样例规则中标注字段的词语的约束条件进行逐轮放松的调整。

[0151] 在对目标样例规则进行泛化处理的过程中，可以交替选择上下文泛化和词节点泛化作为每一轮泛化处理的路径，以平衡目标样例规则的多轮泛化处理。

[0152] 在选择泛化处理的策略时，可以在开发环境中让系统交替输出两类处理策略的可选项，供开发者选择，从而保持两种泛化方式的平衡利用。示例性的，上一轮对目标样例规则Rule(n)进行词节点泛化处理，且泛化后的Rule(n)对应的抽取结果满足质量测试的通过条件，在这一轮泛化处理时，只显示属于上下文泛化的若干处理策略的菜单选项。

[0153] 通过上述措施，可以保证规则泛化有序进行迭代，从而最终获得满足字段抽取质量要求的字段抽取引擎。此外，迭代泛化过程是人机耦合的，系统提供泛化的路径菜单，路径选择最终决定于开发者，开发者可以充分利用经验，选择最佳路径进行快速泛化，避免穷尽全部路径，导致用时过长。

[0154] 在一实施例中，目标样例规则包括多个词节点，示例性的，其中带有两个标注字段

(即:<标注字段>)的目标样例规则模式初始的上下文全匹配词节点序列(token list)可以表示为:

[0155] [token-1]<[token-2]>…<[token-i]…[token-j]>…[token-n]

[0156] 此时,目标样例规则中共有n个词节点,两个目标标注字段。

[0157] 当通过上下文泛化作为泛化处理的路径时,一种处理策略可以为:从目标样例规则的开头和/或结尾删除若干与抽取字段不相交的词节点。

[0158] 删除开头词节点:<[token-2]>…<[token-i]…[token-j]>…[token-n]

[0159] 删除结尾词节点:[token-1]<[token-2]>…<[token-i]…[token-j]>…[token-n-1]

[0160] 或者,处理策略可以为:将多个词节点中若干非关键词节点,替换为最大可选项,并调整更新最大可选项。其中,非关键词节点不携带字段标签,最大可选项表示被替换的非关键词节点位置可允许的词节点的最大数量。

[0161] 示例性的,替换后的目标样例规则可以表示为:

[0162] [token-1]<[token-2]>…[ ]*k…[token-j]…[token-n]

[0163] 这里,[ ]*k为最大可选项,表示最多有k个词节点(token)。该最大可选项用于为被替换的非关键词节点左右词节点之间的距离设定限制。例如:[洗][ ]*3[澡]的规则模式可以匹配出现“洗……澡”的字符串,只要“洗”和“澡”之间最多不超过3个词节点相隔。因此,该规则模式可以匹配“洗澡/洗个澡/洗一个澡”,但却不能匹配短语“洗[一][个][痛快][的]澡”。这样看来,这个距离限制还需要进一步泛化放宽。

[0164] 可见,[ ]*k中k的取值决定了两个词节点之间的距离限制泛化的上下位链条:

[0165] [ ]*1→[ ]*2→[ ]*3→[ ]*4→……→[ ]*

[0166] 这里,[ ]*1等价于正则表达式中常用的问号表示法[ ]?,表示被替换的非关键词节点可有可无;[ ]*表示不限制词节点数量。

[0167] 通过该处理策略进行泛化处理时,首先将非关键词节点替换为[ ]*1,后续更新时,根据上下文链条逐步调整最大可选项指示的最大数量[ ]*k。

[0168] 或者,处理策略可以为:将目标样例规则拆分为若干子样例规则。

[0169] [token-1][token-2]…[token-i]…[token-j]…[token-n]可被拆分为:

[0170] [token-1][token-2]…[token-i]…[token-j]

[0171] [token-j]…[token-n]

[0172] 在一实施例中,以上下文泛化作为泛化处理的路径时,上述几种处理策略可配置上下文链条。这种情况下,目标样例规则在多轮泛化过程中,可表示为:

[0173] [token-1][token-2]…[token-i]…[token-j]…[token-n]

[0174] 删除开头或结尾的词节点:[token-2]…[token-i]…[token-j]…[token-n]

[0175] 用[ ]*k替换非关键词节点:[token-2]…[ ]*k…[token-j]…[token-n]

[0176] 规则拆分:[token-2]…[ ]*k…[token-j];[token-j]…[token-n]

[0177] 在一实施例中,目标样例规则包括多个词节点,当通过词节点泛化作为泛化处理的路径时,处理策略可以为:对多个词节点中若干关键词节点,进行同义词拓展。其中,关键词节点为标注字段中的词节点。

[0178] 对于可以枚举的关键词节点同义词,词节点泛化可以直接用逻辑或枚举同义词。

例如:关键词节点[飞机票]可扩展为[飞机票|轮船票|火车票|高铁票]或[*票],这里,*匹配词节点的任意多个汉字,代码[*票]匹配后缀为“票”的任何词节点。

[0179] 或者,处理策略可以为:将多个词节点中若干关键词节点替换为其对应的本体入口特征(ontology feature),依据本体上下位链条指示的顺序,对本体入口特征进行泛化迭代。入口特征是词典连接本体概念知识库的符号标签,本体入口特征为本体上下文链条中底层的概念;本体上下位链条为多个存在上下位关系的概念构成的链条。本体知识库可以调用开源的知识库“知网”(HowNet)。示例性的,关键词节点“北京”、“南京”在“知网”的本体入口特征都是“city”。

[0180] 对于任一难以枚举同义词的关键词节点,在第一次对该关键词节点进行泛化处理时,可以将该关键词节点替换为该关键词节点对应的本体入口特征。后续继续对该关键词节点对应的本体入口特征进行泛化时,可以依据本体上下位链条(taxonomy)对本体入口特征进行更新,从而使得更新后的特征相比更新前的特征具有更宽泛的约束条件。示例性的,根据本体上下文链条,可以将“city”更新为“place”,这实际上在NLP规则匹配过程中引进了最直接的概念常识推理。

[0181] 本体上下位链条可以指示词节点对应的本体入口特征从下位到上位的泛化顺序。示例性的,“老虎”这个词在“知网”在本体入口特征是beast,其本体上下位链条为:beast→animal→AnimalHuman→animate→physical→thing。箭头指示了本体概念一步步泛化的过程。

[0182] 图8为本申请一实施例提供的样例规则泛化处理的整体流程示意图,如图8所示,首先可以准备数据集,也就是前述语料集:开发集、召回测试集和初始样例集。对初始样例集中的语料进行标注之后,得到经过标注的样例集,并可针对每一已标注的样例语料,生成对应的样例规则。

[0183] 逐个选中每一样例规则,作为目标样例规则Rule(n),并对目标样例规则进行一系列逐轮泛化工作,包括:掐头去尾、特征替换、同义词拓展、规则拆分、[]*k替换非关键词节点等。每一轮泛化处理后,在开发集上进行精确率回归测试。针对经过泛化处理的目标样例规则,以及当前规则集中的规则,编译出新的字段抽取引擎,并以字段抽取引擎对开发集进行字段抽取,获得第一抽取结果。通过diff工具判断第一抽取结果与开发集对应的基准抽取结果是否存在差异槽位信息点。一方面,若存在,可以判断差异槽位信息点是否符合精确率要求。一种情况下,不符合精确率要求,可以回滚,以本轮泛化前的目标样例规则,进行下一轮泛化处理。另一种情况下,符合精确率要求,可以接着进行召回回归测试。另一方面,若不存在差异槽位信息点,说明本轮泛化没有导致精确率的变化,也可以接着进行召回回归测试。

[0184] 通过字段抽取引擎对召回测试集进行字段抽取,获得第二抽取结果,并可以判断第二抽取结果中槽位信息点数量,超出召回测试集对应的现存基准槽位信息点数量的比例,是否达到比例阈值。一方面,未达到比例阈值,说明不符合召回要求,可以回滚,以本轮泛化前的目标样例规则,进行下一轮泛化处理。另一方面,达到比例阈值,说明符合召回要求。两项回归测试均通过后,可以更新基准抽取结果和基准槽位信息点数量,并继续下一轮的泛化流程。

[0185] 在缺乏标注数据的业务场景中,本申请方案可以在少量经过标注的样例语料的基

础上,先自动生成样例规则,然后泛化得到多个字段抽取规则,并以多个字段抽取规则编译或解释得到字段抽取引擎。本方案大大降低了纯手工高代码开发的人力成本和时间成本,也可以避免纯人工编码出现的句法错误,降低了开发者的培训门槛和成本。

[0186] 下面列举实际应用场景,对本申请实施例提供的方案进行说明。

[0187] 实施例1

[0188] 针对智能助理应用的技能,开发问句理解系统所需的字段抽取引擎。

[0189] 步骤1:准备半自动开发流程所需的语料集。例如:智能助理相关技能的问题集可能根据技能类型分为自动订票技能、天气问答技能等,可以根据技能类型准备相应的语料集。

[0190] 步骤2:初始化字段抽取引擎NLP-Tagger。

[0191] 步骤3:为开发集建立基准抽取结果,初始化为空;为召回测试集建立基准槽位信息点数量,初始化为0。

[0192] 步骤4:对初始样例集中的语料进行标注,获得已标注的样例语料。

[0193] 例如:对于自动订票技能,样例语料的标注如下:

[0194] 我想订一张[明天下午:Time]从[南京:FromLocation]到[北京:ToLocation]的[飞机票:VehicleType]

[0195] 想订[五月20号:Time]从[武汉:FromLocation]到[广州:ToLocation]的[火车票:VehicleType]

[0196] 从[南京:FromLocation]至[安庆:ToLocation]的[轮船票:VehicleType],[后天上午:Time]的

[0197] 要一张[北京:FromLocation]直达[莫斯科:ToLocation]的[飞机票:VehicleType],[三月20日:Time]就好

[0198] .....

[0199] 步骤5:全自动生成样例规则。例示如下:

[0200] [我][想][订][一][张]<[明天][下午]:Time>[从]<[南京]:FromLocation>[到]<[北京]:ToLocation>[的]<[飞机票]:VehicleType>

[0201] [想][订]<[五月][20][号]:Time>[从]<[武汉]:FromLocation>[到]<[广州]:ToLocation>[的]<[火车票]:VehicleType>

[0202] [从]<[南京]:FromLocation>[至]<[安庆]:ToLocation>[的]<[轮船票]:VehicleType>,<[后天][上午]:Time>[的]

[0203] [要][一][张]<[北京]:FromLocation>[直达]<[莫斯科]:ToLocation>[的]<[飞机票]:VehicleType>[,]<[三月][20][日]:Time>[就][好]

[0204] .....

[0205] 步骤6:在所有样例规则中,逐一选择规则Rule(n),从n=1到n=m,以步骤7的方式进行泛化处理。当穷尽所有样例规则n=m时,完成样例规则的泛化迭代过程,最终形成的规则集合示例如下:

[0206] <[day][time]?:Time>[]*5<[place]:FromLocation>[到|至|直达]<[place]:ToLocation>[]*5<[*票]:VehicleType>

[0207] <[month][number][号|日]:Time>[]*5<[place]:FromLocation>[到|至|直达]<

[place]:ToLocation>[]*5<[*票]:VehicleType>

[0208] <[place]:FromLocation>[到|至|直达]<[place]:ToLocation>[]*5<[*票]:VehicleType>[]*5<[day]?[time]:Time>

[0209] .....

[0210] 其中Time字段可以进一步合并同类项为一个宏代码如下：

[0211] @TimeZiduan＝<[month]?[number]?[day|号|日][time]?:Time>,

[0212] 合并后的头两条规则合二为一,用@表示调用宏代码,则最终定型的规则集合为：

[0213] @TimeZiduan[]*5<[place]:FromLocation>[到|至|直达]<[place]:ToLocation>[]*5<[*票]:VehicleType>

[0214] <[place]:FromLocation>[到|至|直达]<[place]:ToLocation>[]*5<[*票]:VehicleType>[]*5@TimeZiduan

[0215] .....

[0216] 半自动规则开发毕。

[0217] 上述经过泛化的规则集经编译或解释执行,可以成功捕捉样例外的许多案例,例如：

[0218] 想订张五月八号下午的从郑州到武汉的高铁票

[0219] 来一张八月10号自南昌到北京的飞机票

[0220] 劳驾买一张青岛至天津的轮船票,要9月9号下午的

[0221] 麻烦你给我订购一张六月8日广州直达北京高铁票

[0222] .....

[0223] 步骤7:该步骤是一个不断循环和迭代规则Rule(n)的过程。例如,Rule(1)半自动泛化主要流程如下所示,下列泛化的每一步均通过前述的两项回归测试,直到最后定型。

[0224] [我][想][订][一][张][明天下午:Time][从][南京:FromLocation][到][北京:ToLocation][的][飞机票:VehicleType]

[0225] →<[明天][下午]:Time>[从]<[南京]:FromLocation>[到]<[北京]:ToLocation>[的]<[飞机票]:VehicleType>//上下文泛化:掐头,序列变短

[0226] →<[day][subday]:Time>[从]<[city]:FromLocation>[到]<[city]:ToLocation>[的]<[飞机票]:VehicleType>//词泛化:代入本体入口标签city/day/subday等

[0227] →<[day][subday]:Time>[]*5<[city]:FromLocation>[到]<[city]:ToLocation>[]*5<[飞机票]:VehicleType>//上下文泛化:用[]*k替换非关键词

[0228] →<[day][subday]:Time>[]*5<[city]:FromLocation>[到|至|直达]<[city]:ToLocation>[]*5<[*票]:VehicleType>//词泛化:同义词扩展

[0229] →<[day][subday]?:Time>[]*5<[city]:FromLocation>[到|至|直达]<[city]:ToLocation>[]*5<[*票]:VehicleType>//上下文泛化:可选项[]?的出现

[0230] →<[day][time]?:Time>[]*5<[place]:FromLocation>[到|至|直达]<[place]:ToLocation>[]*5<[*票]:VehicleType>//词泛化:标签上下位链条city→place;subday→time

[0231] 实施例2

[0232] 在法律领域,需要对刑事判决书做信息抽取,以便为所有刑事判决案例自动构建

一个刑事判决知识图谱。这样的知识图谱可以为案例查询和调研提供全方位的准确情报，这是传统的关键词检索无法做到的。

[0233] 该实施例的一项基本信息抽取业务是：从刑事判决书中标注8种实体。这也是一个典型的领域NLP应用场景，输入处理对象是中文文本数据(刑事判决书)，输出的是8类字段：defendant，gender，birthday，birthplace，ethnicgroup，education，employer，address等。与很多领域NLP任务一样，该项目只有刑事判决书的原始历史文档，缺乏标注数据。因此，本发明NLP是一个合适的应用。

[0234] 法律领域与智能助理技能是完全不同的领域场景，同一个方案的有效实施证明该方案对于不同领域业务场景应用的普适性。

[0235] 该实施例的半自动开发流程的第一步是全自动生成的初始样例规则，如下所示：

[0236] [被告人][陈小红:defendant][，][女:gender][，][1970年二月3日:birthday][出生][于][A省B市:birthplace][，][汉族:ethnicgroup][，][中专文化:education][，][工作单位][嘉兴xx营销策划有限公司:employer][，][住][A省B市:address]

[0237] 规则泛化的具体实施非常类似实施例1里面的泛化流程。按照这个流程步步迭代，最终结果是下列泛化规则集：

[0238] [被告人|被告]<[]*2[name]:defendant>

[0239] <[女|男]:gender>[，]

[0240] [，]<[]*5[日|号]:birthday>[出生]

[0241] [住|住址]<[]*3[place|name]:birthplace>

[0242] [，]<[]?[*5族]:ethnicgroup>

[0243] <[school]:education>[文化|程度|毕业|肄业]

[0244] [工作单位]<[]*8[noun]:employer>[，]

[0245] 图9是本发明一实施例的一种字段抽取引擎的生成装置的框图，如图9所示，该装置可以包括：生成模块910、泛化模块920、获取模块930以及编译模块940。

[0246] 生成模块910，用于针对样例集中每一已标注的样例语料，生成与所述样例语料对应的样例规则，获得多个样例规则；

[0247] 泛化模块920，用于逐个选中每一样例规则，作为目标样例规则，并对所述目标样例规则进行泛化处理，直至所述目标样例规则完成泛化处理；

[0248] 获取模块930，用于将完成泛化处理的目标样例规则作为字段抽取规则，获得多个字段抽取规则；

[0249] 编译模块940，用于基于所述多个字段抽取规则编译或解释得到字段抽取引擎，并获得基于所述规则模块的字段抽取引擎。

[0250] 上述装置中各个模块的功能和作用的实现过程具体详见上述字段抽取引擎的生成方法中对应步骤的实现过程，在此不再赘述。

[0251] 在本申请所提供的几个实施例中，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申请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，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。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，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系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0252] 另外，在本申请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。

[0253] 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，本申请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)执行本申请各个实施例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-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

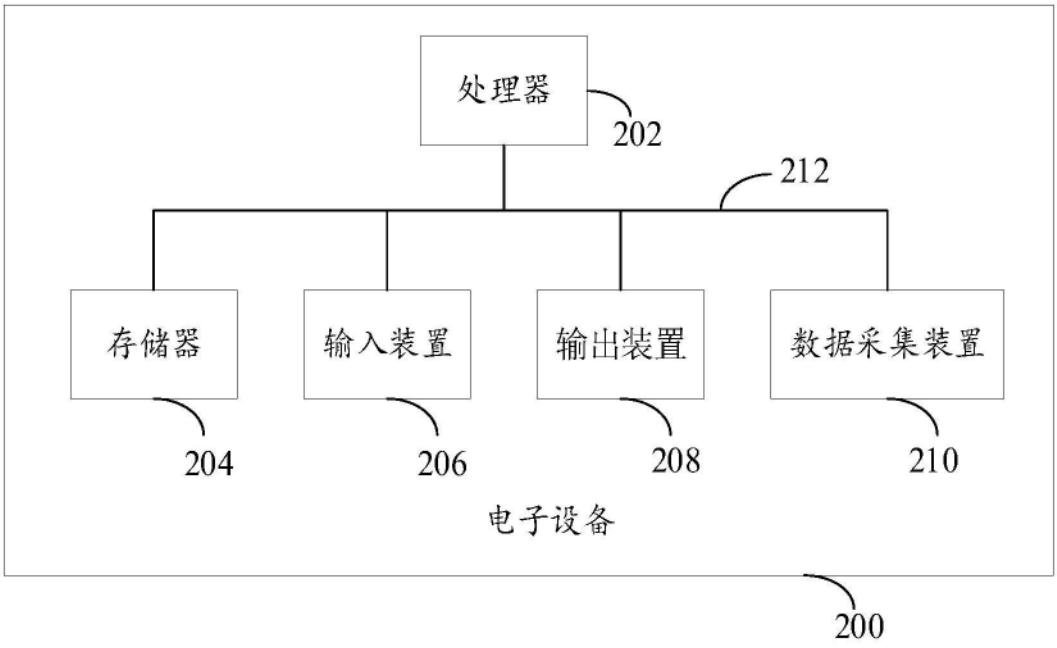

处理器
202
212
存储器
输入装置
输出装置
数据采集装置
204
206
208
210
电子设备
200

图1

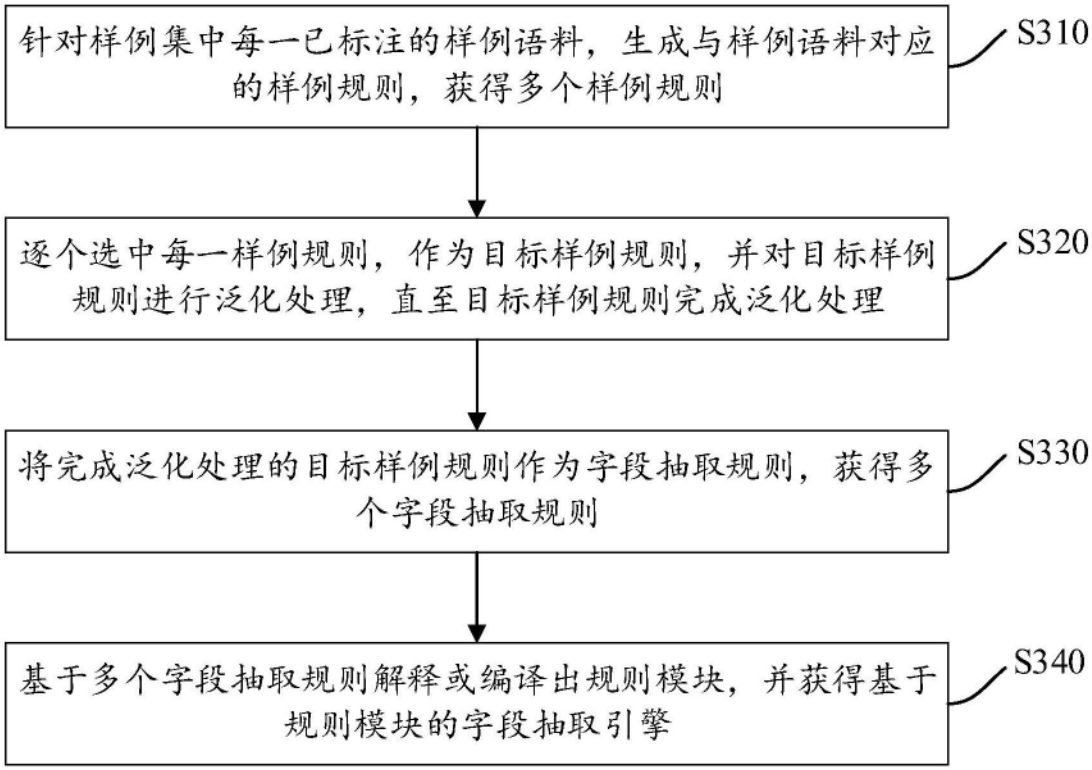

针对样例集中每一已标注的样例语料，生成与样例语料对应的样例规则，获得多个样例规则
S310
逐个选中每一样例规则，作为目标样例规则，并对目标样例规则进行泛化处理，直至目标样例规则完成泛化处理
S320
将完成泛化处理的目标样例规则作为字段抽取规则，获得多个字段抽取规则
S330
基于多个字段抽取规则解释或编译出规则模块，并获得基于规则模块的字段抽取引擎
S340

图2

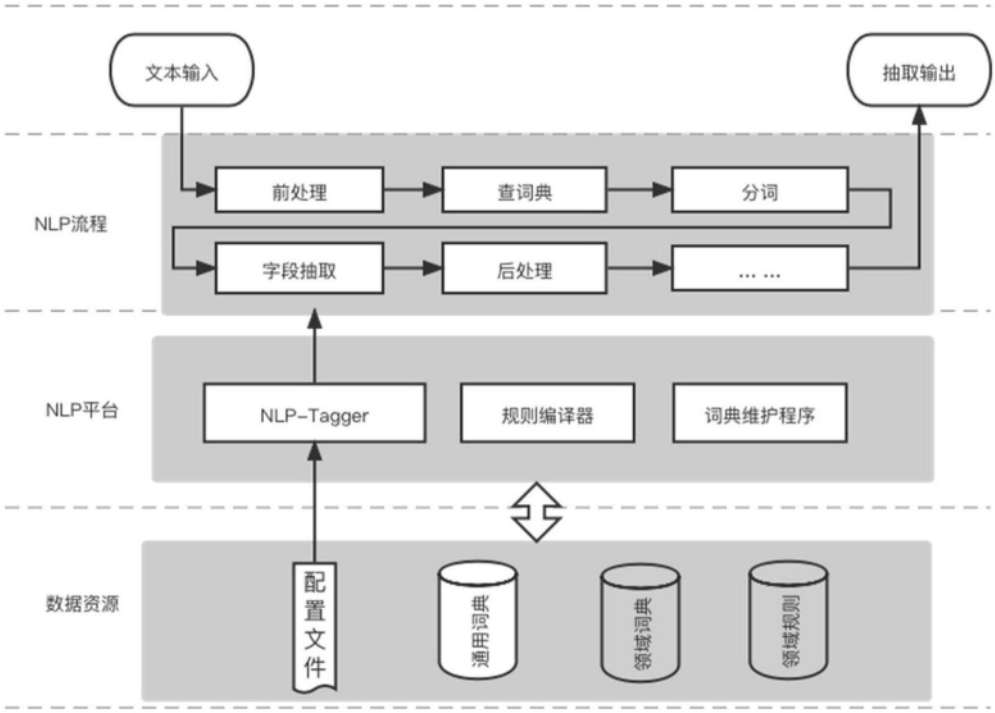

文本输入
抽取输出
NLP流程
前处理
查词典
分词
字段抽取
后处理
... ...
NLP平台
NLP-Tagger
规则编译器
词典维护程序
数据资源
配置文件
通用词典
领域词典
领域规则

图3

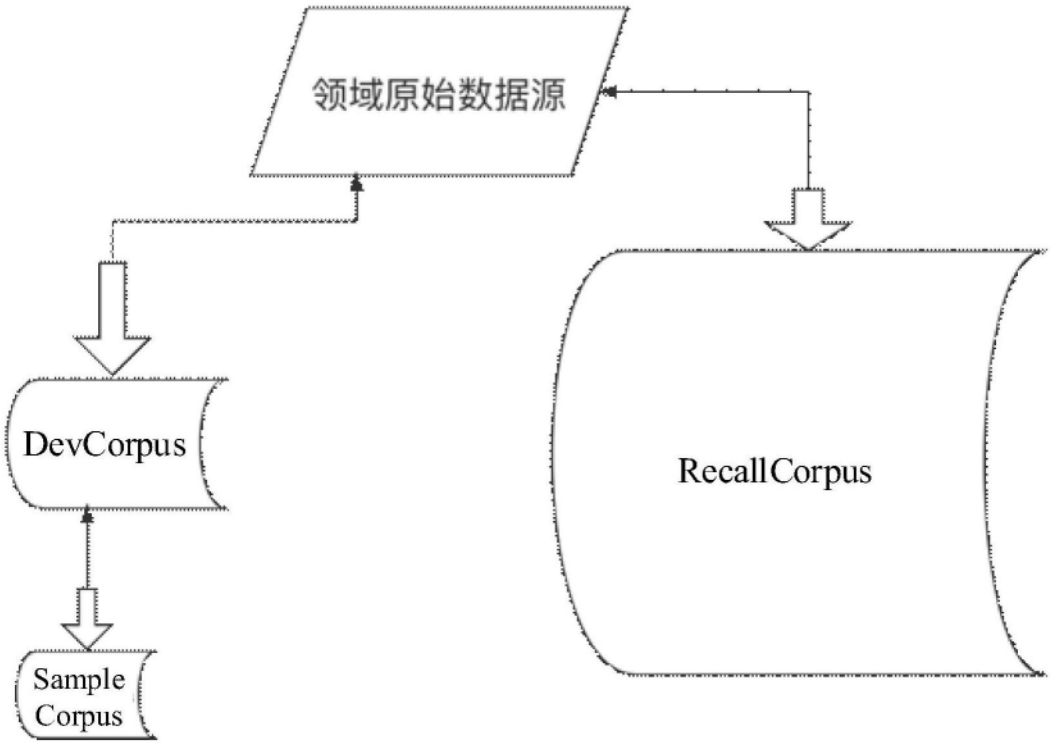

领域原始数据源
DevCorpus
RecallCorpus
Sample Corpus

图4

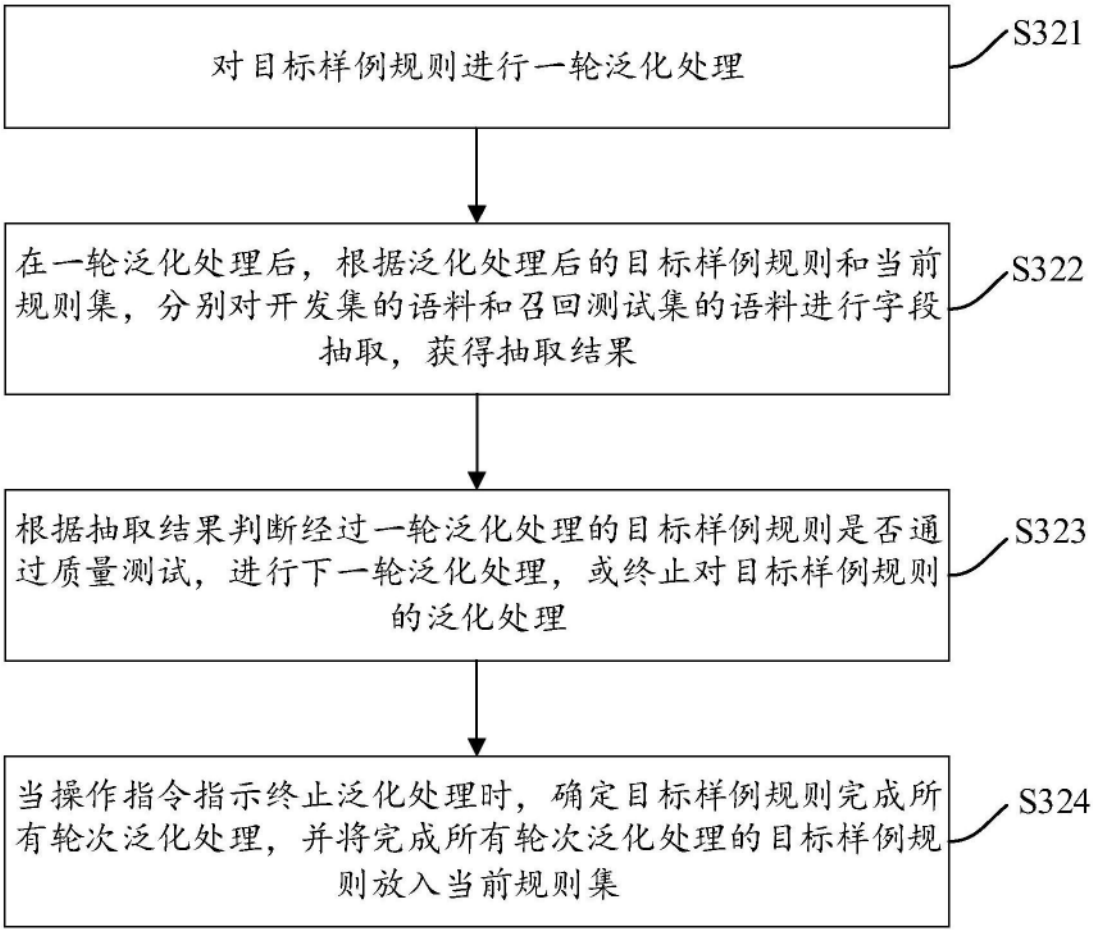

对目标样例规则进行一轮泛化处理
S321
在一轮泛化处理后，根据泛化处理后的目标样例规则和当前规则集，分别对开发集的语料和召回测试集的语料进行字段抽取，获得抽取结果
S322
根据抽取结果判断经过一轮泛化处理的目标样例规则是否通过质量测试，进行下一轮泛化处理，或终止对目标样例规则的泛化处理
S323
当操作指令指示终止泛化处理时，确定目标样例规则完成所有轮次泛化处理，并将完成所有轮次泛化处理的目标样例规则放入当前规则集
S324

图5

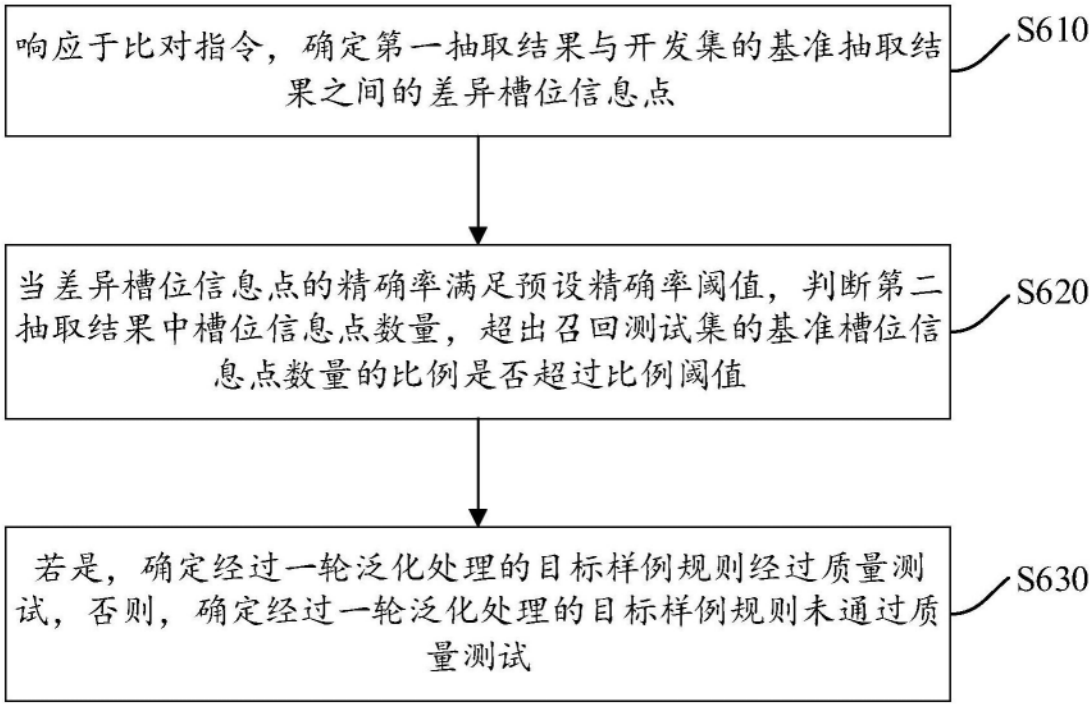

响应于比对指令，确定第一抽取结果与开发集的基准抽取结果之间的差异槽位信息点
S610
当差异槽位信息点的精确率满足预设精确率阈值，判断第二抽取结果中槽位信息点数量，超出召回测试集的基准槽位信息点数量的比例是否超过比例阈值
S620
若是，确定经过一轮泛化处理的目标样例规则经过质量测试，否则，确定经过一轮泛化处理的目标样例规则未通过质量测试
S630

图6

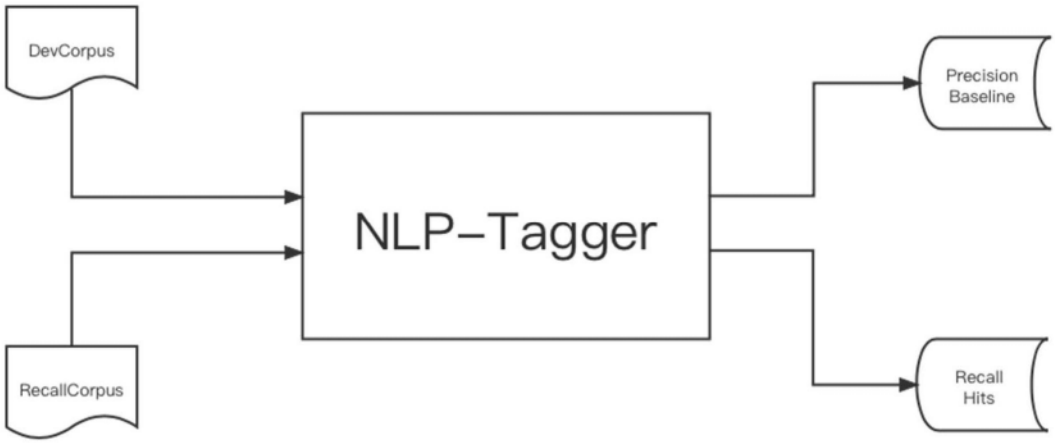

DevCorpus
RecallCorpus
NLP-Tagger
Precision Baseline
Recall Hits

图7

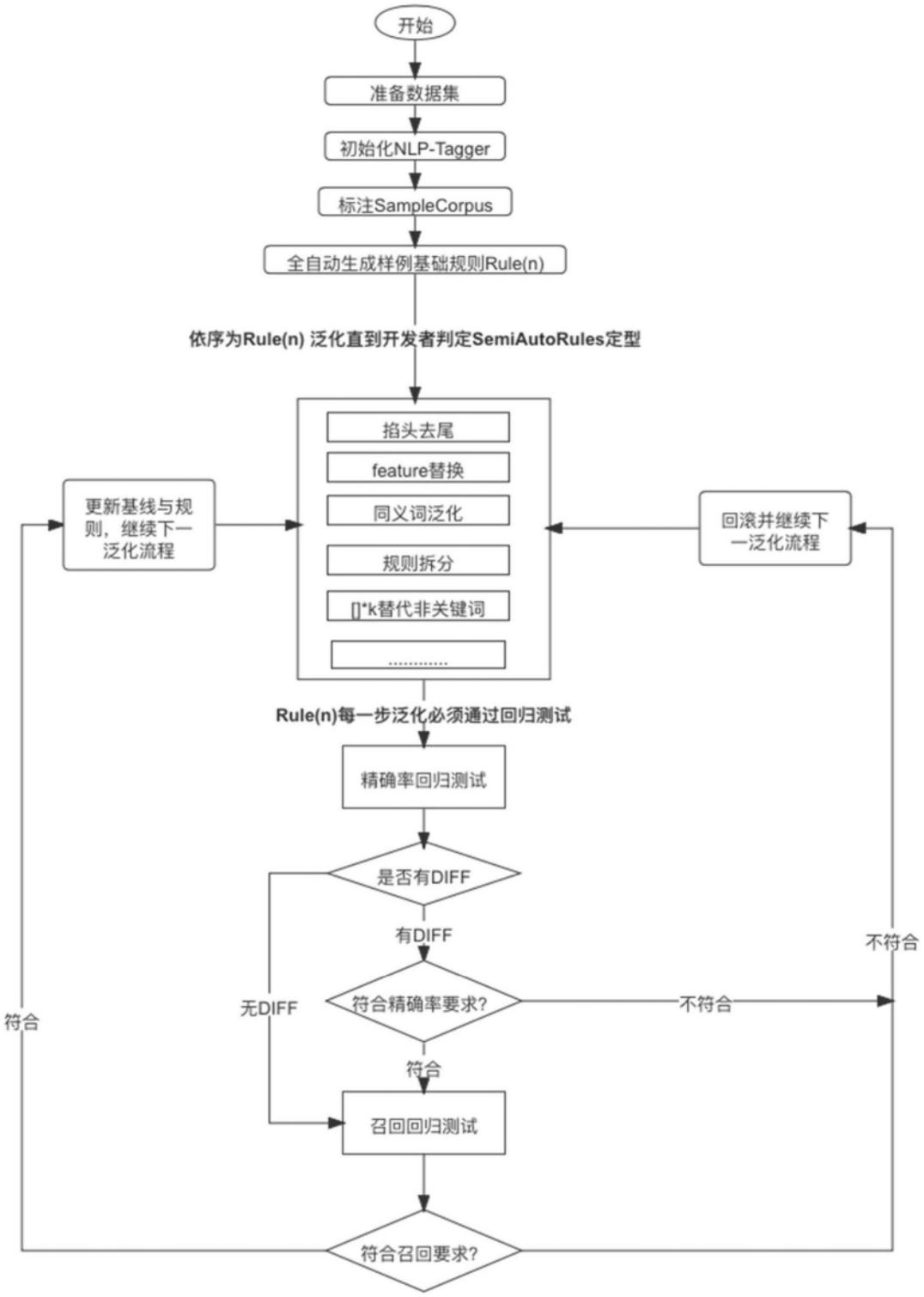

开始
准备数据集
初始化NLP-Tagger
标注SampleCorpus
全自动生成样例基础规则Rule(n)
依序为Rule(n) 泛化直到开发者判定SemiAutoRules定型
掐头去尾
feature替换
同义词泛化
规则拆分
[]*k替代非关键词
.....
更新基线与规则，继续下一泛化流程
回滚并继续下一泛化流程
Rule(n)每一步泛化必须通过回归测试
精确率回归测试
是否有DIFF
有DIFF
无DIFF
符合精确率要求?
不符合
不符合
符合
符合
召回回归测试
符合召回要求?

图8

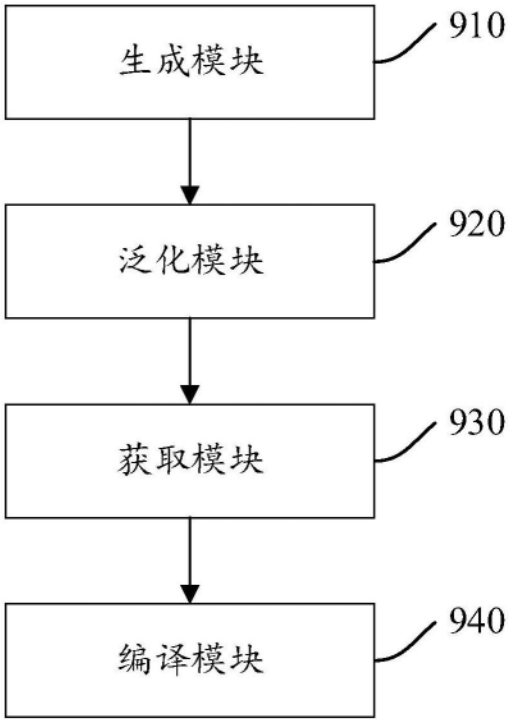


图9