(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115017271 B
(45)授权公告日 2022.11.08

(21)申请号 202210944442.6

(22)申请日 2022.08.08

(65)同一申请的已公布的文献号
申请公布号 CN 115017271 A

(43)申请公布日 2022.09.06

(73)专利权人 杭州实在智能科技有限公司
地址 310000 浙江省杭州市余杭区余杭街道文一西路1818-2号6幢6层

(72)发明人 马富欣 孙林君

(74)专利代理机构 浙江永鼎律师事务所 33233
专利代理师 周希良

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/335(2019.01)
G06F 16/35(2019.01)
G06F 40/194(2020.01)
G06F 40/205(2020.01)
G06F 40/232(2020.01)
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06N 3/04(2006.01)

(56)对比文件
CN 111126781 A,2020.05.08
CN 111178052 A,2020.05.19
CN 110969008 A,2020.04.07
US 9875235 B1,2018.01.23
US 2020234183 A1,2020.07.23
陈云等.基于受限自然语言和模块组合的代码自动生成.《计算机工程》.2008,(第20期),第58-60页.
次曲.浅析一种面向室内智能机器人导航的路径自然语言处理方法.《科技风》.2017,(第10期),第8页.
Gota Dan etal..Multi-Channel Chatbot and Robotic Process Automation.《IEEE》.2022,全文.

审查员 齐智超

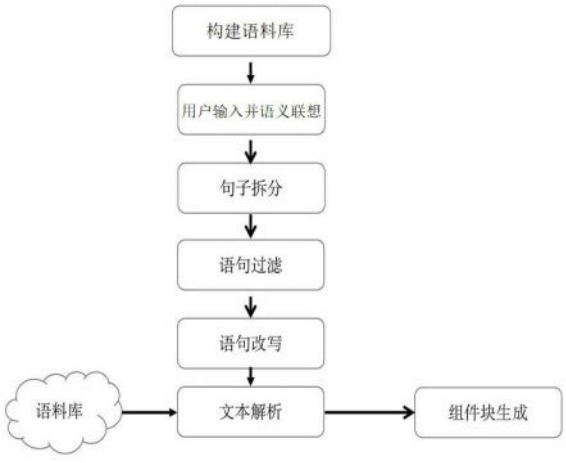
权利要求书2页 说明书10页 附图3页

(54)发明名称

用于智能生成RPA流程组件块的方法及系统

(57)摘要

本发明属于RPA产品技术领域,具体涉及用于智能生成RPA流程组件块的方法及系统。方法包括S1,构建组件语料库;S2,用户输入语句,通过文本相似度算法从组件语料库中获取N条最相似的语句推荐给用户;S3,对用户输入的语句进行拆分,获得拆分后的语句;S4,将拆分后的语句中,与组件操作无关的语句进行过滤,获得过滤后的语句;S5,对过滤后的语句进行语义改写,获得改写后的语句;S6,对改写后的语句进行解析,并根据解析结果生成组件块。本发明具有能够实现用户通过输入自然语言描述,即可自动生成流程并对属性进行填充的操作,使用户的入门门槛降低,同时减少用户手动填写属性时间成本的特点。



CN 115017271 B

1.用于智能生成RPA流程组件块的方法,其特征在于,包括如下步骤;

S1,构建组件语料库;

S2,用户输入语句,通过文本相似度算法从组件语料库中获取N条最相似的语句推荐给用户;

S3,对用户输入的语句进行拆分,获得拆分后的语句;

S4,将拆分后的语句中,与组件操作无关的语句进行过滤,获得过滤后的语句;

S5,对过滤后的语句进行语义改写,获得改写后的语句;

S6,对改写后的语句进行解析,并根据解析结果生成组件块;

步骤S2包括如下步骤:

S21,在用户输入过程中,实时的将用户输入的语句与组件语料库中的数据,通过文本相似度算法进行匹配,得到相似度匹配结果;

S22,按照相似度匹配结果的数值进行排序,将排名靠前的n个语句实时输出到提示框中并推荐给用户;

S23,若存在和用户需求一致的语句,则用户点击获取;若不存在和用户需求一致的语句,则用户不进行选择;

步骤S3包括如下步骤:

利用序列到序列模型将用户输入的语句拆分为多个简短的语句;

步骤S6包括如下步骤:

S61,使用命名实体识别算法对改写后的语句进行属性提取,提取的属性标签由人工制定;

S62,将改写后的语句和组件语料库中的所有数据,通过文本相似度算法进行相似度计算,得到相似度排名靠前的N条语句;

S63,将所述N条语句进输入排序算法模型,得到最相似的一条语句;

S64,确定最相似语句所涉及的组件、组件需要的属性及属性个数,并和NER提取的属性进行对比;若组件需要的组件属性及个数与提取的属性相匹配,则对组件属性进行填充,否则保留组件原有的属性;

S65,基于步骤S64过程,生成组件块;所述组件块包括组件名和组件代码。

2.根据权利要求1所述的用于智能生成RPA流程组件块的方法,其特征在于,步骤S1包括如下步骤:

S11,根据项目实施材料,对项目中涉及的流程拆分为若干个组件块;

S12,对组件块中涉及的组件进行统计并整理,并由人工根据组件进行句子的构造;

S13,构造后的语句根据组件涉及的属性进行NER的标注,最终形成<组件、句子、属性>三元组。

3.根据权利要求1所述的用于智能生成RPA流程组件块的方法,其特征在于,步骤S4包括如下步骤:

S41,对拆分后的语句采用文本分类模型进行二分类,得到介于(0,1)间的概率数值;

S42,若概率数值大于预先设定的阈值,则将对应的语句过滤。

4.根据权利要求1所述的用于智能生成RPA流程组件块的方法,其特征在于,步骤S5包括如下步骤:

S51,通过实体替换和文本纠错算法将过滤后的语句进行拼写错误纠正;

S52,对语句中存在的属性缺失和指代词语,通过规则和端到端的神经网络进行属性补充和指代消解。

5.用于智能生成RPA流程组件块的系统,用于实现权利要求1-4任一项所述的用于智能生成RPA流程组件块的方法,其特征在于,所述用于智能生成RPA流程组件块的系统包括;

组件语料库构建模块,用于构建组件语料库;

语义联想模块,用于用户输入语句,通过文本相似度算法从组件语料库中获取N条最相似的语句推荐给用户;

语句拆分模块,用于对用户输入的语句进行拆分,获得拆分后的语句;

语句过滤模块,用于将拆分后的语句中,与组件操作无关的语句进行过滤,获得过滤后的语句;

语句改写模块,用于对过滤后的语句进行语义改写,获得改写后的语句;

文本解析模块,用于对改写后的语句进行解析,并根据解析结果生成组件块。

6.根据权利要求5所述的用于智能生成RPA流程组件块的系统,其特征在于,所述组件语料库构建模块具体如下:

根据项目实施材料,对项目中涉及的流程拆分为若干个组件块;

对组件块中涉及的组件进行统计并整理,并由人工根据组件进行句子的构造;

构造后的语句根据组件涉及的属性进行NER的标注,最终形成<组件、句子、属性>三元组。

7.根据权利要求5所述的用于智能生成RPA流程组件块的系统,其特征在于,所述语义联想模块具体如下:

在用户输入过程中,实时的将用户输入的语句与组件语料库中的数据,通过文本相似度算法进行匹配,得到相似度匹配结果;

按照相似度匹配结果的数值进行排序,将排名靠前的n个语句实时输出到提示框中并推荐给用户;

若存在和用户需求一致的语句,则用户点击获取;若不存在和用户需求一致的语句,则用户不进行选择。

用于智能生成RPA流程组件块的方法及系统

技术领域

[0001] 本发明属于RPA产品技术领域,具体涉及用于智能生成RPA流程组件块的方法及系统。

背景技术

[0002] RPA(Robotic Process Automation),可以称之为机器人流程自动化,是一种软件技术。它可以通过模拟人操作电脑,替代人工执行一些有规则的、机械性和重复性的工作,将企业中的人力资源释放出来,减轻企业的用人成本,可以极大提升工作效率和准确性。

[0003] 随着信息数字化的高速发展,RPA在各个行业中得到了广泛的应用,已经成为助力企业组织向“智能自动化转型”,构建业务流程自动化的重要引擎之一。

[0004] 目前,市面上有各种各样的RPA产品,这些产品都会存在一个RPA设计器,其中RPA设计器由大量的组件构成(如打开Excel,打开网页),这些组件对日常常用的操作进行封装,用户可以在RPA设计器中通过拖拽组件的方式进行组合,不同的组件组合构成一个功能不同的RPA流程包(流程包是指通过组合一系列组件构成的针对某个具体业务的组件集),通过执行构建的流程包实现流程自动化。

[0005] 但由于组件数量过多,对于第一次使用RPA的用户来说,不清楚自己的任务可能会设计哪些组件,存在一定的使用难度和入门门槛。

[0006] 目前流程组件生成的系统主要包括以下两类:一类是基于视频分析的流程生成方法,该类系统通过用户操作组件的视频与系统产生的操作日志通过深度学习技术对视频流进行识别并生成流程;另一类是基于NLP技术的流程生成方法,该类系统分为三种,第一种通过使用人工构建的话术模板对用户的输入文本进行切分得到流程,然后利用序列标注模型来识别组件及属性,最后根据组件及属性来生成组件代码;第二种通过会话方式来创建流程,首先人为的控制会话的起始与终止位置来切分会话,然后将切分后的句子生成句向量,最后通过相似度的方法来得到对应的流程,但是该方法不设计组件属性的提取与填充;第三种方法首先通过人工收集RPA流程包并构建一个专家系统,然后使用文本分类来识别组件,利用实体识别来提取属性,最后生成组件代码。

[0007] 基于视频流的流程生成技术的一个前提是需要一个操作视频,但对于初次使用RPA的用户来说,由于组件数量过多,要完成一个完整可执行的流程操作存在一定的难度。

[0008] 基于NLP的流程生成系统,存在以下缺点:

[0009] 1.现有的系统只能在用户全部写完需求后才能生成对应的组件块,不能在用户输入时实时动态的根据输入提供联想提示来引导用户来快速实现自己的需求或者类似需求,提高用户体验。

[0010] 2.现有系统的句子拆分通过标点符号或者通过配置话术模板来对输入进行切分。只通过标点切分会容易误切分句子导致文本上下文语义缺失,比如“打开A表,B表”,通过标点切分后为“打开A表”,“B 表”两个短句,会缺少Excel操作中的“打开”动作,会丢失重要的文本语义信息,进而导致组件识别错误;通过话术模板方法来进行句子切分虽然可以避免

使用标点切分的弊端，但是话术模板需要人工收集、配置并需要不断更新，耗费大量的人力成本。

[0011] 3.现有系统不支持对用户输入的自然语言描述进行提前过滤及改写机制，当存在口语化输入或者误输入同音词时，由于与构建语料库中的数据不在同一分布下从而影响输入的质量，导致无法基于用户的描述生成最优的组件；当输入的文本为闲聊语句时，此时需要将其过滤，不为其生成对应的组件信息。

[0012] 4.现有流程组件生成系统中的文本解析模块主要包含组件识别和属性填充两部分。其中组件识别被当作意图识别并使用文本分类模型建模预测，但由于设计器中的组件数量过多，使用文本分类模型会导致组件识别的准确性较差；另一方面如果一句话对应多个组件时，使用分类模型到的组件标签还会存在组件先后顺序的问题，标签乱序会对流程的执行结果存在很大影响。例如分类识别出的组件为“写入单元格”和“关闭Excel”，如果关闭操作在前，写入操作在后，由于组件的先后顺序不同，导致RPA流程执行的结果也不同。

[0013] 因此，设计一种能够通过将深度学习技术与RPA结合，实现用户通过输入自然语言描述即可自动生成流程并对属性进行填充的操作，使用户的入门门槛降低，同时减少用户手动填写属性时间成本的用于智能生成RPA流程组件块的方法及系统，就显得十分重要。

[0014] 例如，申请号为CN202110927454.3的中国专利文献描述的一种基于状态转移概率模型的RPA组件推荐方法，包括：该方法通过统计业务场景中各类操作对应功能组件的使用关系，得到每个功能组件到其他组件的转移概率；以转移概率为依据为每一步RPA流程创建推荐组件，并通过高频组件的类别分布，推荐组件类别；虽然提到的组件及类别推荐基于转移概率模型实现，模型训练数据量越大，模型准确率越高；为避免过拟合应当选择当前业务场景中多个业务人员的操作数据为输入，以防止个人的不规范操作习惯影响模型整体的准确率；另外，可以帮助创建者快速找到适用组件，加速流程构建，但是其缺点在于，当现有用户初次使用RPA时，由于组件数量过多，要完成一个完整可执行的流程操作存在一定难度，导致用户体验效果差的问题。

发明内容

[0015] 本发明是为了克服现有技术中，当现有用户初次使用RPA时，由于组件数量过多，要完成一个完整可执行的流程操作存在一定难度，导致用户体验效果差的问题，提供了一种能够通过将深度学习技术与RPA结合，实现用户通过输入自然语言描述即可自动生成流程并对属性进行填充的操作，使用户的入门门槛降低，同时减少用户手动填写属性时间成本的用于智能生成RPA流程组件块的方法及系统。

[0016] 为了达到上述发明目的，本发明采用以下技术方案：

[0017] 用于智能生成RPA流程组件块的方法，包括如下步骤；

[0018] S1，构建组件语料库；

[0019] S2，用户输入语句，通过文本相似度算法从组件语料库中获取N条最相似的语句推荐给用户；

[0020] S3，对用户输入的语句进行拆分，获得拆分后的语句；

[0021] S4，将拆分后的语句中，与组件操作无关的语句进行过滤，获得过滤后的语句；

[0022] S5，对过滤后的语句进行语义改写，获得改写后的语句；

[0023] S6,对改写后的语句进行解析,并根据解析结果生成组件块。

[0024] 作为优选,步骤S1包括如下步骤:

[0025] S11,根据项目实施材料,对项目中涉及的流程拆分为若干个组件块;

[0026] S12,对组件块中涉及的组件进行统计并整理,并由人工根据组件进行句子的构造;

[0027] S13,构造后的语句根据组件涉及的属性进行NER的标注,最终形成<组件、句子、属性>三元组。

[0028] 作为优选,步骤S2包括如下步骤:

[0029] S21,在用户输入过程中,实时的将用户输入的语句与组件语料库中的数据,通过文本相似度算法进行匹配,得到相似度匹配结果;

[0030] S22,按照相似度匹配结果的数值进行排序,将排名靠前的n个语句实时输出到提示框中并推荐给用户;

[0031] S23,若存在和用户需求一致的语句,则用户点击获取;若不存在和用户需求一致的语句,则用户不进行选择。

[0032] 作为优选,步骤S3包括如下步骤:

[0033] 利用序列到序列模型将用户输入的语句拆分为多个简短的语句。

[0034] 作为优选,步骤S4包括如下步骤:

[0035] S41,对拆分后的语句采用文本分类模型进行二分类,得到介于(0,1)间的概率数值;

[0036] S42,若概率数值大于预先设定的阈值,则将对应的语句过滤。

[0037] 作为优选,步骤S5包括如下步骤:

[0038] S51,通过实体替换和文本纠错算法将过滤后的语句进行拼写错误纠正;

[0039] S52,对语句中存在的属性缺失和指代词语,通过规则和端到端的神经网络进行属性补充和指代消解。

[0040] 作为优选,步骤S6包括如下步骤:

[0041] S61,使用命名实体识别算法对改写后的语句进行属性提取,提取的属性标签由人工制定;

[0042] S62,将改写后的语句和组件语料库中的所有数据,通过文本相似度算法进行相似度计算,得到相似度排名靠前的N条语句;

[0043] S63,将所述N条语句进输入排序算法模型,得到最相似的一条语句;

[0044] S64,确定最相似语句所涉及的组件、组件需要的属性及属性个数,并和NER提取的属性进行对比;若组件需要的组件属性及个数与提取的属性相匹配,则对组件属性进行填充,否则保留组件原有的属性;

[0045] S65,基于步骤S64过程,生成组件块;所述组件块包括组件名和组件代码。

[0046] 本发明还提供了用于智能生成RPA流程组件块的系统,包括;

[0047] 组件语料库构建模块,用于构建组件语料库;

[0048] 语义联想模块,用于用户输入语句,通过文本相似度算法从组件语料库中获取N条最相似的语句推荐给用户;

[0049] 语句拆分模块,用于对用户输入的语句进行拆分,获得拆分后的语句;

[0050] 语句过滤模块，用于将拆分后的语句中，与组件操作无关的语句进行过滤，获得过滤后的语句；

[0051] 语句改写模块，用于对过滤后的语句进行语义改写，获得改写后的语句；

[0052] 文本解析模块，用于对改写后的语句进行解析，并根据解析结果生成组件块。

[0053] 作为优选，所述组件语料库构建模块具体如下：

[0054] 根据项目实施材料，对项目中涉及的流程拆分为若干个组件块；

[0055] 对组件块中涉及的组件进行统计并整理，并由人工根据组件进行句子的构造；

[0056] 构造后的语句根据组件涉及的属性进行NER的标注，最终形成<组件、句子、属性>三元组。

[0057] 作为优选，所述语义联想模块具体如下：

[0058] 在用户输入过程中，实时的将用户输入的语句与组件语料库中的数据，通过文本相似度算法进行匹配，得到相似度匹配结果；

[0059] 按照相似度匹配结果的数值进行排序，将排名靠前的n个语句实时输出到提示框中并推荐给用户；

[0060] 若存在和用户需求一致的语句，则用户点击获取；若不存在和用户需求一致的语句，则用户不进行选择。

[0061] 本发明与现有技术相比，有益效果是：(1)本发明通过将深度学习技术与RPA结合，实现用户通过输入自然语言描述即可自动生成流程并对属性进行填充的操作，使不熟悉RPA设计器的用户可以通过输入自然语言描述就可自动生成组件，降低了用户的入门门槛，使零门槛入门成为可能，并且属性的自动填充能有效减少用户手动填写属性的时间成本；(2)本发明通过构建高质量的组件语料库，将组件、自然语言描述、组件属性通过三元组展示出来；基于文本相似度模型的语义联想模块，实时的为用户提供联想句，提高联想句对用户实际需求的命中率，减少用户确定实际所需输入内容的耗时；基于规则+序列到序列模型相结合的语句拆分模块，对用户输入的文本进行细粒度的切分，来解决传统方式导致切分错误的问题；然后利用文本分类模型对切分后的语句预先过滤，减轻模型服务的负载，并提高组件生成的准确率；对过滤后的语句采用基于文本纠错的文本改写对用户在输入过程中出现的同谐音、混淆音、形近字、多漏字等错误进行纠正，并使用基于端到端的神经网络进行实体补充及指代消解，提高后续相似度匹配的精度；最后，通过基于文本相似度和NER模型对改写后的句子进行智能解析，识别需求所涉及的组件及组件所需要的属性并自动填充，并生成最终的RPA流程，降低用户对设计器图形界面的操作时间，提供更便利的用户体验，使现有的RPA产品更智能。

附图说明

[0062] 图1为本发明中用于智能生成RPA流程组件块的方法的一种流程图；

[0063] 图2为本发明中文本解析过程的一种流程图；

[0064] 图3为本发明实施例所提供的语义联想过程的一种功能展示图；

[0065] 图4为本发明实施例所提供的文本解析过程实际业务应用的一种流程图。

具体实施方式

[0066] 为了更清楚地说明本发明实施例，下面将对照附图说明本发明的具体实施方式。显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图，并获得其他的实施方式。

[0067] 实施例1：

[0068] 如图1所示的用于智能生成RPA流程组件块的方法，包括如下步骤；

[0069] S1，构建组件语料库；

[0070] S2，用户输入语句，通过文本相似度算法从组件语料库中获取N条最相似的语句推荐给用户；

[0071] S3，对用户输入的语句进行拆分，获得拆分后的语句；

[0072] S4，将拆分后的语句中，与组件操作无关的语句进行过滤，获得过滤后的语句；

[0073] S5，对过滤后的语句进行语义改写，获得改写后的语句；

[0074] S6，对改写后的语句进行解析，并根据解析结果生成组件块。

[0075] 步骤S1包括如下步骤：

[0076] S11，根据项目实施材料，对项目中涉及的流程拆分为若干个组件块；

[0077] S12，对组件块中涉及的组件进行统计并整理，并由人工根据组件进行句子的构造；

[0078] S13，构造后的语句根据组件涉及的属性进行NER的标注，最终形成<组件、句子、属性>三元组。

[0079] 语料库的构建是本发明的基础，形式为一个三元组，由<组件，句子，属性> 构成，比如<打开Excel，打开桌面上的男装表，桌面、男装>该三元组，其中“打开Excel”为RPA中的组件，“打开桌面上的男装表”是“打开Excel”所对应的句子，“桌面”及“男装”是“打开Excel”组件所需要填充的属性值。RPA组件的收集通过领域内的业务专家和专业的实施人员共同对已有的流程包按照不同的场景进行梳理得到，然后根据组件构造一定量的相似句，最后对构造的句子由人工进行实体的标注，得到最终的三元组。

[0080] 步骤S2包括如下步骤：

[0081] S21，在用户输入过程中，实时的将用户输入的语句与组件语料库中的数据，通过文本相似度算法进行匹配，得到相似度匹配结果；

[0082] S22，按照相似度匹配结果的数值进行排序，将排名靠前的n个语句实时输出到提示框中并推荐给用户；

[0083] S23，若存在和用户需求一致的语句，则用户点击获取；若不存在和用户需求一致的语句，则用户不进行选择。

[0084] 本发明通过语义联想功能来提高联想句对用户实际需求的命中率，减少用户确定实际所需输入内容的耗时，辅助其快速搭建流程。

[0085] 步骤S3包括如下步骤：

[0086] 利用序列到序列模型将用户输入的语句拆分为多个简短的语句。能够采用上述拆分方法，能够解决使用标点进行切分后导致切分错误的问题以及避免通过话术模板拆分方式造成人力浪费的问题。

[0087] 步骤S4包括如下步骤：

[0088] S41，对拆分后的语句采用文本分类模型进行二分类，得到介于(0,1)间的概率数值；

[0089] S42，若概率数值大于预先设定的阈值，则将对应的语句过滤。

[0090] 提前过滤无关信息可以减轻模型的负载；另外，提前过滤某些用户闲聊语句或和超出设计器能力范围内的无理需求，能够提高用户的体验感受。

[0091] 步骤S5包括如下步骤：

[0092] S51，通过实体替换和文本纠错算法将过滤后的语句进行拼写错误纠正；

[0093] S52，对语句中存在的属性缺失和指代词语，通过规则和端到端的神经网络进行属性补充和指代消解。

[0094] 通过基于实体库+文本纠错算法将用户的输入进行一定的拼写纠正，使改写后的数据尽可能的和语料库中的数据同分布，保证模型的准确度和泛化性。另一方面，对句子中存在属性缺失和指代词语采用规则+端到端的神经网络来进行属性补充和指代消解。

[0095] 如图2所示，步骤S6包括如下步骤：

[0096] S61，使用命名实体识别算法对改写后的语句进行属性提取，提取的属性标签由人工制定；

[0097] S62，将改写后的语句和组件语料库中的所有数据，通过文本相似度算法进行相似度计算，得到相似度排名靠前的N条语句；

[0098] S63，将所述N条语句进输入排序算法模型，得到最相似的一条语句；

[0099] S64，确定最相似语句所涉及的组件、组件需要的属性及属性个数，并和NER提取的属性进行对比；若组件需要的组件属性及个数与提取的属性相匹配，则对组件属性进行填充，否则保留组件原有的属性；

[0100] S65，基于步骤S64过程，生成组件块；所述组件块包括组件名和组件代码。

[0101] 本发明还提供了用于智能生成RPA流程组件块的系统，包括；

[0102] 组件语料库构建模块，用于构建组件语料库；

[0103] 语义联想模块，用于用户输入语句，通过文本相似度算法从组件语料库中获取N条最相似的语句推荐给用户；

[0104] 语句拆分模块，用于对用户输入的语句进行拆分，获得拆分后的语句；

[0105] 语句过滤模块，用于将拆分后的语句中，与组件操作无关的语句进行过滤，获得过滤后的语句；

[0106] 语句改写模块，用于对过滤后的语句进行语义改写，获得改写后的语句；

[0107] 文本解析模块，用于对改写后的语句进行解析，并根据解析结果生成组件块。

[0108] 组件语料库构建模块具体如下：

[0109] 根据项目实施材料，对项目中涉及的流程拆分为若干个组件块；

[0110] 对组件块中涉及的组件进行统计并整理，并由人工根据组件进行句子的构造；

[0111] 构造后的语句根据组件涉及的属性进行NER的标注，最终形成<组件、句子、属性>三元组。

[0112] 语义联想模块具体如下：

[0113] 在用户输入过程中，实时的将用户输入的语句与组件语料库中的数据，通过文本

相似度算法进行匹配,得到相似度匹配结果;

[0114] 按照相似度匹配结果的数值进行排序,将排名靠前的n个语句实时输出到提示框中并推荐给用户;

[0115] 若存在和用户需求一致的语句,则用户点击获取;若不存在和用户需求一致的语句,则用户不进行选择。

[0116] 语句拆分模块具体如下:

[0117] 利用序列到序列模型将用户输入的语句拆分为多个简短的语句。

[0118] 语句过滤模块具体如下:

[0119] 对拆分后的语句采用文本分类模型进行二分类,得到介于(0,1)间的概率数值;

[0120] 若概率数值大于预先设定的阈值,则将对应的语句过滤。

[0121] 提前过滤无关信息可以减轻模型的负载;另外,提前过滤某些用户闲聊语句或和超出设计器能力范围内的无理需求,能够提高用户的体验感受。

[0122] 语句改写模块具体如下:

[0123] 通过实体替换和文本纠错算法将过滤后的语句进行拼写错误纠正;

[0124] 对语句中存在的属性缺失和指代词语,通过规则和端到端的神经网络进行属性补充和指代消解。

[0125] 通过基于实体库+文本纠错算法将用户的输入进行一定的拼写纠正,使改写后的数据尽可能的和语料库中的数据同分布,保证模型的准确度和泛化性。另一方面,对句子中存在属性缺失和指代词语采用规则+端到端的神经网络来进行属性补充和指代消解。

[0126] 文本解析模块具体如下:

[0127] 使用命名实体识别算法对改写后的语句进行属性提取,提取的属性标签由人工制定;

[0128] 将改写后的语句和组件语料库中的所有数据,通过文本相似度算法进行相似度计算,得到相似度排名靠前的N条语句;

[0129] 将所述N条语句进输入排序算法模型,得到最相似的一条语句;

[0130] 确定最相似语句所涉及的组件、组件需要的属性及属性个数,并和NER提取的属性进行对比;若组件需要的组件属性及个数与提取的属性相匹配,则对组件属性进行填充,否则保留组件原有的属性;

[0131] 基于步骤S64过程,生成组件块;所述组件块包括组件名和组件代码。

[0132] 基于本发明的技术方案,在具体实施和操作过程中的一个典型业务流程如下:

[0133] 1.语料库构建

[0134] 首先构建一个高质量的语料库。根据项目实施材料,对项目中涉及的流程拆分一个个小的组件块,然后对组件块中涉及的组件进行统计并整理,然后由人工根据组件进行句子的构造,构造后的语句根据组件涉及的属性进行NER的标注,最终形成<组件、句子、属性>三元组。

[0135] 具体为:

[0136] 通过项目梳理得到了一个单组件“删除重复”,然后根据此组件构造3个相似句,如“删除PH值页C列的重复项”、“删除值班表姓名列的重复项”、“删除值班表sheet2中第三列的重复项”;接着对这3条相似句进行NER的标注,以第一个相似句为例,标注后的结果为“PH

值”为Sheet名称、“C列”为列名;最后形成<删除重复、“删除PH值页C列的重复项”、{“PH值”:”Sheet名称”,”C列”:”列名”}>三元组。

[0137] 2.语义联想

[0138] 语义联想主要是对用户输入实时的提供语义提示功能。功能展示如图3所示。

[0139] 具体为:

[0140] 在用户输入过程中,实时的将用户输入与构建的语料库中数据通过文本相似度算法(BM25)进行匹配,相似度匹配的结果为一个数值,然后按照数值进行排序,将排名前8个语句实时输出到提示框中,如果存在和用户需求一致的语句则用户可以点击获取。比如当输入打开时,系统会给出上图中的8个语句,当用户的目标是想打开Excel的时候,可以点击“打开csv表格”,减少用户的输入。

[0141] 3.句子拆分

[0142] 句子拆分是利用序列到序列模型将用户输入较长的自然语言拆分为多个简短的语句,通过拆分后的多个短句依次进行文本解析。

[0143] 以“打开Excel在第一行写入ID,姓名,年龄,住址,然后在B2写入张三”为例,如果仅仅按照标点(“,”)进行切分,切分结果为[“打开Excel在第一行写入ID”,“姓名”,“年龄”,“住址”,“然后在B2写入张三”] 5个短句子,当进行输入过滤时,“姓名”,“年龄”和“住址”会被过滤导致写入数据的缺失。因此需要将用户输入输入到序列到序列模型中,输入为原始的语句(用户输入),输出为切分后的句子列表,即“打开Excel“,”在第一行写入ID,姓名,年龄,住址”和”然后在B2写入张三”三个短句,然后对这三个短句分别进行后面的过滤、改写、解析操作。

[0144] 4.语句过滤

[0145] 语句过滤把和组件无关的语句删除,禁止进入后面的解析流程。语句过滤通过对句子拆分后返回的句子列表采用文本分类模型(fasttext)进行二分类,二分类的标签为“过滤”和“不过滤”两种。具体为,将切分后的语句输入到二分类模型中,得到一个介于(0,1)之间的概率值,如果概率值大于0.5,则将用户输入过滤,否则进行语句改写+文本解析模块得到组件块。

[0146] 5.语句改写

[0147] 用户的输入在很大程度上会影响后续文本解析模型的精度,因此需要对拆分后的句子列表进行一定程度的改写,使其和语料库中的数据尽可能的类似。语句改写一方面通过使用实体替换及文本纠错的方式处理拼写错误;另一方面通过规则+基于最大熵模型的指代消解算法对句子中的指示代词进行处理,给出最后改写后的结果用于文本解析中。

[0148] 具体为:以用户想输入“把十面埋伏写入到A1单元格中然后把该单元格的字体颜色设置为红色”这句话,但是在输入时却误写为“把四面埋伏写入到A1单元格中然后把该单元格的字体颜色设置为红色”为例。此时,首先利用基于BERT的文本纠错算法对其进行一定的改写,改为“把十面埋伏写入到A1单元格中然后把该单元格的字体颜色设置为红色”,然后利用指代消岐算法将句子中的“该单元格”修改为“A1单元格”,进一步提高组件属性识别和填充的准确性。

[0149] 6.文本解析

[0150] 文本解析模块以改写后的句子作为输入,并返回文本解析的结果,结果包括组件

和组件对应的属性，然后根据业务将组件和属性结合，生成最后的组件块，组件块包括组件名以及组件代码。

[0151] 以“打开Excel在第一行写入ID，姓名，年龄，住址，然后在B2写入张三”为例，经过上述1～5步骤后，该句被拆分为三个短句，即“打开Excel”，“在第一行写入ID，姓名，年龄，住址”和”然后在B2写入张三”。

[0152] 1)首先对第i(i为计数器，从1～S，S为短句个数)个短句和语料库中的所有数据进行相似度的计算，计算流程如图4所示，得到该语句和N(N为语料库中的数据量)个相似度得分，然后对这N个得分进行排序，得到前50个相似度最高的语句，作为候选集。具体为，以本步骤中的例句为例，循环遍历上述两个语句(S1，S2，S3)，分别将S1、S2和S3输入到相似度模型并召回50条相似语句，表示为($S1_{top1}$，$S1_{top2}$...$S1_{top50}$)、($S2_{top1}$，$S2_{top2}$...$S2_{top50}$)和($S3_{top1}$，$S3_{top2}$...$S3_{top50}$)。

[0153] 2)对S1、S2和S3得到的50个候选集利用排序模型(比如序列attention模型)对其进一步排序 ，对每个句子召回的候选集分别得到50个(0，1)的概率值，然后对概率值排序，输出概率最高的一条语句，得到该语句所对应的组件，其中S1对应的组件为“打开Excel”，S2对应的组件为“写入行”，S3对应的组件为“写入单元格”。

[0154] 3)对第i个短句进行NER实体的识别，得到实体属性值。其中S1得到的属性值为空，S2得到的属性值为{“行数”：“1”，“数据”：“ID，姓名，年龄，住址”}，S3得到的属性值为{“单元格”：“B2”}。

[0155] 4)对得到的组件和实体属性值通过业务规则进行结合 ，具体来说，通过1)2)3)三个步骤将得到组件及组件属性，并以字典的形式表示出来，以本步骤中的实例为例，句子S1对应的组件和属性值分别为{“打开Excel”，[ ]}，句子S2对应的为{“写入行”：[“行值”：“1”，数据”：“ID，姓名，年龄，住址”]}，句子S3对应的为{“写入单元格”：[单元格”：“B2”]}，通过业务逻辑规则拼接得到最终的组件块和代码。业务逻辑规则如下：如果组件需要的属性值与NER识别出的属性值数量一致，则将NER识别出的属性值替换组件代码的默认属性；如果组件属性值数量不一致时，则保留原有的组件代码。

[0156] 5)重复进行上述步骤，直至计数器i与短句个数相等。

[0157] 本发明通过将深度学习技术与RPA结合，实现用户通过输入自然语言描述即可自动生成流程并对属性进行填充的操作，降低了用户的入门门槛，减少用户手动填写属性的时间成本。

[0158] 本发明创造性的设计了一种基于用户输入来自动生成流程组件块的机制和方法，该方法包含语义联想、句子切分、语句过滤、文本解析，该方法可以有效提高组件识别的准确率和属性填充的覆盖率。

[0159] 本发明将语义联想应用于流程生成系统中，能够在用户输入时实时的推荐相关描述，辅助其快速搭建自己的流程。

[0160] 本发明采用文本纠错与指代消解相结合的技术，对语句进行实体补充及指代消解等改写操作，提高后续相似度匹配的精度。

[0161] 本发明设计了一种使用文本相似度和NER相结合的文本解析方法来识别RPA组件和组件属性，能够提高组件识别的准确性。

[0162] 以上所述仅是对本发明的优选实施例及原理进行了详细说明，对本领域的普通技

术人员而言，依据本发明提供的思想，在具体实施方式上会有改变之处，而这些改变也应视为本发明的保护范围。

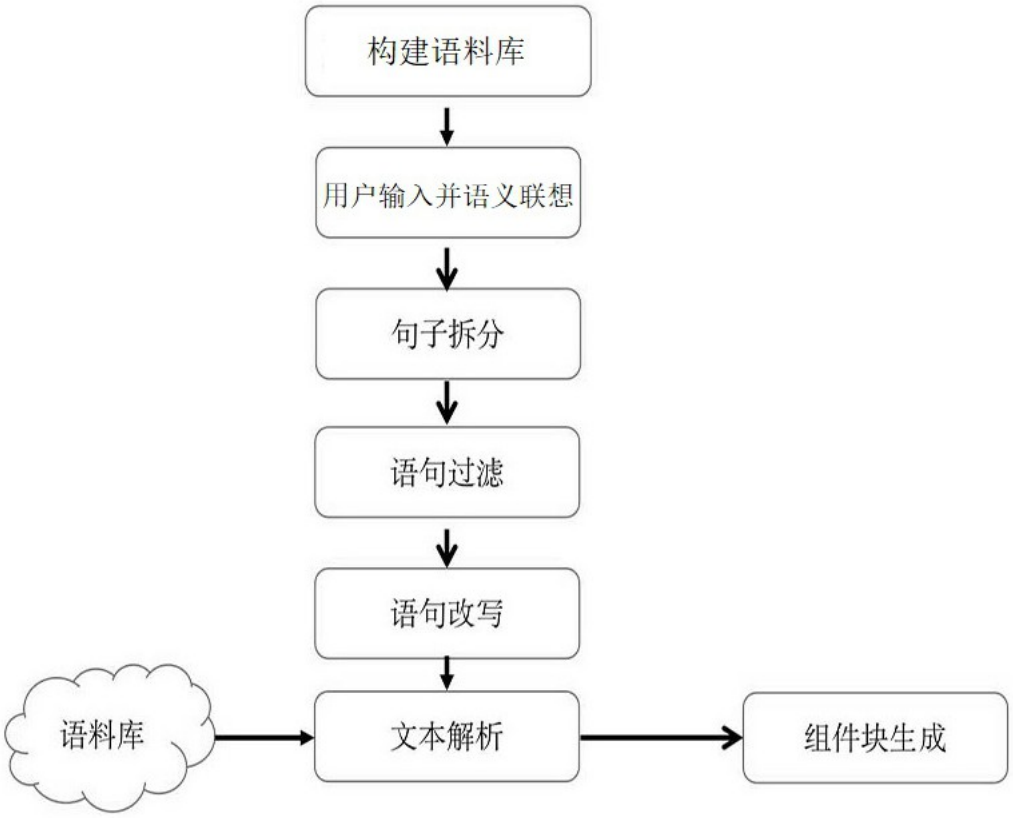


图1

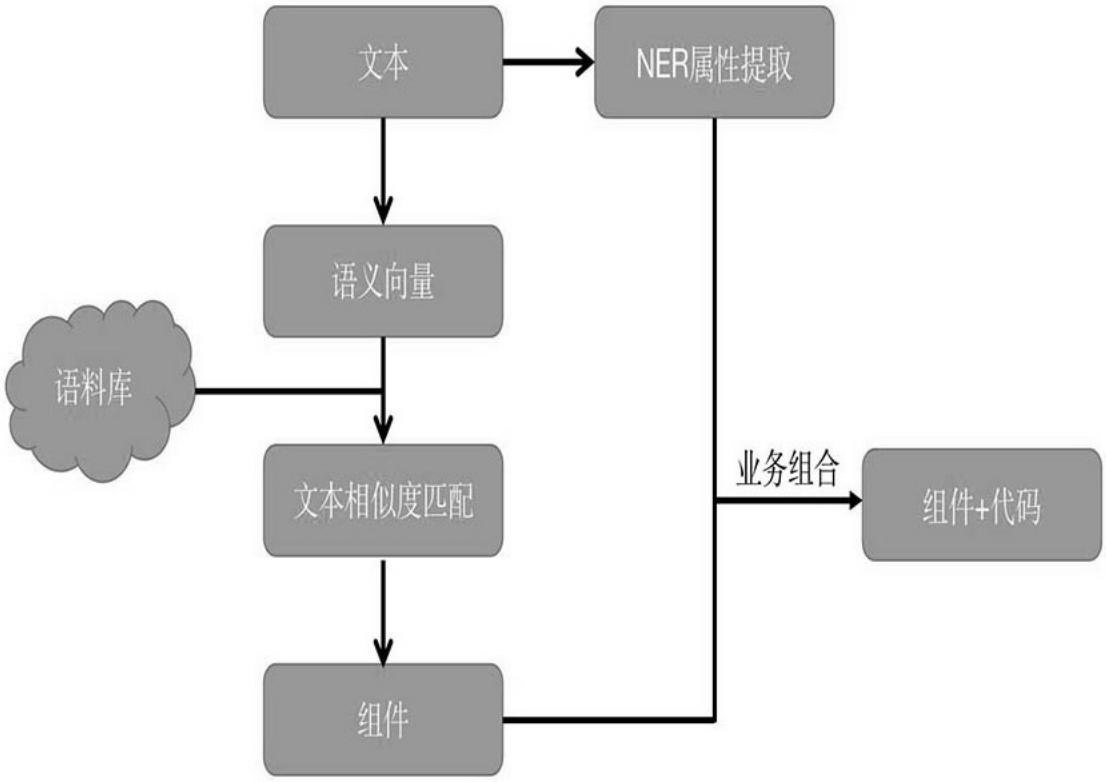

文本
NER属性提取
语义向量
语料库
文本相似度匹配
业务组合
组件+代码
组件

图2

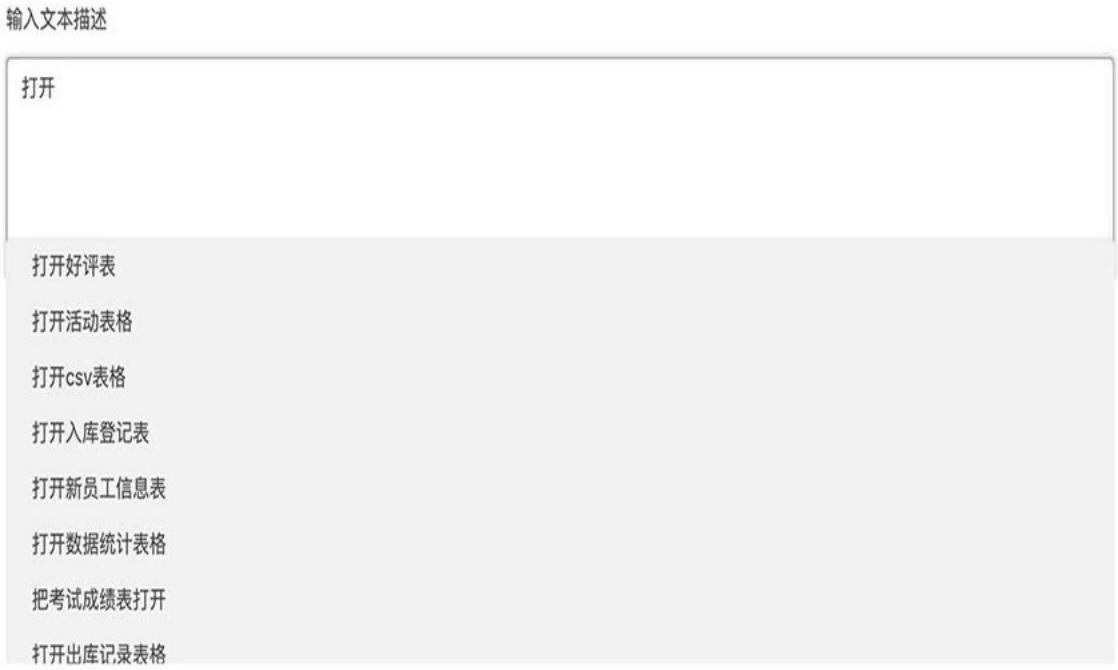

输入文本描述
打开
打开好评表
打开活动表格
打开csv表格
打开入库登记表
打开新员工信息表
打开数据统计表格
把考试成绩表打开
打开出库记录表格

图3

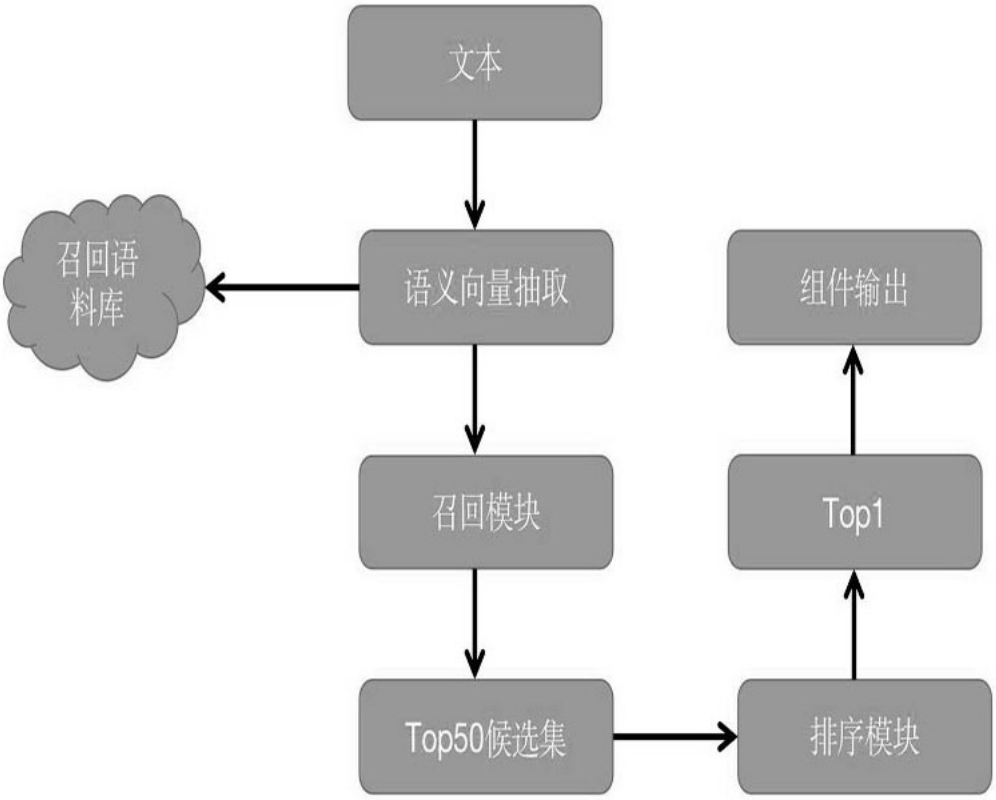


图4