(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 114118077 A
(43)申请公布日 2022.03.01

(21)申请号 202111389734.X

(22)申请日 2021.11.22

(71)申请人 深圳深度赋智科技有限公司
地址 518000 广东省深圳市南山区粤海街道科技园社区科苑路8号讯美科技广场1号楼815

(72)发明人 曹勇 吴承霖 张杨 陈焕坤

(74)专利代理机构 北京知果之信知识产权代理有限公司 11541
代理人 高科

(51)Int.Cl.
G06F 40/284(2020.01)
G06N 5/02(2006.01)
G06N 20/00(2019.01)

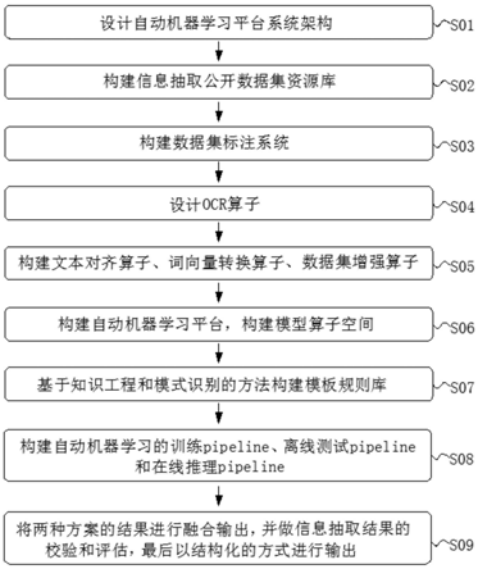
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种基于自动机器学习平台的智能信息抽取系统构建方法

(57)摘要

本发明涉及自然语言处理的文档信息抽取技术领域,公开了一种基于自动机器学习平台的智能信息抽取系统构建方法,基于知识工程的方法和自动机器学习平台复合抽取的方法来完成信息抽取的任务,基于自动机器学习平台完成模型算子的自动选择,自动对用户的输入进行预处理、建模分析、标准输出和服务提供,同时,基于知识工程的方法用于对特定字段的抽取,自动机器学习平台极大地优化现有基于深度学习的信息抽取效果,而基于传统的知识工程的方法提升信息抽取的覆盖度和不同场景的抽取效果,通过综合两类抽取方法,对于文档的结构信息、上下文信息、特殊信息都能够有更加全面的定位和认知。



CN 114118077 A

1.一种基于自动机器学习平台的智能信息抽取系统构建方法,其特征在于,包括以下步骤:

S01:设计自动机器学习平台系统架构,包括前端设计、算法设计、后台部署三个板块;

S02:构建信息抽取公开数据集资源库,同时融合用户提供的数据集形成增强数据集;

S03:构建数据集标注系统,用户对需要抽取的部分重要信息进行标注,将标注好的数据作为训练数据;

S04:设计OCR算子,实现多种类型文档的智能解析,转换为信息抽取系统可用的数据集格式;

S05:构建文本对齐算子、词向量转换算子、数据集增强算子,对数据集进行预处理和数据增强;

S06:构建自动机器学习平台,基于主流的bert类算子、bilstm算子、crf算子等构建模型算子空间,设计完备的算子超参数搜索空间,通过实验论证各参数的最优取值范围;

S07:基于知识工程和模式识别的方法构建模板规则库,从传统的信息抽取方法上实现抽取流程;

S08:构建自动机器学习的训练pipeline、离线测试pipeline和在线推理pipeline,同时完成微服务部署;

S09:将两种方案的结果进行融合输出,并做信息抽取结果的校验和评估,最后以结构化的方式进行输出。

2.根据权利要求1所述的一种基于自动机器学习平台的智能信息抽取系统构建方法,其特征在于,步骤S01中,设计自动机器学习平台的UI界面,包括登录注册、上传数据、选择项目类型、构建任务、自动配置模型参数组合、自动构建模型算子组合、执行训练pipeline、执行离线测试pipeline、部署在线服务pipeline、配置数据导出模型、查看解决方案等功能模块。

3.根据权利要求1所述的一种基于自动机器学习平台的智能信息抽取系统构建方法,其特征在于,步骤S02中,增强后的数据集按照一定的比例进行训练集、测试集的划分,且训练集不仅用于深度学习模型的训练,同时也输入到专家系统用于特征分析、模板构建和规则制定。

4.根据权利要求1所述的一种基于自动机器学习平台的智能信息抽取系统构建方法,其特征在于,步骤S03中,提供用户标注的平台,用户直接上传无标签数据,同时通过标注平台进行智能标注,同时优化标注平台的操作流程、标注效率,实现同类信息自动标注、相关信息推荐标注。

5.根据权利要求1所述的一种基于自动机器学习平台的智能信息抽取系统构建方法,其特征在于,步骤S06中,构建完备的超参数搜索空间和模型算子空间,每次试验通过优化算法自动选择一组解决方案进行训练,得到训练结果后再调整解决方案的算子选择,不断迭代得到最优模型。

6.根据权利要求5所述的一种基于自动机器学习平台的智能信息抽取系统构建方法,其特征在于,模型训练过程中,超参数的定义方式为,定义一个全范围的搜索空间,包括学习率、迭代轮次、批处理大小、分字策略、数据集划分比例,在这个空间中,每一次实验就按照一定的优化策略对每一类超参数确定一个取值,去迭代模型,得到模型结果后,模型选择

一个更好的解决方案的值。

7.根据权利要求1所述的一种基于自动机器学习平台的智能信息抽取系统构建方法，其特征在于，步骤S07中，根据特征工程的方法对数据集进行分析，总结出信息抽取的规则集合，按照集合去抽取对应字段的信息，同时对抽取结果进行评估，调整规则集合，对规则集合中的元素进行增加、删除和修改，以迭代模型和优化抽取效果。

8.根据权利要求7所述的一种基于自动机器学习平台的智能信息抽取系统构建方法，其特征在于，基于自然语言文本中的模式识别和模式匹配方法从海量文本中抽取不同种类的信息，不局限于使用单一模式进行信息抽取，基于深度学习模型和模板规则同时进行抽取，对不同字段涉及不同的抽取方案，最终将抽取结果进行汇总，作为最终输出。

9.根据权利要求8所述的一种基于自动机器学习平台的智能信息抽取系统构建方法，其特征在于，对于模板规则的方法进行不断迭代，每一轮迭代都需要对抽取效果进行评估后，根据指标结果进行动态调整；

信息抽取模型的指标定义为精确率、召回率和F1值三类，其中精确率是信息抽取正确的字段和所有抽取到的字段数的比率，召回率是指抽取正确的字段和所有抽取正确的字段的比率；

为了同时考虑查全率和查准率，引入F1值指标，F1值定义为正确率和召回率的调和平均值，其计算公式为：

F1值＝正确率*召回率*2/(正确率+召回率)。

10.根据权利要求9所述的一种基于自动机器学习平台的智能信息抽取系统构建方法，其特征在于，信息抽取后，对抽取效果进行校验，添加多重校验机制，通过校验算子对抽取结果进行格式化整理和校验，允许用户在线校验抽取结果，记录并保存抽取正确的字段用以迭代算法模型，优化抽取效果。

一种基于自动机器学习平台的智能信息抽取系统构建方法

技术领域

[0001] 本发明涉及自然语言处理的文档信息抽取技术领域，具体为一种基于自动机器学习平台的智能信息抽取系统构建方法。

背景技术

[0002] 信息抽取是指从海量的自然语言语料库中，抽取出特定的事件或事实信息，对海量文档中的内容实现自动分类、重要信息提取、生成摘要信息和重构文本结构等。随着自然语言处理技术的不断突破和发展，信息抽取技术已经在众多领域解决了具有基础性的地位，可以较好地解决文本、信息、知识获取、知识加工、文档组织、企业管理等应用场景中的文本处理问题。

[0003] 目前的信息抽取按照建模过程的差异，可大致分为三种：一种是基于知识工程的方法，借助于专家对于文本语料库的认知和分析，人工制作模板和规则去匹配海量文本以实现信息抽取。这种方法的缺点是需要耗费大量的时间成本和人力成本，复用性不高，无法处理新的字段信息。第二种基于传统的机器学习方法，通过机器学习方法(例如隐马尔科夫模型、LSTM模型等)来推导抽取规则和抽取方式，例如中英文人命的抽取、地名的抽取等，具备一定的泛化性，但是抽取性能较差，无法实现多字段的抽取和上下文的理解。第三种方法是基于深度神经网络来实现抽取，基于大规模训练语料、预训练模型、深度神经网络来训练一个泛化性较好的抽取器，例如Bert模型、Transformer模型等。这种方法能够在前两种方法中取得平衡，既提升抽取效果，又能够降低人工成本，但是存在算法复杂度较高、抽取效果仍然有限的缺点。

[0004] 针对上述问题，本发明提供了一种基于自动机器学习平台的智能信息抽取系统构建方法。

发明内容

[0005] 本发明的目的在于提供一种基于自动机器学习平台的智能信息抽取系统构建方法，自动机器学习平台极大地优化现有基于深度学习的信息抽取效果，而基于传统的知识工程的方法提升信息抽取的覆盖度和不同场景的抽取效果，通过综合两类抽取方法，对于文档的结构信息、上下文信息、特殊信息都能够有更加全面的定位和认知，从而解决了背景技术中的问题。

[0006] 为实现上述目的，本发明提供如下技术方案：一种基于自动机器学习平台的智能信息抽取系统构建方法，包括以下步骤：

[0007] S01：设计自动机器学习平台系统架构，包括前端设计、算法设计、后台部署三个板块；

[0008] S02：构建信息抽取公开数据集资源库，同时融合用户提供的数据集形成增强数据集；

[0009] S03：构建数据集标注系统，用户可对需要抽取的部分重要信息进行标注，将标注

好的数据作为训练数据；

[0010] S04：设计OCR算子，实现多种类型文档的智能解析，转换为信息抽取系统可用的数据集格式；

[0011] S05：构建文本对齐算子、词向量转换算子、数据集增强算子，对数据集进行预处理和数据增强；

[0012] S06：构建自动机器学习平台，基于主流的bert类算子、bilstm算子、crf算子等构建模型算子空间，设计完备的算子超参数搜索空间，通过实验论证各参数的最优取值范围；

[0013] S07：基于知识工程和模式识别的方法构建模板规则库，从传统的信息抽取方法上实现抽取流程；

[0014] S08：构建自动机器学习的训练pipeline、离线测试pipeline和在线推理pipeline，同时完成微服务部署；

[0015] S09：将两种方案的结果进行融合输出，并做信息抽取结果的校验和评估，最后以结构化的方式进行输出。

[0016] 进一步地，步骤S01中，设计自动机器学习平台的UI界面，包括登录注册、上传数据、选择项目类型、构建任务、自动配置模型参数组合、自动构建模型算子组合、执行训练pipeline、执行离线测试pipeline、部署在线服务pipeline、配置数据导出模型、查看解决方案等功能模块。

[0017] 进一步地，步骤S02中，增强后的数据集按照一定的比例进行训练集、测试集的划分，且训练集不仅用于深度学习模型的训练，同时也输入到专家系统用于特征分析、模板构建和规则制定。

[0018] 进一步地，步骤S03中，提供用户标注的平台，用户直接上传无标签数据，同时通过标注平台进行智能标注，同时优化标注平台的操作流程、标注效率，实现同类信息自动标注、相关信息推荐标注。

[0019] 进一步地，步骤S06中，构建完备的超参数搜索空间和模型算子空间，每次试验通过优化算法自动选择一组解决方案进行训练，得到训练结果后再调整解决方案的算子选择，不断迭代得到最优模型。

[0020] 进一步地，模型训练过程中，超参数的定义方式为，定义一个全范围的搜索空间，包括学习率、迭代轮次、批处理大小、分字策略、数据集划分比例，在这个空间中，每一次实验就按照一定的优化策略对每一类超参数确定一个取值，去迭代模型，得到模型结果后，模型选择一个更好的解决方案的值。

[0021] 进一步地，步骤S07中，根据特征工程的方法对数据集进行分析，总结出信息抽取的规则集合，按照集合去抽取对应字段的信息，同时对抽取结果进行评估，调整规则集合，对规则集合中的元素进行增加、删除和修改，以迭代模型和优化抽取效果。

[0022] 进一步地，基于自然语言文本中的模式识别和模式匹配方法从海量文本中抽取不同种类的信息，不局限于使用单一模式进行信息抽取，基于深度学习模型和模板规则同时进行抽取，对不同字段涉及不同的抽取方案，最终将抽取结果进行汇总，作为最终输出。

[0023] 进一步地，对于模板规则的方法进行不断迭代，每一轮迭代都需要对抽取效果进行评估后，根据指标结果进行动态调整；

[0024] 信息抽取模型的指标定义为精确率、召回率和F1值三类，其中精确率是信息抽取

正确的字段和所有抽取到的字段数的比率，召回率是指抽取正确的字段和所有抽取正确的字段的比率；

[0025] 为了同时考虑查全率和查准率，引入F1值指标，F1值定义为正确率和召回率的调和平均值，其计算公式为：

[0026] F1值＝正确率*召回率*2/(正确率+召回率)。

[0027] 进一步地，信息抽取后，对抽取效果进行校验，添加多重校验机制，通过校验算子对抽取结果进行格式化整理和校验，允许用户在线校验抽取结果，记录并保存抽取正确的字段用以迭代算法模型，优化抽取效果。

[0028] 与现有技术相比，本发明的有益效果如下：

[0029] 1、本发明提供的一种基于自动机器学习平台的智能信息抽取系统构建方法，结合了当前自然语言处理中各种场景下信息抽取任务最经典的模型算法进行建模分析，采用了所有的经典网络结构算法构建搜索空间，将目前主流的词向量转换模型均进行构建，动态地调整网络中的各参数值，实现自动优化，而不需要人工手动调参，极大地解决人力成本的问题，根据任务不断进行搜索空间的探索，得到最优的算子搜索空间定义，在完成初步的搜索空间定义后，再逐步优化，较好地解决了信息抽取效果不佳的问题。

[0030] 2、本发明提供的一种基于自动机器学习平台的智能信息抽取系统构建方法，基于知识工程的方法和自动机器学习平台复合抽取的方法来完成信息抽取的任务，基于自动机器学习平台完成模型算子的自动选择，自动对用户的输入进行预处理、建模分析、标准输出和服务提供，同时，基于知识工程的方法用于对特定字段的抽取，自动机器学习平台极大地优化现有基于深度学习的信息抽取效果，而基于传统的知识工程的方法提升信息抽取的覆盖度和不同场景的抽取效果，通过综合两类抽取方法，对于文档的结构信息、上下文信息、特殊信息都能够有更加全面的定位和认知。

附图说明

[0031] 构成本申请的一部分的附图用来提供对本申请的进一步理解，使得本申请的其它特征、目的和优点变得更明显。本申请的示意性实施例附图及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

[0032] 图1为本发明的基于自动机器学习平台的智能信息抽取系统构建方法的整体流程图；

[0033] 图2为本发明的基于自动机器学习平台的智能信息抽取算法模块图；

[0034] 图3为本发明的基于自动机器学习平台的智能信息抽取系统的流程图。

具体实施方式

[0035] 为了使本技术领域的人员更好地理解本申请方案，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清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分的实施例，而不是全部的实施例。基于本申请中的实施例，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，都应当属于本申请保护的范围。

[0036] 需要说明的是，本申请的说明书和权利要求书及上述附图中的术语“第一”、“第

二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0037] 在本申请中，术语“上”、“下”、“左”、“右”、“前”、“后”、“顶”、“底”、“内”、“外”、“中”、“竖直”、“水平”、“横向”、“纵向”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。这些术语主要是为了更好地描述本申请及其实施例，并非用于限定所指示的装置、元件或组成部分必须具有特定方位，或以特定方位进行构造和操作。

[0038] 并且，上述部分术语除了可以用于表示方位或位置关系以外，还可能用于表示其他含义，例如术语“上”在某些情况下也可能用于表示某种依附关系或连接关系。对于本领域普通技术人员而言，可以根据具体情况理解这些术语在本申请中的具体含义。

[0039] 另外，术语“多个”的含义应为两个以及两个以上。

[0040] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0041] 请参阅图1，一种基于自动机器学习平台的智能信息抽取系统构建方法，包括以下步骤：

[0042] S01：设计自动机器学习平台系统架构，包括前端设计、算法设计、后台部署三个板块；

[0043] S02：构建信息抽取公开数据集资源库，同时融合用户提供的数据集形成增强数据集；

[0044] S03：构建数据集标注系统，用户可对需要抽取的部分重要信息进行标注，将标注好的数据作为训练数据；

[0045] S04：设计OCR算子，实现多种类型文档的智能解析，转换为信息抽取系统可用的数据集格式；

[0046] S05：构建文本对齐算子、词向量转换算子、数据集增强算子，对数据集进行预处理和数据增强；

[0047] S06：构建自动机器学习平台，基于主流的bert类算子、bilstm算子、crf算子等构建模型算子空间，设计完备的算子超参数搜索空间，通过实验论证各参数的最优取值范围；

[0048] S07：基于知识工程和模式识别的方法构建模板规则库，从传统的信息抽取方法上实现抽取流程；

[0049] S08：构建自动机器学习的训练pipeline、离线测试pipeline和在线推理pipeline，同时完成微服务部署；

[0050] S09：将两种方案的结果进行融合输出，并做信息抽取结果的校验和评估，最后以结构化的方式进行输出。

[0051] 设计自动机器学习平台的UI界面，包括登录注册、上传数据、选择项目类型、构建任务、自动配置模型参数组合、自动构建模型算子组合、执行训练pipeline、执行离线测试pipeline、部署在线服务pipeline、配置数据导出模型、查看解决方案等功能模块。

[0052] 获取用户标注数据集后，对数据进行基于特征工程方法的分析，分析数据集的特

性和分布规律，研究文档的特性，对异常数据集进行清洗和增强。

[0053] 将增强后的数据集按照一定的比例进行训练集、测试集的划分，作为本发明的进一步改进，训练集不仅用于深度学习模型的训练，同时也输入到专家系统用于特征分析、模板构建和规则制定。

[0054] 构建文本对齐算子，按照用户的不同业务需求将文本扩充为等长的输入，同时针对不同的语种构建自适应语料分割算子，例如对于中文文本按照字符进行分割，对于英文文本按照单词进行分割，最后输出标准、等长的文本内容。

[0055] 构建数据集增强算子，对数据集中分布不均衡的样本和异常样本进行增强和过滤，保证数据集样本的均匀分布，基于开源词向量模型构建输入标准化算子，输入到模型中。将数据集从原始文本转换为标准的字符id和对应的词向量矩阵。

[0056] 构建模型算子空间，将主流的信息抽取的模型均编入自动机器学习平台，模型算子包括Bilstm算子、CRF算子、RNN算子、RNN-CNN算子、Bert算子等，作为本发明的进一步改进，本方法不局限于使用单一算子进行建模，而是，构建大规模模型算子空间，执行每次任务使用多次实验，每次随机选择一个算子进行建模分析，得到一个解决方案。多次实验后选择最优的模型提交给用户，用户就基于最优的模型获得信息抽取服务。

[0057] 对于Bert模型，为了进一步提升模型性能，需要从开源网站下载Bert模型及其改进版的模型的网络结构、词汇词典、网络权重参数、分词配置文件进行分词、网络结构和权重初始化的步骤。根据任务场景的不同，选择不同类型和不同大小的预训练模型。

[0058] 实现基于模板、规则的信息抽取自动化和标准化流程，对训练集进行分析后，得到信息分布规律和抽取的方法，基于开源的工具包制定规则集合。

[0059] 同时，基于自然语言文本中的模式识别和模式匹配方法从海量文本中抽取不同种类的信息。作为本发明的进一步改进，本方法不局限于使用单一模式进行信息抽取，而是，基于深度学习模型和模板规则同时进行抽取，对不同字段涉及不同的抽取方案，最终将抽取结果进行汇总，作为最终输出。

[0060] 对于模板规则的方法进行不断迭代，每一轮迭代都需要对抽取效果进行评估后，根据指标结果进行动态调整。

[0061] 信息抽取模型的指标定义为精确率、召回率和F1值三类。精确率和召回率是广泛用于信息抽取和统计学分类领域的两个度量值，用来评价结果的质量。其中精确率是信息抽取正确的字段和所有抽取到的字段数的比率；召回率是指抽取正确的字段和所有抽取正确的字段的比率。为了同时考虑查全率和查准率，引入F1值指标，F1值定义为正确率和召回率的调和平均值，其计算公式为：F1值＝正确率*召回率*2/(正确率+召回率)。

[0062] 模型训练过程中，需要定义各类不同的超参数进行迭代。作为本发明的进一步改进，本方法不再使用人工方式调整超参数和网络结构，而是，定义一个全范围的搜索空间，包括学习率、迭代轮次、批处理大小、分字策略、数据集划分比例等。在这个空间中，每一次实验就按照一定的优化策略对每一类超参数确定一个取值，去迭代模型。得到模型结果后，模型需要选择一个更好的解决方案的值。

[0063] 信息抽取任务划分为普通字段抽取和特定字段抽取。普通字段包括人名、地名、金额、时间、日期。特定字段需要根据场景进行定义，例如简历抽取包括教育背景、公司名、技能点、兴趣爱好等。合同抽取包括项目名称、项目标的、合同名称、合同风险等。作为本发明

的进一步改进，本方法不局限于传统模式的信息抽取，而是，对抽到的信息和对应的类别添加关系判断算子，判断该信息是否为该字段的正确信息，如果是正确抽取，则结构化保存数据；反之则重新抽取或放弃该字段，以保证抽取效果。

[0064] 平台需要对抽取效果进行校验，作为本发明的进一步改进，本方法不局限于传统模式的主观评判方式，而是，平台添加多重校验机制，通过校验算子对抽取结果进行格式化整理和校验，同时允许用户在线校验抽取结果。记录并保存抽取正确的字段用以迭代算法模型，优化抽取效果。

[0065] 综上所述：本发明提供了一种基于自动机器学习平台的智能信息抽取系统构建方法，结合了当前自然语言处理中各种场景下信息抽取任务最经典的模型算法进行建模分析，采用了所有的经典网络结构算法构建搜索空间，将目前主流的词向量转换模型均进行构建，动态地调整网络中的各参数值，实现自动优化，而不需要人工手动调参，极大地解决人力成本的问题，根据任务不断进行搜索空间的探索，得到最优的算子搜索空间定义，在完成初步的搜索空间定义后，再逐步优化，较好地解决了信息抽取效果不佳的问题；基于知识工程的方法和自动机器学习平台复合抽取的方法来完成信息抽取的任务，基于自动机器学习平台完成模型算子的自动选择，自动对用户的输入进行预处理、建模分析、标准输出和服务提供，同时，基于知识工程的方法用于对特定字段的抽取，自动机器学习平台极大地优化现有基于深度学习的信息抽取效果，而基于传统的知识工程的方法提升信息抽取的覆盖度和不同场景的抽取效果，通过综合两类抽取方法，对于文档的结构信息、上下文信息、特殊信息都能够有更加全面的定位和认知。

[0066] 需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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。

[0067] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

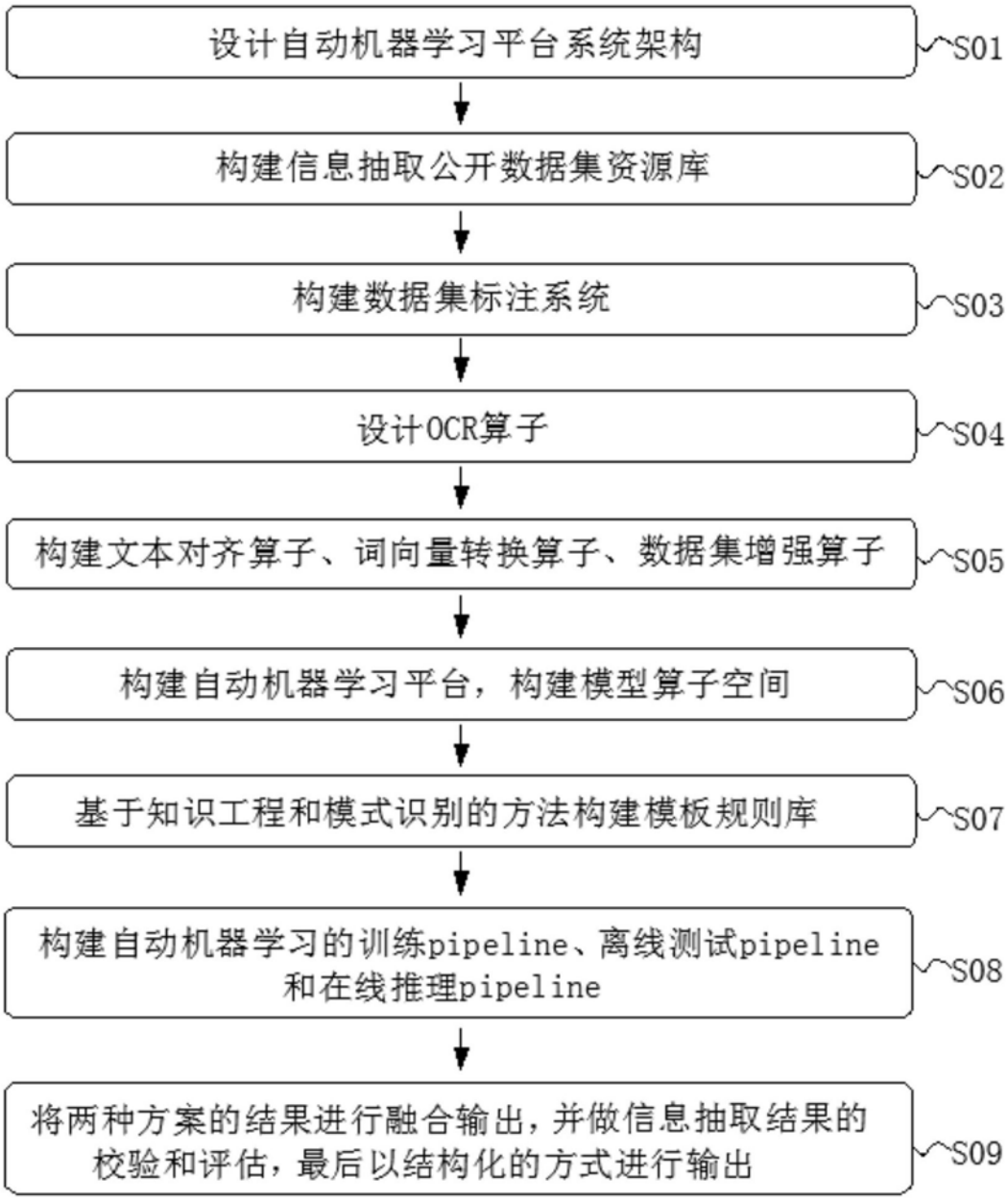

设计自动机器学习平台系统架构
S01
构建信息抽取公开数据集资源库
S02
构建数据集标注系统
S03
设计OCR算子
S04
构建文本对齐算子、词向量转换算子、数据集增强算子
S05
构建自动机器学习平台，构建模型算子空间
S06
基于知识工程和模式识别的方法构建模板规则库
S07
构建自动机器学习的训练pipeline、离线测试pipeline和在线推理pipeline
S08
将两种方案的结果进行融合输出，并做信息抽取结果的校验和评估，最后以结构化的方式进行输出
S09

图1

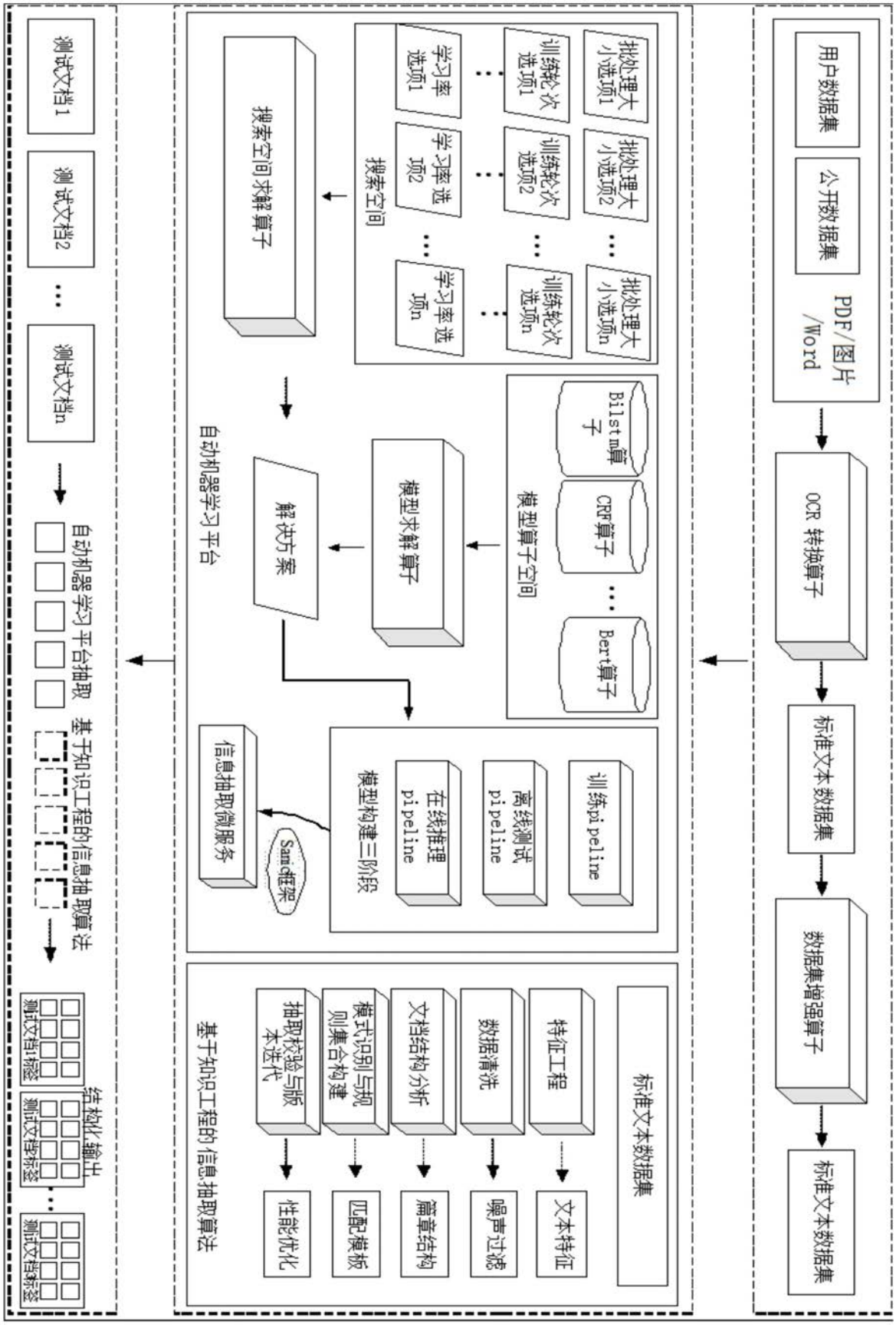

用户数据集
公开数据集
PDF/图片/Word
OCR 转换算子
标准文本数据集
数据集增强算子
标准文本数据集
批处理大小选项1
批处理大小选项2
批处理大小选项n
训练轮次选项1
训练轮次选项2
训练轮次选项n
学习率选项1
学习率选项2
学习率选项n
搜索空间
搜索空间求解算子
Bilstm算子
CRF算子
Bert算子
模型算子空间
模型求解算子
解决方案
训练pipeline
离线测试pipeline
在线推理pipeline
模型构建三阶段
Sanic框架
信息抽取微服务
自动机器学习平台
标准文本数据集
特征工程
数据清洗
文档结构分析
模式识别与规则集合构建
抽取校验与版本迭代
文本特征
噪声过滤
篇章结构
匹配模板
性能优化
基于知识工程的信息抽取算法
测试文档1
测试文档2
测试文档n
自动机器学习平台抽取
基于知识工程的信息抽取算法
结构化输出
测试文档1标签
测试文档2标签
测试文档n标签

图2

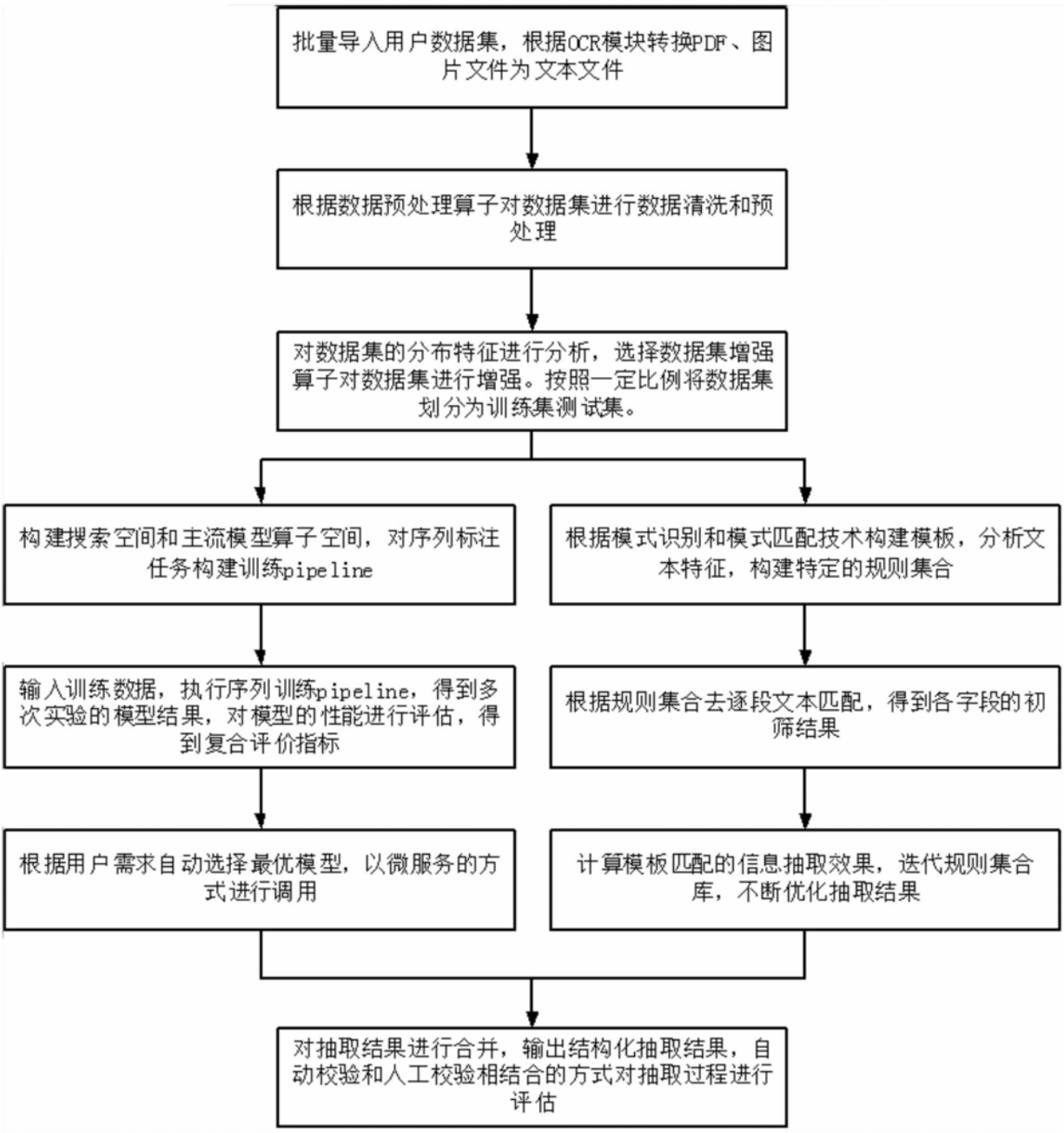

批量导入用户数据集，根据OCR模块转换PDF、图片文件为文本文件
根据数据预处理算子对数据集进行数据清洗和预处理
对数据集的分布特征进行分析，选择数据集增强算子对数据集进行增强。按照一定比例将数据集划分为训练集测试集。
构建搜索空间和主流模型算子空间，对序列标注任务构建训练pipeline
根据模式识别和模式匹配技术构建模板，分析文本特征，构建特定的规则集合
输入训练数据，执行序列训练pipeline，得到多次实验的模型结果，对模型的性能进行评估，得到复合评价指标
根据规则集合去逐段文本匹配，得到各字段的初筛结果
根据用户需求自动选择最优模型，以微服务的方式进行调用
计算模板匹配的信息抽取效果，迭代规则集合库，不断优化抽取结果
对抽取结果进行合并，输出结构化抽取结果，自动校验和人工校验相结合的方式对抽取过程进行评估

图3