(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110795543 B
(45)授权公告日 2023.09.22

(21)申请号 201910828781.6

(22)申请日 2019.09.03

(65)同一申请的已公布的文献号
申请公布号 CN 110795543 A

(43)申请公布日 2020.02.14

(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 周辉阳

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
专利代理师 祝亚男

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/36(2019.01)

(56)对比文件
CN 109271506 A,2019.01.25
CN 109885660 A,2019.06.14
Xiao Huang, Jingyuan Zhang, Dingcheng Li, Ping Li.Knowledge Graph Embedding Based Question Answering.《Proceedings of the Twelfth ACM International Conference on Web Search and Data MiningJanuary》.2019,第2页第2章到第8页第6章.

审查员 蔡智勇

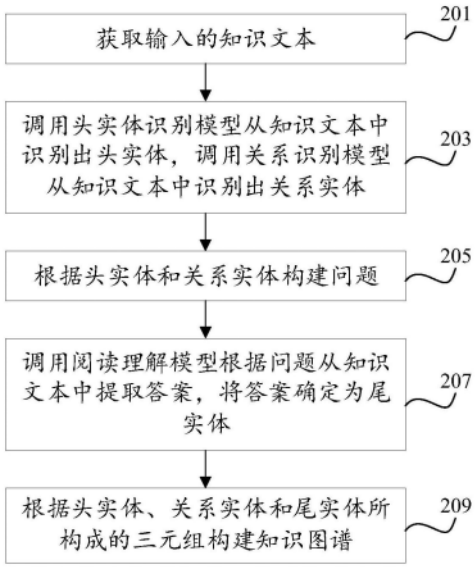
权利要求书2页 说明书15页 附图10页

(54)发明名称

基于深度学习的非结构化数据抽取方法、装置及存储介质

(57)摘要

本申请公开了一种基于深度学习的非结构化数据抽取方法,所述方法应用于人工智能的自然语言处理领域,所述方法包括:获取输入的知识文本;调用头实体识别模型从所述知识文本中识别出头实体,调用关系识别模型从所述知识文本中识别出关系实体;根据所述头实体和所述关系实体构建问题;调用阅读理解模型根据所述问题从所述知识文本中提取答案,将所述答案确定为尾实体;根据所述头实体、所述关系实体和所述尾实体所构成的三元组构建知识图谱。该方法实现了自动构建问题以调用阅读理解模型进行自动化的非结构化数据提取,从而实现自动化的非结构化数据提取的效果。



CN 110795543 B

1.一种基于深度学习的非结构化数据抽取方法,其特征在于,所述方法包括:

获取输入的知识文本;

调用头实体识别模型从所述知识文本中识别出头实体,调用关系识别模型从所述知识文本中识别出关系实体;

根据所述头实体和所述关系实体构建问题;

获取所述问题对应的所述头实体的第一词向量和所述关系实体的第二词向量;

根据所述第一词向量,在知识图谱的已有三元组中确定出候选三元组;

根据所述第一词向量和所述第二词向量,在所述候选三元组中确定出目标实体;

当所述目标实体不满足真实性条件时,调用阅读理解模型根据所述问题从所述知识文本中提取答案,将所述答案确定为尾实体;

当所述目标实体满足所述真实性条件时,将所述知识文本进行向量化,得到每个句子的词向量序列;

计算所述目标实体的第三词向量和所述每个句子的词向量序列之间的相似度;

从所述相似度最高的句子中提取出所述答案,将所述答案确定为尾实体;

根据所述头实体、所述关系实体和所述尾实体所构成的三元组构建知识图谱。

2.根据权利要求1所述的方法,其特征在于,所述根据所述头实体和所述关系实体构建问题,包括:

确定所述头实体的第一实体类型和所述关系实体的第二实体类型;

从多个候选问题模板中,确定与所述第一实体类型和所述第二实体类型对应的问题模板;

将所述头实体和所述关系实体按照所述问题模板进行组合,得到所述问题。

3.根据权利要求2所述的方法,其特征在于,所述头实体或所述关系实体为至少两个;

所述方法还包括:

根据至少两个所述头实体或所述关系实体的排列组合,拆解得到至少两组所述头实体和所述关系实体之间的一对一组合。

4.根据权利要求1所述的方法,其特征在于,所述根据所述第一词向量,在知识图谱的已有三元组中确定出候选三元组,包括:

遍历所述知识图谱的已有三元组,确定出头实体等于所述第一词向量的三元组作为所述候选三元组;

或,

遍历所述知识图谱的已有三元组,确定出头实体包括所述第一词向量的三元组作为所述候选三元组。

5.根据权利要求1所述的方法,其特征在于,所述根据所述第一词向量和所述第二词向量,在所述候选三元组中确定出目标实体,包括:

根据所述第一词向量和所述第二词向量计算预测向量;

计算所述预测向量和所述候选三元组对应的标签向量之间的距离,将所述距离最小的候选三元组确定为所述目标实体。

6.根据权利要求1至5任一所述的方法,所述方法还包括:

当所述目标实体满足所述真实性条件时,计算所述目标实体和所述知识文本中的每个

句子的最长公共子序列；

从具有最长的所述最长公共子序列的句子中提取出所述答案。

7.一种基于深度学习的非结构化数据抽取装置，其特征在于，所述装置包括：

获取模块，用于获取输入的知识文本；

调用模块，用于调用头实体识别模型、关系识别模型；

识别模块，用于在调用头实体识别模型后从所述知识文本中识别出头实体，在调用关系识别模型后从所述知识文本中识别出关系实体；

构建模块，用于根据所述头实体和所述关系实体构建问题；

所述获取模块，还用于获取所述问题对应的所述头实体的第一词向量和所述关系实体的第二词向量；

确定模块，用于根据所述第一词向量，在知识图谱的已有三元组中确定出候选三元组；根据所述第一词向量和所述第二词向量，在所述候选三元组中确定出目标实体；

判断模块，用于判断目标实体是否满足真实性条件；

提取模块，用于当所述目标实体不满足真实性条件时，调用阅读理解模型根据所述问题从所述知识文本中提取答案，将所述答案确定为尾实体；

所述确定模块，还用于当所述目标实体满足所述真实性条件时，将所述知识文本进行向量化，得到每个句子的词向量序列；计算所述目标实体的第三词向量和所述每个句子的词向量序列之间的相似度；从所述相似度最高的句子中提取出所述答案，将所述答案确定为尾实体；

所述构建模块，还用于根据所述头实体、所述关系实体和所述尾实体所构成的三元组构建知识图谱。

8.根据权利要求7所述的装置，其特征在于，所述构建模块还包括：确定子模块和组合子模块；

所述确定子模块，用于确定所述头实体的第一实体类型和所述关系实体的第二实体类型；从多个候选问题模板中，确定与所述第一实体类型和所述第二实体类型对应的问题模板；

所述组合子模块，用于将所述头实体和所述关系实体按照所述问题模板进行组合，得到所述问题。

9.根据权利要求8所述的装置，其特征在于，所述头实体或所述关系实体为至少两个；

所述装置还包括拆解模块；

所述拆解模块，用于根据至少两个所述头实体或所述关系实体的排列组合，拆解得到至少两组所述头实体和所述关系实体之间的一对一组合。

10.一种计算机设备，所述计算机设备包括：处理器和存储器，所述存储器中存储有至少一段程序，所述至少一段程序由所述处理器加载并执行，以实现如权利要求1至6任一项所述的基于深度学习的非结构化数据抽取方法。

11.一种计算机可读存储介质，其特征在于，所述存储介质中存储有至少一段程序，所述至少一段程序由处理器加载并执行，以实现如权利要求1至6任一项所述的基于深度学习的非结构化数据抽取方法。

基于深度学习的非结构化数据抽取方法、装置及存储介质

技术领域

[0001] 本申请涉及人工智能的自然语言处理领域，特别涉及一种基于深度学习的非结构化数据抽取方法、装置及存储介质。

背景技术

[0002] 人工智能技术是一门综合学科，自然语言处理(Natural Language Processing，NLP)是人工智能研究的一大方向，非结构化数据抽取是自然语言处理中的一个课题。该课题的主要目的是从一段长文本(比如句子、段落或短篇章级别)中抽取出客观的三元组信息。比如：“小明(Charles Aránguiz)，1989年4月17日出生于智利圣地亚哥”这句话中可以抽取的三元组信息如下：[小明-出生地-圣地亚哥，小明-出生日期-1989年4月17日，小明-国籍-智利]。

[0003] 相关技术中，采用深度学习的方法来进行非结构化数据抽取。比如，采用BERT模型去做阅读理解。BERT模型的工作原理包括：向BERT模型输入一个问题和一个答案文本，由BERT模型根据该问题在答案文本中尝试寻找答案。若BERT模型成功寻找到答案，则输出答案在答案文本中的起始字符位置和结束字符位置。

[0004] 但是针对纯开放式的非结构化数据抽取，只会给定一个或者几个段落，并没有问题给出，因此无法通过BERT模型完成知识抽取任务。

发明内容

[0005] 本申请实施例提供了一种基于深度学习的非结构化数据抽取方法、装置及存储介质，可以解决纯开放式的非结构化数据抽取，只会给定一个或者几个段落，并没有问题给出，因此无法通过BERT模型完成知识抽取任务的问题。所述技术方案如下：

[0006] 根据本申请的一个方面，提供了一种基于深度学习的非结构化数据抽取方法，所述方法包括：

[0007] 获取输入的知识文本；

[0008] 调用头实体识别模型从所述知识文本中识别出头实体，调用关系识别模型从所述知识文本中识别出关系实体；

[0009] 根据所述头实体和所述关系实体构建问题；

[0010] 调用阅读理解模型根据所述问题从所述知识文本中提取答案，将所述答案确定为尾实体；

[0011] 根据所述头实体、所述关系实体和所述尾实体所构成的三元组构建知识图谱。

[0012] 根据本申请的另一方面，提供了一种基于深度学习的非结构化数据抽取装置，所述装置包括：

[0013] 获取模块，用于获取输入的知识文本；

[0014] 调用模块，用于调用头实体识别模型、关系识别模型和阅读理解模型；

[0015] 识别模块，用于在调用头实体识别模型后从所述知识文本中识别出头实体，在调

用关系识别模型后从所述知识文本中识别出关系实体；

[0016] 构建模块，用于根据所述头实体和所述关系实体构建问题；根据所述头实体、所述关系实体和所述尾实体所构成的三元组构建知识图谱；

[0017] 提取模块，用于在调用阅读理解模型后根据所述问题从所述知识文本中提取答案；

[0018] 确定模块，用于将所述答案确定为尾实体；根据所述头实体、所述关系实体和所述尾实体所构成的三元组构建知识图谱。

[0019] 根据本申请的另一方面，提供了一种计算机设备，所述计算机设备包括：处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上方面所述基于深度学习的非结构化数据抽取方法。

[0020] 根据本申请的另一方面，提供了一种计算机可读存储介质，所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上方面所述的基于深度学习的非结构化数据抽取方法。

[0021] 本申请实施例提供的技术方案带来的有益效果至少包括：

[0022] 通过调用头实体模型从知识文本中识别出头实体，调用关系识别模型从知识文本中识别出关系实体，根据头实体和关系实体构建问题，利用构建的问题来调用阅读理解模型根据问题从知识文本中提取答案。解决了相关技术中的阅读理解模型无法直接应用于开放式的非结构化数据提取的问题，实现了自动构建问题以调用阅读理解模型进行自动化的非结构化数据提取，从而实现自动化的非结构化数据提取的效果。

附图说明

[0023] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，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。

[0024] 图1是本申请一个示例性实施例提供的服务器的实施环境框图；

[0025] 图2是本申请一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图；

[0026] 图3是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图；

[0027] 图4是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图；

[0028] 图5是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图；

[0029] 图6是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图；

[0030] 图7是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法

的流程图；

[0031] 图8是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图；

[0032] 图9是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的界面示意图；

[0033] 图10是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的界面示意图；

[0034] 图11是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的界面示意图；

[0035] 图12是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的界面示意图；

[0036] 图13是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的界面示意图；

[0037] 图14是本申请另一个示例性实施例提供的基于深度学习的非结构化数据抽取装置的框图；

[0038] 图15是本申请另一个示例性实施例提供的服务器的结构示意图。

具体实施方式

[0039] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0040] 首先对本申请实施例涉及的若干个名词进行简介：

[0041] 实体：指表示一个概念的基本单位。

[0042] 模板：具有扩展样例的通用句式。

[0043] 双向转换编码器(Bidirectional Encoder Representation from Transformer，BERT)：采用大规模无标注语料训练，获得具有一定阅读能力的神经网络模型。

[0044] 知识图谱(Knowledge Gragh)：在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。

[0045] 问题(Query)：用户的搜索语句，包含用户的语音、文字、图片输入。

[0046] TransE：TransE的直观含义，就是TransE基于实体和关系的分布式向量表示，将每个三元组实例(head，relation，tail)中的关系(relation)看做从头实体(head)到尾实体(tail)的翻译，通过不断调整h、r和t(head、relation和tail的向量)，使(h+r)尽可能与t相等，即h+r＝t。

[0047] 人工智能(Artificial Intelligence，AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

[0048] 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层

面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0049] 自然语言处理(Natural Language Processing,NLP)是人工智能的一个子领域。在NLP的研究领域中存在一个较为困难的课题:非结构化数据抽取。非结构化数据的主要目的是从一段长文本(比如句子、段落或短篇章级别)中抽取出客观的三元组信息。比如:“小明,1989年4月17日出生于智利圣地亚哥,智利职业足球运动员,司职中场,效力于德国足球俱乐部”这句话中可以抽取的三元组信息如下:[小明-出生地-圣地亚哥,小明-出生日期-1989年4月17日,小明-国籍-智利,小明-职业-足球运动员,小明-俱乐部-足球俱乐部]。大量的科研人员做了很多的努力,都没有合理的非结构化数据的抽取方式,或者只能针对特殊应用或者特定领域能达到一个高度,但是对于纯粹的开放式的任务的效果还是很差。

[0050] 本申请实施例提供了一种基于“知识图谱”和“机器学习”的自动化非结构化数据抽取模型(以下简称抽取模型)。该抽取模型的自动化数据抽取过程,包括三个阶段:

[0051] 1)问题提取阶段;

[0052] 在输入一段长文本后,让该抽取模型同时输出多个主语和谓语的起始位置,根据识别出来的主语和谓语分别转换成词向量(Embedding)的形式,再加上偏移位置信息的相对向量,就得到主语和谓语的词向量表示。根据主语和谓语的词向量表示,拆解出多个一对一的问题。

[0053] 2)基于“知识图谱”的非结构化数据抽取过程;

[0054] 针对每个一对一的问题,利用已知的“知识图谱”寻找答案的方式,获得该问题对应的候选答案。然后,根据候选答案在输入的长文本中寻找目标答案。

[0055] 若在输入的长文本中,找到该问题的目标答案符合客观答案条件,则结束流程;若在输入的长文本中,找到该问题的目标答案不符合客观答案条件,则进入下一阶段。

[0056] 3)基于“机器学习”的非结构化数据抽取过程。

[0057] 当基于知识图谱的非结构化数据抽取失败时,利用阅读理解模型在输入的长文本中寻找目标答案。同时,根据阅读理解模型所找到的目标答案,在“知识图谱”中增加三元组。

[0058] 在上述非结构化数据抽取方法的抽取过程结束后,抽取到的知识图谱可以用于实现基于机器学习的问答系统。比如,智能车载系统、智能音箱系统、智能景点讲解系统等等。

[0059] 参考图1,示出了本申请一个示例性实施例提供的计算机系统的结构示意图,该计算机系统包括终端120和服务器140。

[0060] 终端120与服务器140之间通过有线或者无线网络相互连接。

[0061] 可选地,终端120可以包括笔记本电脑、台式电脑、智能手机、平板电脑、智能音箱、智能机器人中的至少一种。

[0062] 终端120包括第一存储器和第一处理器。第一存储器中存储有第一程序;上述第一程序被第一处理器调用执行以实现基于机器学习的问题答复方法。第一存储器可以包括但不限于以下几种: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)、以及电可擦除只读存储器

(Electric Erasable Programmable Read-Only Memory,EEPROM)。

[0063] 第一处理器可以是一个或者多个集成电路芯片组成。可选地,第一处理器可以是通用处理器,比如,中央处理器(Central Processing Unit,CPU)或者网络处理器(Network Processor,NP)。可选地,第一处理器用于通过调用服务器140提供的问答模型144来实现本申请提供的基于机器学习的问题答复方法。

[0064] 可选地,终端120中包括显示器;显示器用于显示问题或者答案。

[0065] 可选地,终端120中包括麦克风;麦克风用于采集语音形式的问题。

[0066] 可选地,终端120中包括扬声器;扬声器用于播放语音形式的答案。

[0067] 服务器140包括第二存储器和第二处理器。第二存储器中存储有第二程序,上述第二程序被第二处理器调用来实现本申请提供的自动化非结构化数据抽取方法以及问答方法。示例性的,第二存储器中存储有问答模型144,上述问答模型144被第二处理器调用以实现基于机器学习的问题答复方法中服务器侧执行的步骤。可选地,第二存储器可以包括但不限于以下几种:RAM、ROM、PROM、EPROM、EEPROM。

[0068] 第二存储器中还存储有知识图谱142和自动化的非结构化数据抽取模型146。当第二处理器执行基于机器学习的问题答复方法中服务器侧的步骤时,第二处理器调用问答模型144从知识图谱142中寻找得到问题的正确答案对应的向量序列。

[0069] 可选地,第二处理器通过调用第二存储器中存储非结构化数据抽取模型146,以实现上述非结构化数据抽取方法。可选地,第二处理器可以是通用处理器,比如,CPU或者NP。

[0070] 示意性的,本申请提供的基于机器学习的问题答复方法可以应用于车载语音系统、智能音箱、智能客服、儿童陪伴机器人、智能问答软件、百科问答软件等问答产品(终端)中。

[0071] 图2示出了本申请一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图。该方法可以由图1所示的服务器来执行。所述方法包括:

[0072] 步骤201,获取输入的知识文本;

[0073] 服务器中存储有自动化的非结构化数据抽取模型(下文简称抽取模型)。当存在待学习的知识文本时,将待学习的知识文本输入至该抽取模型中。

[0074] 知识文本是一个长文本,知识文本包括一个或多个段落。或者,知识文本包括一篇文章。或者,知识文本包括多个句子。

[0075] 步骤203,调用头实体识别模型从知识文本中识别出头实体,调用关系识别模型从知识文本中识别出关系实体;

[0076] 该抽取模型中包括:头实体(head)识别模型和关系(relation)识别模型。该头实体识别模型从知识文本中识别出一个或多个主语作为头实体,该关系识别模型从知识文本中识别出一个或多个谓词作为关系。

[0077] 步骤205,根据头实体和关系实体构建问题;

[0078] 该抽取模型根据头实体和关系实体来构建问题。当主语和谓语均为一个时,可以构建出一个问题;当主语或谓词为至少两个时,可以构建出多个问题。

[0079] 步骤207,调用阅读理解模型根据问题从知识文本中提取答案,将答案确定为尾实体;

[0080] 该阅读理解模型的输入为问题和知识文本,输出为答案在知识文本中的位置。该

阅读理解模型是基于深度学习的自然语言处理模型。

[0081] 示例性的，该阅读理解模型为BERT模型。

[0082] 步骤209，根据头实体、关系实体和尾实体所构成的三元组构建知识图谱。

[0083] 在阅读理解模型提取到答案后，将答案作为尾实体。该抽取模型将头实体、关系实体、尾实体所形成的三元组(head，relation，tail)添加至知识图谱中。

[0084] 综上所述，本实施例提供的方法，通过调用头实体模型从知识文本中识别出头实体，调用关系识别模型从知识文本中识别出关系实体，根据头实体和关系实体构建问题，利用构建的问题来调用阅读理解模型根据问题从知识文本中提取答案。解决了相关技术中的阅读理解模型无法直接应用于开放式的非结构化数据提取的问题，实现了自动构建问题以调用阅读理解模型进行自动化的非结构化数据提取，从而实现自动化的非结构化数据提取的效果。

[0085] 图3示出了本申请一个示例性实施例提供的基于深度学习的非结构化数据抽取方法的流程图。该方法可以由图1所示的服务器来执行。所述方法包括：

[0086] 步骤301，获取输入的知识文本；

[0087] 服务器中存储有自动化的非结构化数据抽取模型(下文简称抽取模型)。当存在待学习的知识文本时，将待学习的知识文本输入至该抽取模型中。

[0088] 该知识文本是非结构化数据。知识文本是一个长文本，知识文本包括一个或多个段落。或者，知识文本包括一篇文章。或者，知识文本包括多个句子。示例性的，该知识文本是百科知识、网页、电子书籍、景点介绍中的至少一种文本。

[0089] 示例性的，若输入的知识文本较长时，服务器按照段落为单位，对知识文本进行拆解，将每个段落作为一个知识文本进行处理。

[0090] 步骤302，调用头实体识别模型从知识文本中识别出头实体；

[0091] 该抽取模型中包括：头实体识别模型41，如图4所示。该头实体识别模型41从知识文本中识别出一个或多个主语作为头实体。

[0092] 头实体识别模型的输入是知识文本(比如段落)，输出为头实体在知识文本中的位置。可选地，该头实体的位置采用起始位置来表示，或者，头实体的位置采用起始位置和结束位置来表示。

[0093] 在一个示例中，头实体的个数为一个或多个。

[0094] 步骤303，调用关系识别模型从知识文本中识别出关系实体；

[0095] 该抽取模型中包括：关系识别模型42，如图4所示。该关系识别模型42从知识文本中识别出一个或多个谓词作为关系。

[0096] 关系识别模型的输入是知识文本(比如段落)，输出为关系在知识文本中的位置。可选地，该关系的位置采用起始位置来表示，或者，关系的位置采用起始位置和结束位置来表示。

[0097] 在一个示例中，关系的个数为一个或多个。

[0098] 上述两个步骤的执行顺序的先后关系不加以限定，步骤302可以在步骤303之前执行，步骤303可以在步骤302之前执行，或者两个步骤同时执行。

[0099] 步骤304，确定头实体的第一实体类型和关系实体的第二实体类型；

[0100] 第一实体类型采用词性类别或语义类别来表示，第二实体类型也采用词性类别或

语义类别来表示。

[0101] 比如,语义类别为人,也即第一实体类型为人(person);语义类别为属性,也即第二实体类型为属性(attribute)。

[0102] 步骤305,从多个候选问题模板中,确定与第一实体类型和第二实体类型对应的问题模板;

[0103] 该抽取模型中提供有多个候选问题模板。比如:[person]的[attribute]是什么。每个候选问题模板对应一组(第一实体类型,第二实体类型)的组合。根据抽取出第一实体类型和第二实体类型,可选择出相应的问题模板。

[0104] 步骤306,将头实体和关系实体按照问题模板进行组合,得到问题;

[0105] 在一个示例中,将头实体和关系实体按照问题模板“[person]的[attribute]是什么”进行组合,则得到问题。

[0106] 由于头实体可能为至少两个,关系实体可能为至少两个。当头实体和关系实体中的至少一个为至少两个时,该抽取模型根据至少两个头实体或关系实体的排列组合,拆解得到至少两组头实体和关系实体之间的一对一组合。

[0107] 比如,头实体为3个,关系实体为4个,则头实体和关系实体的组合为12种,能够拆解得到12个问题。针对每个问题可以执行如下步骤。

[0108] 步骤307,获取头实体的第一词向量和关系实体的第二词向量;

[0109] 对于每个问题,获取头实体的第一词向量和关系实体的第二词向量。

[0110] 该抽取模型中设置有已经训练好的TranSE模型。该抽取模型通过TranSE模型,将头实体转化为第一词向量,将关系实体转换为第二词向量。

[0111] 步骤308,根据第一词向量,在知识图谱的已有三元组中确定出候选三元组;

[0112] 根据头实体的第一词向量,在知识图谱的已有三元组中能够确定出至少两种候选三元组。

[0113] 在一个示例中,遍历知识图谱的已有三元组,确定出头实体等于第一词向量的三元组作为候选三元组;在另一个示例中,遍历知识图谱的已有三元组,确定出头实体包括第一词向量的三元组作为候选三元组。

[0114] 步骤309,根据第一词向量和第二词向量,在候选三元组中确定出目标实体;

[0115] 示例性的,该抽取模型根据第一词向量和第二词向量计算预测向量;计算预测向量和候选三元组对应的标签向量之间的距离,将距离最小的候选三元组确定为目标实体。

[0116] 该计算过程可以参考如下公式实现:

[0117] $$\underset{(h,l,t)\in C}{minimize} ||p_l-\hat{p}_l||_2+\beta_1||e_h-\hat{e}_h||_2+\beta_2||f(e_h,p_l)-\hat{e}_t||_2 -\beta_3 sim\left[n(h),HED_{entity}\right]-\beta_4 sim[n(l),HED_{non}]$$

[0118] h为头实体,l为关系,t为尾实体,β_1至β_4为参数。p为关系向量,e为实体向量,C为知识图谱,p_l为关系向量,$\hat{p}_l$为预测的关系向量,e_h为实体向量,$\hat{e}_h$为预测的实体向量,$\hat{e}_t$为预测的尾实体向量。$||p_l-\hat{p}_l||_2$为关系向量和预测的关系向量之间的距离,n(h)和n(p)表示标签向量对应的字符,HED_{entity}为被识别为实体的字符,HED_{non}为被识别为不是实体的字符。sim为计算两个字符相似度的函数,f()函数定义为两参数相加,minimize为选取最小值。

[0119] 将具有最小值的候选三元组确定为目标实体。

[0120] 步骤310,判断目标实体是否满足真实性条件;

[0121] 假定一个阈值来判定一个(头实体,关系,目标实体)之间的关系是否符合客观规律。因为每一种三元组的关系配对总会有一个最小损失值,如果这个最小损失值大于设定阈值(比如0.3),则认定这个关系不符合客观的三元组事实。反之则认定是一个合法的三元组事实。

[0122] 该真实性条件包括:是否小于设定阈值。当满足真实性条件时,进入步骤311;当不满足真实性条件时,进入步骤312。

[0123] 步骤311,当目标实体满足真实性条件时,根据目标实体在知识文本中提取答案;

[0124] 由于目标实体是已知的知识图谱中的实体,而不是从输入的知识文本中提取的实体。也即基于知识图谱得到的答案是知识图谱中的信息,而不是用户提供的原文中的答案,不够智能化。因此,该抽取模型还需要在输入的知识文本中进行答案回溯。

[0125] 本步骤存在两种实现方式:

[0126] 一,基于词向量的相似度计算方法;

[0127] 将知识文本进行向量化,得到每个句子的词向量序列;计算目标实体的第三词向量和每个句子的词向量序列之间的相似度;从相似度最高的句子中提取出答案。

[0128] 比如,知识文本是:“乌鲁木齐地处中国西北地区、新疆中部、亚欧大陆中心、天山山脉中段北麓、准噶尔盆地南缘,毗邻中亚各国,有“亚心之都”的称呼,是第二座亚欧大陆桥中国西部桥头堡和中国向西开放的重要门户[4],并被列入吉尼斯世界纪录大全,是世界上最内陆、距离海洋和海岸线最远的大型城市(2500公里)。”。该抽取模型经过知识图谱的问答方法已经得到了一个三元组关系符合条件:[乌鲁木齐-地理位置-新疆中部,天山北麓]。

[0129] 该抽取模型得到了知识图谱中的答案(目标实体):“新疆中部,天山北麓”,但是这还不够,该抽取模型需要得到知识文本中对应的答案,不然就显得该抽取模型生搬硬套,没有从用户给的数据中抽取相关的答案。该抽取模型对答案和知识文本进行符号分割,然后用词向量去计算与之最接近的话语,找到知识文本中的答案的起始位置,这样就实现了答案回溯。上面例子中,该抽取模型首先把答案拆成:“新疆中部”和“天山北麓”,把知识文本也按照标点符号进行拆分,然后分别映射成词向量的形式(分词,然后查字典,多个词向量相加,再做归一化,这里的字典可以使用开源900万中文词向量)。该抽取模型分别计算每个句子和目标实体之间的余弦相似度,该抽取模型发现知识文本中“新疆中部”和答案“新疆中部”的余弦相似度最接近,因此该句就是知识文本答案的开始位置,同理,该抽取模型计算得到“天山山脉中段北麓”与“天山北麓”的余弦相似度最接近,所以这一句就是答案的结束位置,因此,最接近的对应答案就是:“新疆中部、亚欧大陆中心、天山山脉中段北麓”,因此针对这个结果,该抽取模型最后抽取的结果就是:[乌鲁木齐—地理位置--新疆中部、亚欧大陆中心、天山山脉中段北麓]。

[0130] 二,基于最长公共子序列的相似度计算方法

[0131] 由于第一种方式中映射得到词向量的计算量较大,比如需要下载开源900万中文词向量以及计算,需要内存加载这么大的数据,约16G),则可以使用如下替代方式:该抽取模型计算目标实体和知识文本中的每个句子的最长公共子序列;从具有最长的最长公共子

序列的句子中提取出答案。

[0132] 示例性的，该抽取模型计算目标实体的起始位置和知识文本中的每个句子的最长公共子序列，得到第一位置；计算目标实体的结束位置和知识文本中的每个句子的最长公共子序列，得到第二位置；将第一位置和第二位置之间的文本序列提取为答案。

[0133] 比如，该抽取模型分别计算目标实体(答案)的开始位置和结束位置，在知识文本中与之最匹配的最长公共子序列用于确定答案的起始位置，同样也能得出一样的结论(“新疆中部”与“新疆中部”拥有最长公共子序列，“天山北麓”与“天山山脉中段北麓”拥有最长公共子序列)。

[0134] 步骤312，当目标实体不满足真实性条件时，调用阅读理解模型根据问题从知识文本中提取答案；

[0135] 该阅读理解模型的输入为问题和知识文本，输出为答案在知识文本中的位置。该阅读理解模型是基于深度学习的自然语言处理模型。

[0136] 示例性的，该阅读理解模型为BERT模型。

[0137] 步骤313，将答案确定为尾实体；

[0138] 步骤314，根据头实体、关系实体和尾实体所构成的三元组构建知识图谱。

[0139] 该知识图谱可以用于用户问询的时候方便调取知识来回答。该知识图谱是通用类型的知识图谱，或者，针对某个领域的专用知识图谱。

[0140] 综上所述，本实施例提供的方法，通过调用头实体模型从知识文本中识别出头实体，调用关系识别模型从知识文本中识别出关系实体，根据头实体和关系实体构建问题，利用构建的问题来调用阅读理解模型根据问题从知识文本中提取答案。解决了相关技术中的阅读理解模型无法直接应用于开放式的非结构化数据提取的问题，实现了自动构建问题以调用阅读理解模型进行自动化的非结构化数据提取，从而实现自动化的非结构化数据提取的效果。

[0141] 本实施例提供的方法，通过提取头实体和关系的方式，在存在至少两个头实体和/或至少两个关系时，采用按照排列组合的方式进行拆解，从而很好的解决了多对多的问题构建。

[0142] 本实施例提供的方法，还通过基于语义类型的相似度计算方法，能够通过目标实体从知识文本(原文)中提取出较为准确和原汁原味的答案，实现了较高的人工智能程度。

[0143] 本实施例提供的方法，还通过基于字符类型的相似度计算，能够以较少的计算量来通过目标实体从知识文本(原文)中提取出较为准确和原汁原味的答案，实现了更为简洁且高效的答案提取方式。

[0144] 本实施例提供的方法，还通过当目标实体不满足真实性条件时，调用阅读理解模型根据问题从知识文本中提取答案，当知识图谱无法提取答案时，利用阅读理解模型提取答案，实现了更加全面的答案提取方式。

[0145] 在基于图3所示的实施例中，头实体识别模型和关系识别模型是需要预先训练得到的模型。在训练过程中，首先导出知识图谱中的所有的三元组，然后利用TRANSE算法去训练词向量。将每个三元组实例(head，relation，tail)中的关系relation看做从实体head到实体tail的翻译，通过不断调整h、r和t(head、relation和tail的向量)，使(h+r)尽可能与t相等，即h+r＝t。

[0146] 在基于图3所示的实施例中，上述阅读理解模型是BERT模型。该BERT模型是需要预先得到的模型。图5示出了本申请一个示例性实施例提供的阅读理解模型训练方法的流程图，该方法应用于服务器中国，该方法包括：

[0147] 步骤401，服务器获取训练样本。

[0148] 每组训练样本包括问题样本、知识文本样本和标定位置。一个问题样本是采集得到的一个历史问题；一个历史问题对应一个知识文本中的正确答案、以及知识文本中的非答案内容，知识文本样本是由正确答案和非答案内容混合在一起形成的文档。知识文本样本中包括一个正确答案和至少一个非答案内容。

[0149] 标定位置是正确答案在知识文本样本中句子的位置；其中，标定位置可以包括起始标定位置和终止标定位置，起始标定位置是正确答案在知识文本样本中句子的起始位置，终止标定位置是正确答案在知识文本样本中句子的终止位置。比如，知识文本样本包括两个句子“今天天气晴朗。今天是本月最后一天。”，对上述两个句子进行分词得到顺序排列的分词结果：“今天”、“天气”、“晴朗”、“今天”、“是”、“本月”、“最后”、“一天”；每一个分词经过词嵌入、编码之后，得到对应的词向量，并按照上述分词结果的排列顺序形成知识文本样本的向量序列；若上述第一个句子是正确答案，那么正确答案的起始标定位置为1，即知识文本样本的向量序列中的第一个词向量，终止标定位置为3，即知识文本样本的向量序列中的第三个词向量；因此，上述第一个词向量至第三个词向量组成的子向量序列即为正确答案对应的向量序列。

[0150] 步骤402，服务器通过阅读理解模型对知识文本样本分别进行编码，得到知识文本样本的向量序列。

[0151] 可选地，阅读理解模型是BERT模型。服务器通过阅读理解模型的编码器对知识文本样本进行编码得到知识文本样本的向量序列。

[0152] 可选地，服务器通过阅读理解模型对知识文本样本中的各个句子进行词嵌入，得到知识文本样本的向量序列；其次，服务器通过阅读理解模型对知识文本样本的向量序列中的每一个词向量进行交叉编码，得到编码后的知识文本样本的向量序列。

[0153] 步骤403，服务器通过阅读理解模型预测正确答案在知识文本样本的向量序列中的位置，并确定上述正确答案的位置与标定位置之间的损失。

[0154] 可选地，上述正确答案的位置包括正确答案的起始位置和终止位置；其中，起始位置是阅读理解模型预测得到的正确答案在知识文本样本的向量序列中句子的起始位置，终止位置是阅读理解模型预测得到的正确答案在知识文本样本的向量序列中句子的终止位置。

[0155] 示意性的，阅读理解模型的输出层中包括归一化函数，归一化函数也就是softmax函数；服务器调用阅读理解模型中的softmax函数对知识文本样本的向量序列中的每一个词向量进行概率计算，根据得到的概率值预测出正确答案的句子起始位置和句子终止位置，即预测出正确答案的句子中第一个分词对应的词向量和最后一个分词对应的词向量的位置。

[0156] 服务器中的阅读理解模型中还包括损失函数(Loss Function)，通过损失函数确定出预测得到的正确答案的位置与标注位置之间的损失，即预测得到的正确答案的位置与标准位置之间的一致性。

[0157] 可选地，损失函数可以包括0-1损失(Zero-one Loss)函数、感知损失(Perceptron Loss)函数、铰链损失(Hinge Loss)函数、交叉熵损失函数、平方误差损失(Square Loss)函数、绝对值损失(Absolute Loss)函数、指数误差(Exponential Loss)函数和正则函数中的任意一种。

[0158] 步骤404，服务器通过上述损失对阅读理解模型中的模型参数进行调整，训练阅读理解模型对正确答案的位置预测能力。

[0159] 服务器通过上述损失对阅读理解模型中的模型参数进行调整，使模型参数调整后的阅读理解模型预测得到的正确答案在知识文本样本的向量序列中的位置与标定位置之间的损失更小。

[0160] 示意性的，服务器采用反向传播算法将上述损失反向传播，在反向传播的过程中，根据上述损失对阅读理解模型中的模型参数的值进行调整。

[0161] 示例性的，将上述实施例提供的非结构化数据抽取方法可以应用于实际，给出以下三个实施例。

[0162] 在一个如图6所示的示例性例子中，上述实施例提供的非结构化数据抽取方法可以应用于景点讲解系统中，系统中，该方法包括如下步骤：

[0163] 步骤601，景区相关文章(海量)。

[0164] 采集或收集大量的景区相关文章、资料、书记等文字类信息。信息数量越多越好。

[0165] 步骤602，本申请提供的自动化的非结构化数据抽取模型。

[0166] 利用本申请提供的自动化的非结构化数据抽取模型将步骤601中的景区相关文章转化为步骤603中的景区专有的知识图谱。

[0167] 步骤603，景区专有的知识图谱。

[0168] 利用本申请提供的自动化的非结构化数据抽取模型，获得景区专有的知识图谱。

[0169] 步骤604，游客关于景区提问。

[0170] 游客提出一个关于景区的问题。

[0171] 步骤605，自动搜寻相关知识回复。

[0172] 根据游客提出的问题，利用步骤603中的景区专有的知识图谱自动搜索相关知识回复游客。

[0173] 综上所述，利用上述实施例提供的非结构化数据抽取方法为景区生成专有的知识图谱，当游客提出景区相关的问题时，可以快速检索到相关知识来恢复游客，并且具有很高的准确度。

[0174] 在一个如图7所示的示例性例子中，上述实施例提供的非结构化数据抽取方法可以应用于自动化的知识图谱的构建过程，在构建过程中，该方法包括如下步骤：

[0175] 步骤701，所有领域的百科类知识(海量)。

[0176] 采集或收集所有领域内的百科类知识，数量越多越好。

[0177] 步骤702，本申请提供的自动化的非结构化数据抽取模型。

[0178] 利用本申请提供的自动化的非结构化数据抽取模型将步骤701中的所有领域的百科类知识转化为步骤703中的三元组类型的知识图谱。

[0179] 步骤703，三元组类型的知识图谱。

[0180] 利用本申请提供的自动化的非结构化数据抽取模型，获得三元组类型的知识图

谱。

[0181] 步骤704，用户query。

[0182] 用户提出一个问题。

[0183] 步骤705，服务器解析意图。

[0184] 服务器解析用户的意图。

[0185] 步骤706，查询结果返回用户。

[0186] 根据服务器解析处的用户意图，利用步骤703中的三元组类型的知识图谱查询结果并反馈给用户。

[0187] 综上所述，将上述实施例提供的非结构化数据抽取方法应用于自动化的知识图谱的构建过程，可以从无到有的构建出知识图谱或自动完善知识图谱，利用构建出的知识图谱可以快速准确地解答用户问题。

[0188] 在一个如图8所示的示例性例子中，上述实施例提供的非结构化数据抽取方法可以应用于全自动化的人工智能，该方法包括如下步骤：

[0189] 步骤801，人类所有的知识文章。

[0190] 采集或收集人类所有的知识文章。

[0191] 步骤802，本申请提供的自动化的非结构化数据抽取模型。

[0192] 利用本申请提供的自动化的非结构化数据抽取模型将步骤801中的人类所有的知识文章转化为步骤803中的人类知识图谱。

[0193] 步骤803，人类知识图谱。

[0194] 利用本申请提供的自动化的非结构化数据抽取模型，获得人类知识图谱。

[0195] 步骤804，机器自动学习。

[0196] 机器在人类知识图谱的基础上可以继续自动学习其他人类知识。

[0197] 步骤805，知晓人类知识的人工智能产品。

[0198] 最终获得一个知晓人类知识的人工智能产品。

[0199] 综上所述，将上述实施例提供的非结构化数据抽取方法应用于全自动化的人工智能，将人类所有的只是文章利用本申请提供的自动化的非结构化数据抽取模型生成人类知识图谱，使机器自动学习人类知识，最终获得一个知晓人类知识的人工智能产品。

[0200] 示例性的，将上述实施例提供的非结构化数据抽取方法应用于产品，给出以下实施例。

[0201] 如图9所示，当用户提问“一是什么?”时，根据上述实施例提供的非结构化数据抽取方法，获取问题答案并显示出来。

[0202] 示例性的，如图10所示，用户通过触发批量导入控件进入如图11所示的批量导入界面，在批量导入界面可以上传段落或文章。示例性的，用户上传的文字是“乌鲁木齐，通称乌市，旧称迪化，是新疆维吾尔自治区首府、新疆的政治、经济、文化、科教和交通中心，中国西北地区重要的中心城市和面向中亚西亚的国际商贸中心[1]。截至2018年，全市下辖7个区、1个县，总面积14216.3平方千米，建成区面积436平方千米，常住人口355万人，城镇人口261.57万人，城镇化率74.61，平均海拔是800米。”当用户提问“乌鲁木齐的平均海拔是多少”时，显示如图12所示的界面，显示“乌鲁木齐的高度是800米”。当用户提问“乌鲁木齐的人口总数是多少”时，显示如图13所示的界面，显示“乌鲁木齐的人口是355万人(2015年常

住人口)”。

[0203] 以下为本申请的装置实施例,对于装置实施例中未详细描述的细节,可以结合参考上述方法实施例中相应的记载,本文不再赘述。

[0204] 图14示出了本申请的一个示例性实施例提供的基于深度学习的非结构化数据抽取装置的结构示意图。该装置可以通过软件、硬件或者两者的结合实现成为终端的全部或一部分,该装置包括:获取模块1404、调用模块1407、识别模块1408、构建模块1410、提取模块1405、确定模块1403。

[0205] 获取模块1404,用于获取输入的知识文本;

[0206] 调用模块1407,用于调用头实体识别模型、关系识别模型和阅读理解模型;

[0207] 识别模块1408,用于在调用头实体识别模型后从所述知识文本中识别出头实体,在调用关系识别模型后从所述知识文本中识别出关系实体;

[0208] 构建模块1410,用于根据所述头实体和所述关系实体构建问题;根据所述头实体、所述关系实体和所述尾实体所构成的三元组构建知识图谱;

[0209] 提取模块1405,用于在调用阅读理解模型后根据所述问题从所述知识文本中提取答案;

[0210] 确定模块1403,用于将所述答案确定为尾实体。

[0211] 在一个可选的实施例中,所述构建模块还包括:确定子模块1412和组合子模块1411;

[0212] 所述确定子模块1412,用于确定所述头实体的第一实体类型和所述关系实体的第二实体类型;从多个候选问题模板中,确定与所述第一实体类型和所述第二实体类型对应的问题模板;

[0213] 所述组合子模块1411,用于将所述头实体和所述关系实体按照所述问题模板进行组合,得到所述问题。

[0214] 在一个可选的实施例中,所述头实体或所述关系实体为至少两个;

[0215] 所述装置还包括拆解模块1409;

[0216] 所述拆解模块1409,用于根据至少两个所述头实体或所述关系实体的排列组合,拆解得到至少两组所述头实体和所述关系实体之间的一对一组合。

[0217] 在一个可选的实施例中,所述装置还包括判断模块1406;

[0218] 所述获取模块1404,还用于获取所述头实体的第一词向量和所述关系实体的第二词向量;

[0219] 所述确定模块1403,还用于根据所述第一词向量,在所述知识图谱的已有三元组中确定出候选实体;根据所述第一词向量和所述第二词向量,在所述候选实体中确定出目标实体;

[0220] 所述判断模块1406,用于判断目标实体是否满足阈值条件;

[0221] 所述提取模块1405,还用于当所述目标实体不满足阈值条件时,执行所述调用阅读理解模型根据所述问题从所述知识文本中提取答案的步骤。

[0222] 在一个可选的实施例中,所述确定模块1403,还用于遍历所述知识图谱的已有三元组,确定出头实体等于所述第一词向量的三元组中的尾实体作为所述候选实体;

[0223] 或,

[0224] 遍历所述知识图谱的已有三元组，确定出头实体包括所述第一词向量的三元组中的尾实体作为所述候选实体。

[0225] 在一个可选的实施例中，所述装置还包括计算模块1402；

[0226] 所述计算模块1402，用于根据所述第一词向量和所述第二词向量计算预测向量；计算所述预测向量和所述候选实体对应的标签向量之间的距离；

[0227] 所述确定模块1403，还用于将所述距离最小的候选实体确定为所述目标实体。

[0228] 在一个可选的实施例中，所述提取模块1405，还用于当所述目标实体满足所述阈值条件时，根据所述目标实体在所述知识文本中提取所述答案。

[0229] 在一个可选的实施例中，所述装置还包括向量化模块1401和计算模块1402；

[0230] 所述向量化模块1401，用于将所述知识文本进行向量化，得到每个句子的词向量序列；

[0231] 所述计算模块1402，用于计算所述目标实体的第三词向量和所述每个句子的词向量序列之间的相似度；

[0232] 所述提取模块1405，还用于从所述相似度最高的句子中提取出所述答案。

[0233] 在一个可选的实施例中，所述装置还包括计算模块1402；

[0234] 所述计算模块1402，用于计算所述目标实体和所述知识文本中的每个句子的最长公共子序列；

[0235] 所述提取模块1405，还用于从具有最长的所述最长公共子序列的句子中提取出所述答案。

[0236] 图15是本申请一个实施例提供的服务器的结构示意图。具体来讲：服务器700包括中央处理单元(英文：Central Processing Unit，简称：CPU)701、包括随机存取存储器(英文：random access memory，简称：RAM)702和只读存储器(英文：read-only memory，简称：ROM)703的系统存储器704，以及连接系统存储器704和中央处理单元701的系统总线705。服务器700还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统(I/O系统)706，和用于存储操作系统713、应用程序714和其他程序模块715的大容量存储设备707。

[0237] 基本输入/输出系统706包括有用于显示信息的显示器708和用于用户输入信息的诸如鼠标、键盘之类的输入设备709。其中显示器708和输入设备709都通过连接到系统总线705的输入/输出控制器710连接到中央处理单元701。基本输入/输出系统706还可以包括输入/输出控制器710以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入/输出控制器710还提供输出到显示屏、打印机或其他类型的输出设备。

[0238] 大容量存储设备707通过连接到系统总线705的大容量存储控制器(未示出)连接到中央处理单元701。大容量存储设备707及其相关联的计算机可读介质为服务器700提供非易失性存储。也就是说，大容量存储设备707可以包括诸如硬盘或者只读光盘(英文：Compact Disc Read-Only Memory，简称：CD-ROM)驱动器之类的计算机可读介质(未示出)。

[0239] 不失一般性，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括RAM、ROM、可擦除可编程只读存储器(英文：Erasable Programmable Read-Only Memory，简称：EPROM)、电可擦除可编程只读存储器(英文：Electrically Erasable Programmable Read-

Only Memory,简称:EEPROM)、闪存或其他固态存储其技术,CD-ROM、数字通用光盘(英文:Digital Versatile Disc,简称:DVD)或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然,本领域技术人员可知计算机存储介质不局限于上述几种。上述的系统存储器704和大容量存储设备707可以统称为存储器。

[0240] 根据本申请的各种实施例,服务器700还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即服务器700可以通过连接在系统总线705上的网络接口单元711连接到网络712,或者说,也可以使用网络接口单元711来连接到其他类型的网络或远程计算机系统(未示出)。

[0241] 本申请还提供一种计算机设备,该计算机设备包括:处理器和存储器,该存储介质中存储有至少一条指令、至少一段程序、代码集或指令集,该至少一条指令、至少一段程序、代码集或指令集由处理器加载并执行以实现上述各方法实施例提供的基于深度学习的非结构化数据抽取方法。

[0242] 本申请还提供一种计算机可读存储介质,该存储介质中存储有至少一条指令、至少一段程序、代码集或指令集,该至少一条指令、至少一段程序、代码集或指令集由处理器加载并执行以实现上述各方法实施例提供的基于深度学习的非结构化数据抽取方法。

[0243] 应当理解的是,在本文中提及的“多个”是指两个或两个以上。“和/或”,描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

[0244] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0245] 以上仅为本申请的可选实施例,并不用以限制本申请,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。

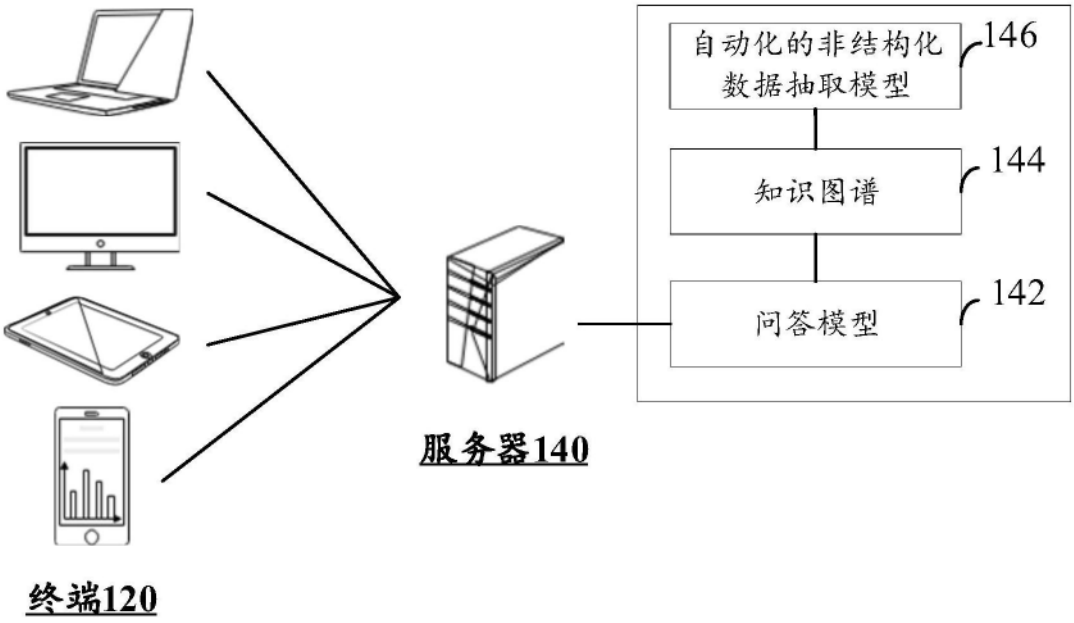

自动化的非结构化数据抽取模型
146
知识图谱
144
问答模型
142
服务器140
终端120

图1

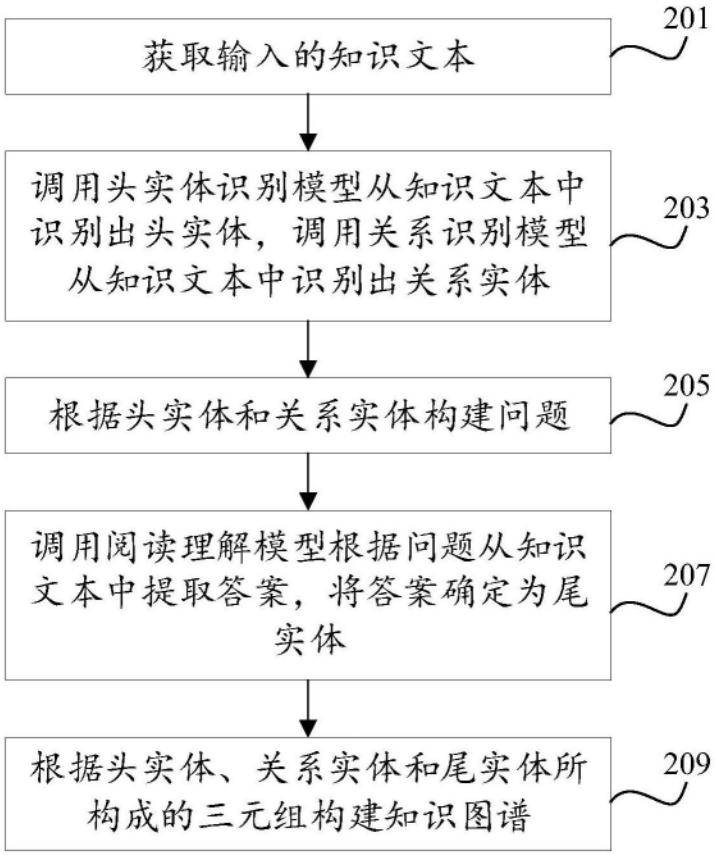

获取输入的知识文本
201
调用头实体识别模型从知识文本中识别出头实体，调用关系识别模型从知识文本中识别出关系实体
203
根据头实体和关系实体构建问题
205
调用阅读理解模型根据问题从知识文本中提取答案，将答案确定为尾实体
207
根据头实体、关系实体和尾实体所构成的三元组构建知识图谱
209

图2

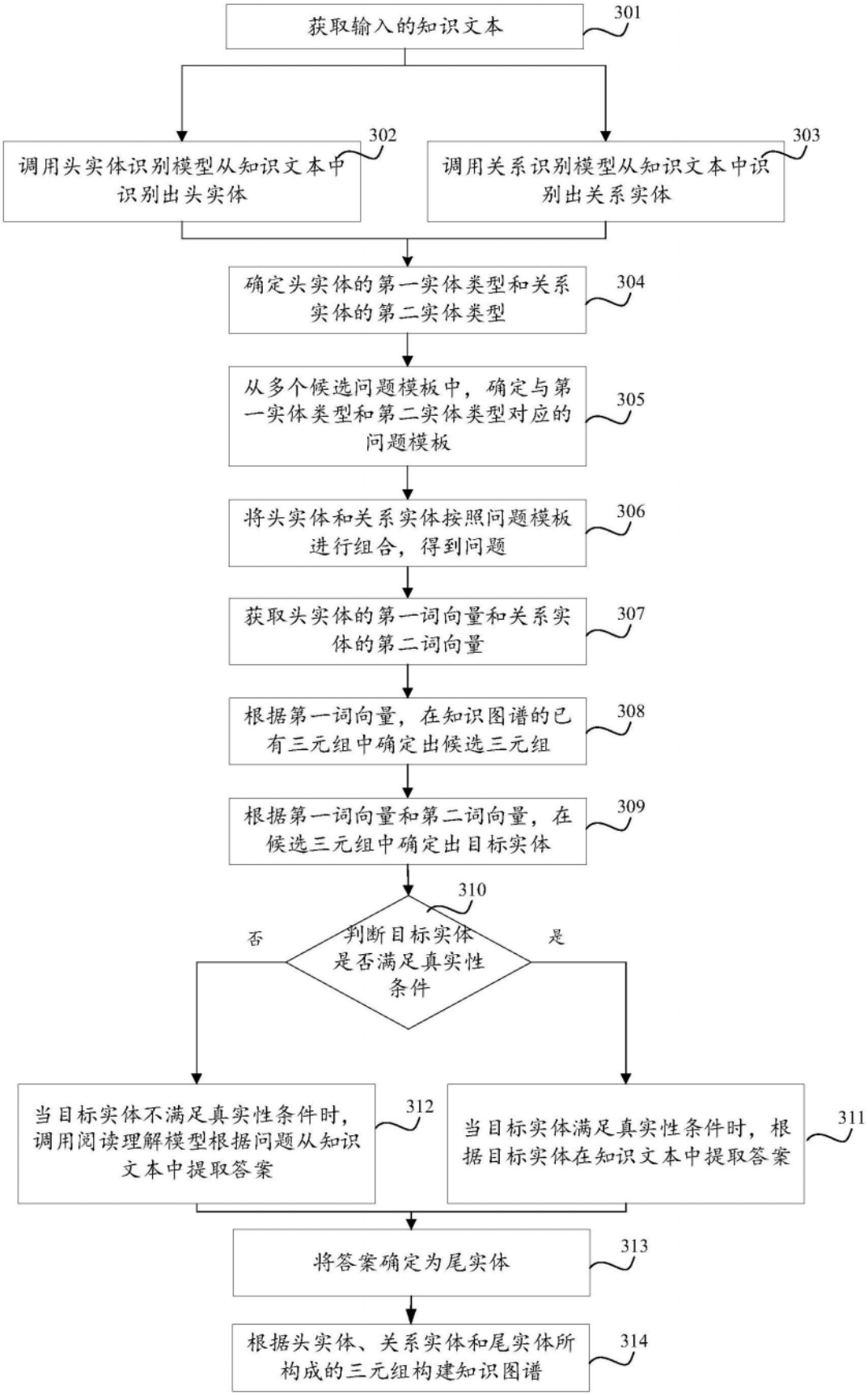

获取输入的知识文本
301
调用头实体识别模型从知识文本中识别出头实体
302
调用关系识别模型从知识文本中识别出关系实体
303
确定头实体的第一实体类型和关系实体的第二实体类型
304
从多个候选问题模板中，确定与第一实体类型和第二实体类型对应的问题模板
305
将头实体和关系实体按照问题模板进行组合，得到问题
306
获取头实体的第一词向量和关系实体的第二词向量
307
根据第一词向量，在知识图谱的已有三元组中确定出候选三元组
308
根据第一词向量和第二词向量，在候选三元组中确定出目标实体
309
判断目标实体是否满足真实性条件
310
否
是
当目标实体不满足真实性条件时，调用阅读理解模型根据问题从知识文本中提取答案
312
当目标实体满足真实性条件时，根据目标实体在知识文本中提取答案
311
将答案确定为尾实体
313
根据头实体、关系实体和尾实体所构成的三元组构建知识图谱
314

图3

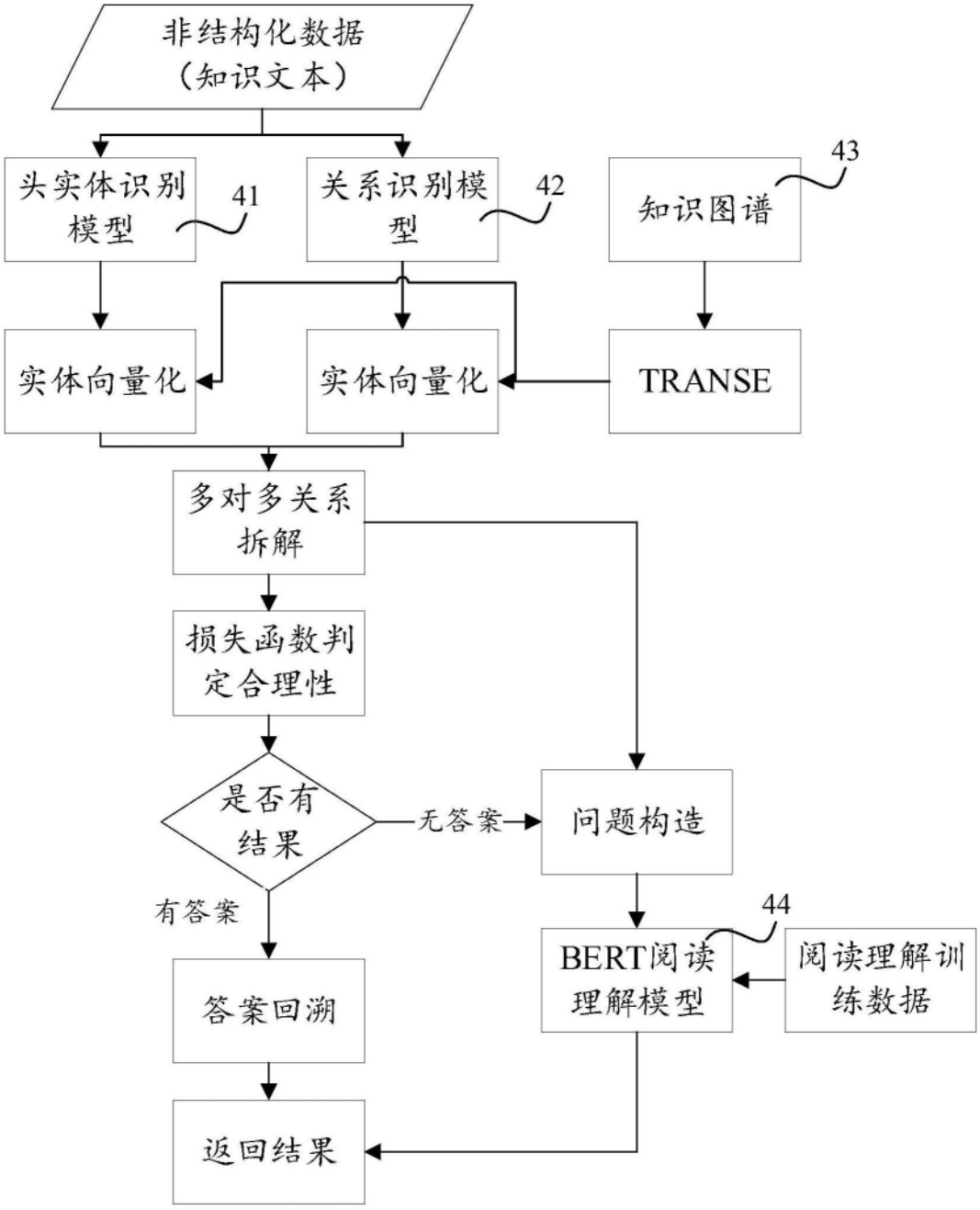

非结构化数据
（知识文本）
头实体识别模型
41
关系识别模型
42
知识图谱
43
实体向量化
实体向量化
TRANSE
多对多关系拆解
损失函数判定合理性
是否有结果
无答案
问题构造
有答案
44
BERT阅读理解模型
阅读理解训练数据
答案回溯
返回结果

图4

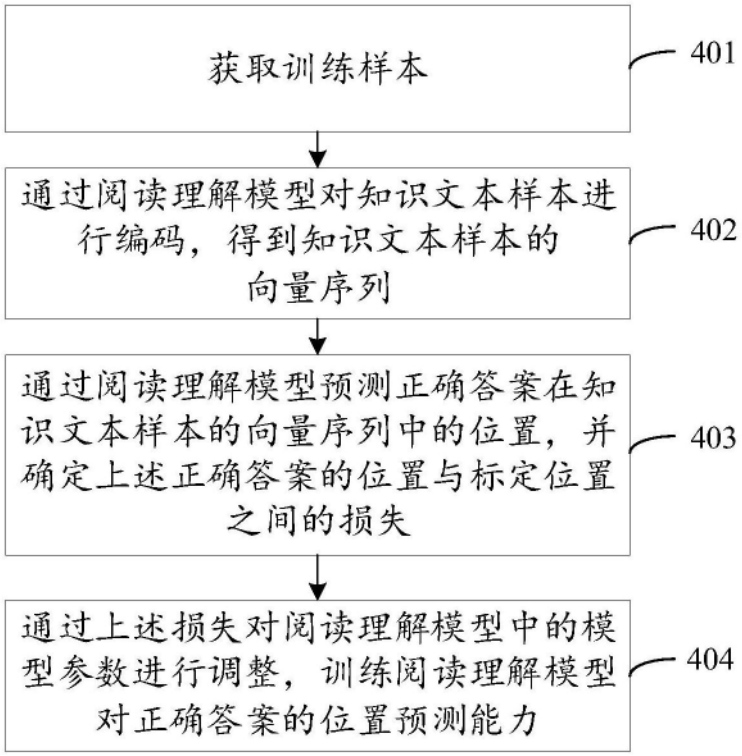

获取训练样本
401
通过阅读理解模型对知识文本样本进行编码，得到知识文本样本的向量序列
402
通过阅读理解模型预测正确答案在知识文本样本的向量序列中的位置，并确定上述正确答案的位置与标定位置之间的损失
403
通过上述损失对阅读理解模型中的模型参数进行调整，训练阅读理解模型对正确答案的位置预测能力
404

图5

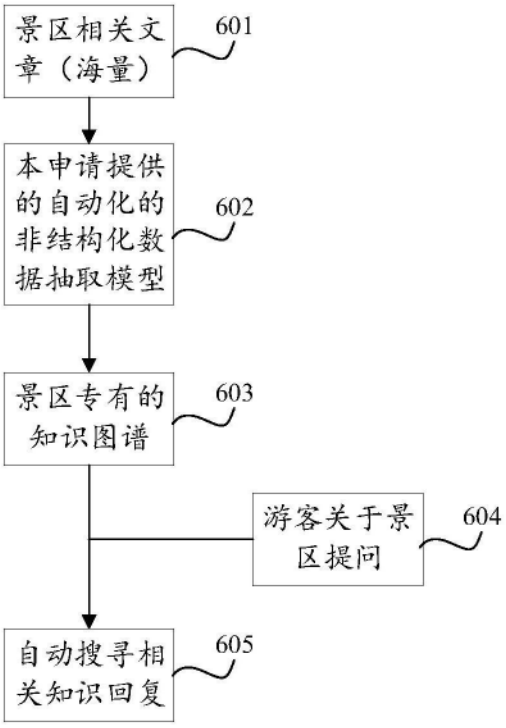

景区相关文章（海量）
601
本申请提供的自动化的非结构化数据抽取模型
602
景区专有的知识图谱
603
游客关于景区提问
604
自动搜寻相关知识回复
605

图6

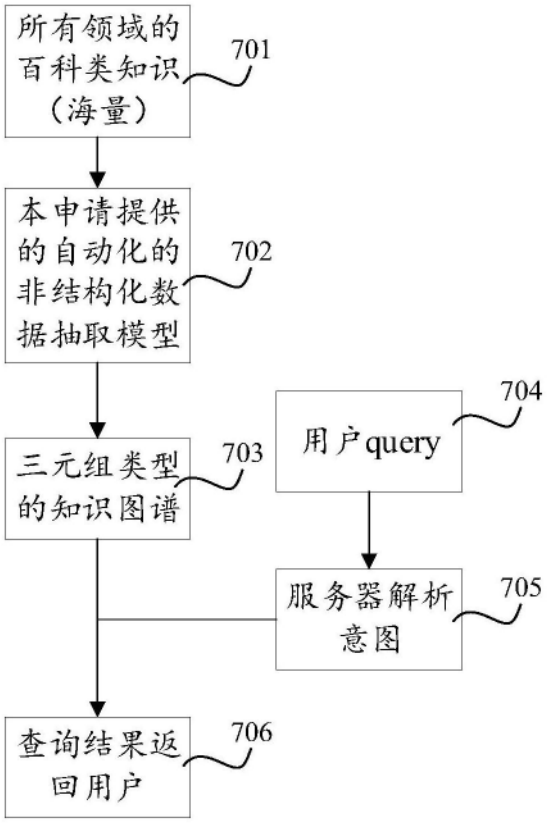

所有领域的百科类知识（海量）
701
本申请提供的自动化的非结构化数据抽取模型
702
三元组类型的知识图谱
703
用户query
704
服务器解析意图
705
查询结果返回用户
706

图7

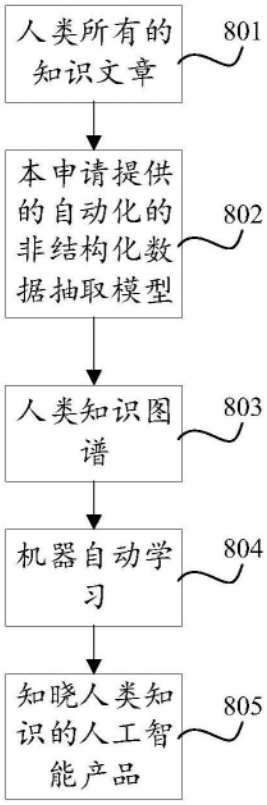

人类所有的知识文章
801
本申请提供的自动化的非结构化数据抽取模型
802
人类知识图谱
803
机器自动学习
804
知晓人类知识的人工智能产品
805

图8

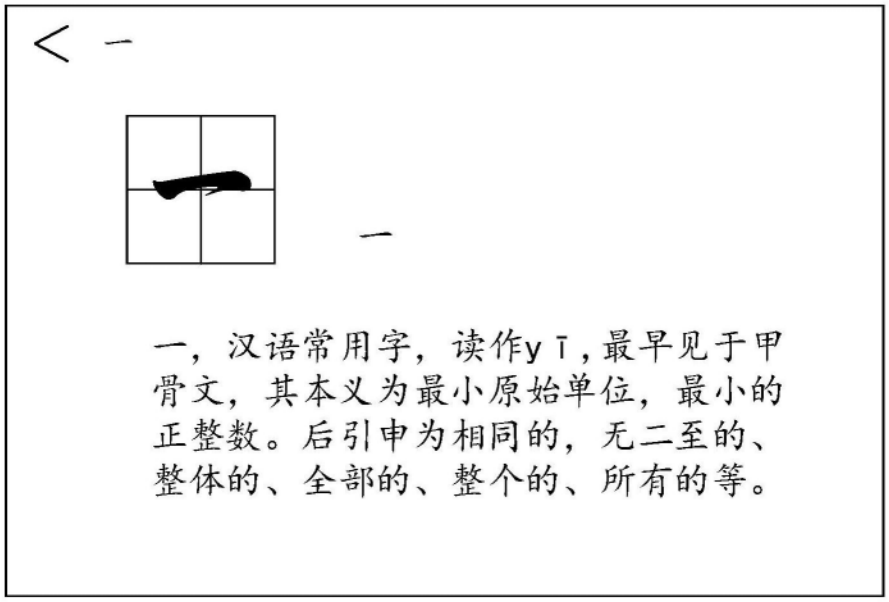

一
一，汉语常用字，读作yī，最早见于甲骨文，其本义为最小原始单位，最小的正整数。后引申为相同的，无二至的、整体的、全部的、整个的、所有的等。

图9


www.dingdang.xxx.45698721/cn
万物起源问答对
① 问答配置
问答管理
质量测试
② 发布
③ 测试验证
④ 线上学习
⑤ 统计
问答管理
添加问答
批量导入
批量导出
No.
问题
答案
1
飞机、电话、电灯的发明者是谁?
飞机的发明者是美国莱特
2
飞机的发明人是谁
飞机的发明时间是1903年
3
飞机的发明时间
飞机的发明时间是1903年
4
飞机电灯电话的发明者各是谁
飞机的发明者是美国莱特

图10

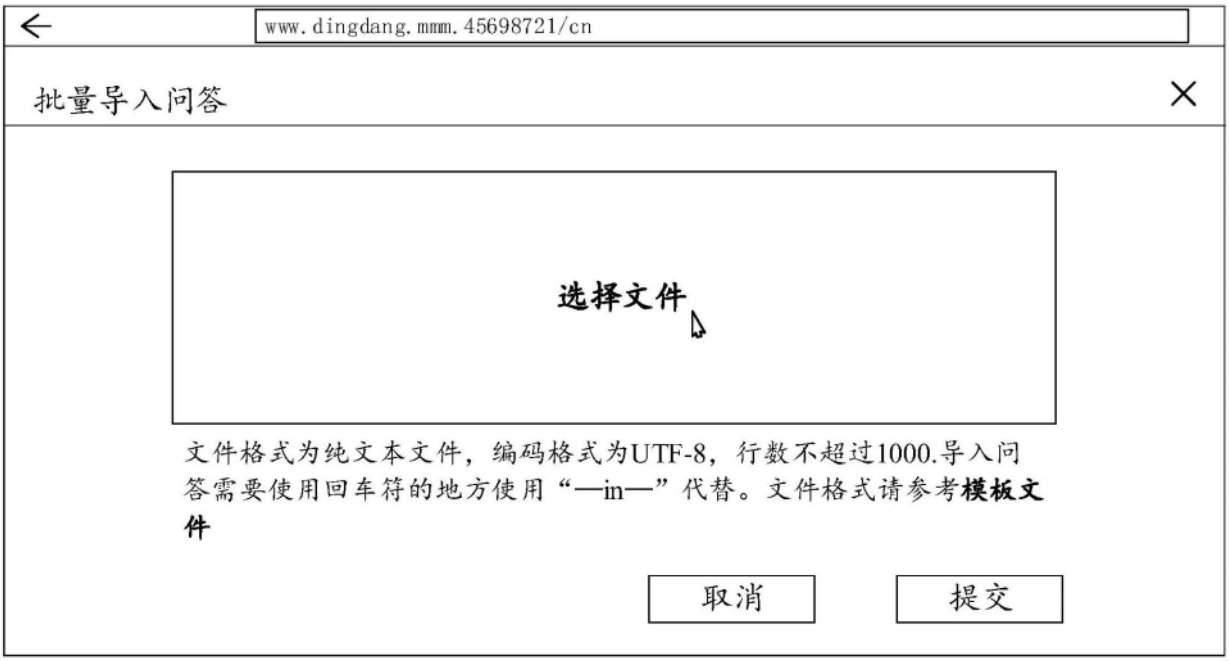

www.dingdang.mmm.45698721/cn
批量导入问答
选择文件
文件格式为纯文本文件，编码格式为UTF-8，行数不超过1000.导入问答需要使用回车符的地方使用“—in—”代替。文件格式请参考模板文件
取消
提交

图11

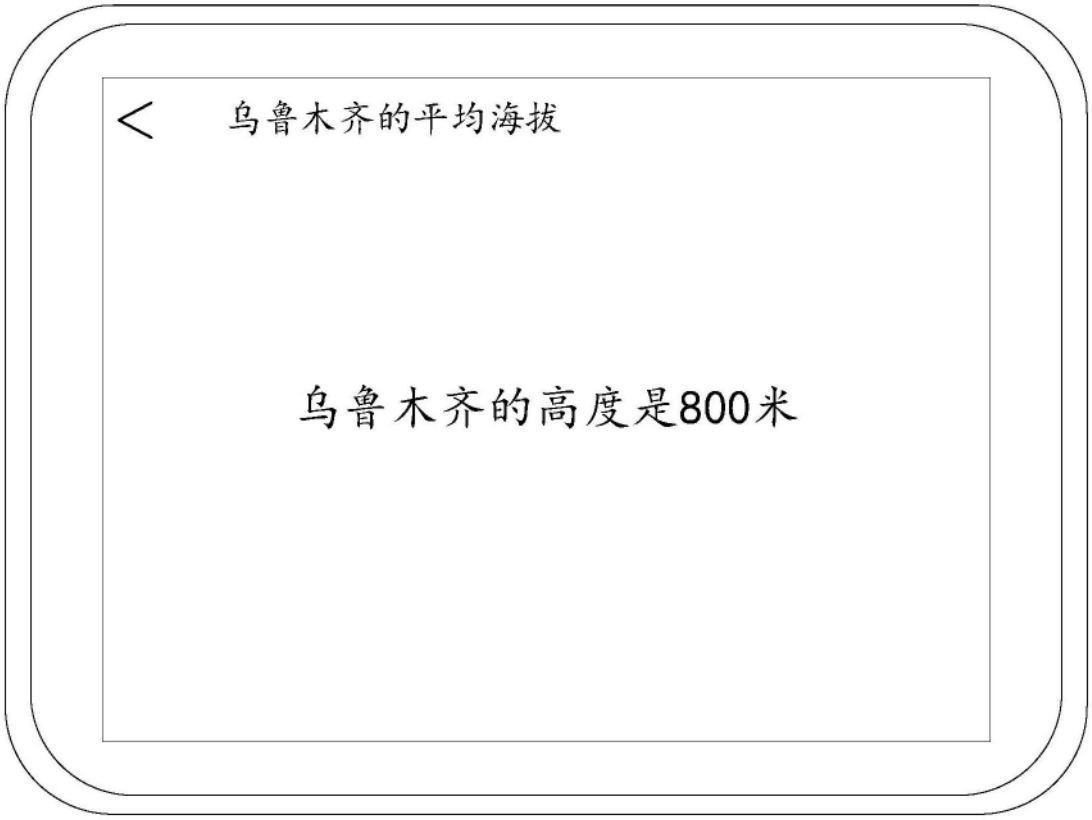

乌鲁木齐的平均海拔
乌鲁木齐的高度是800米

图12

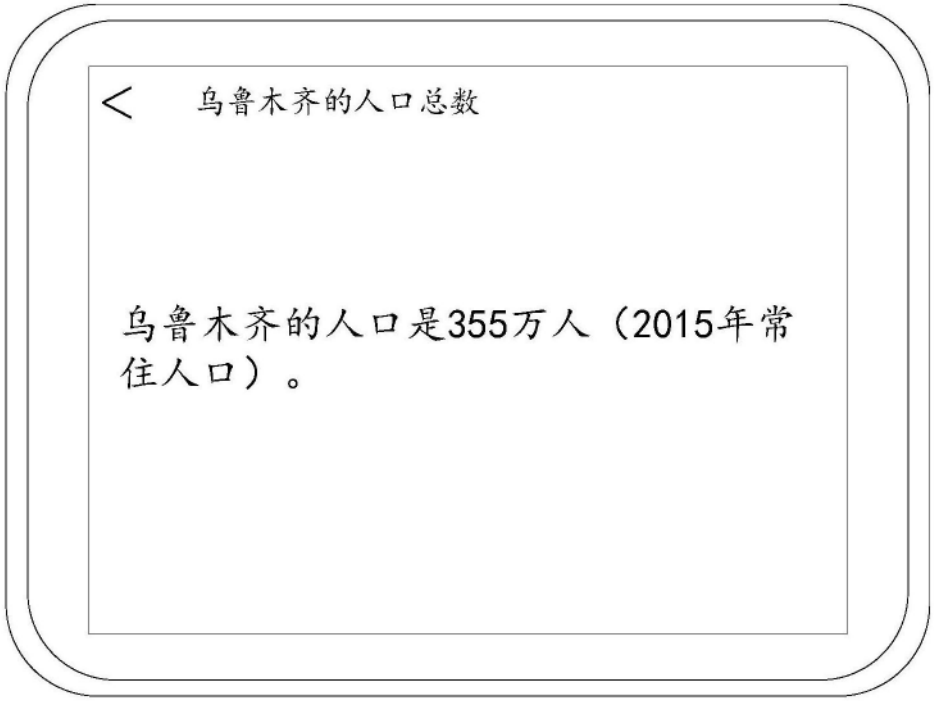

乌鲁木齐的人口总数
乌鲁木齐的人口是355万人（2015年常住人口）。

图13

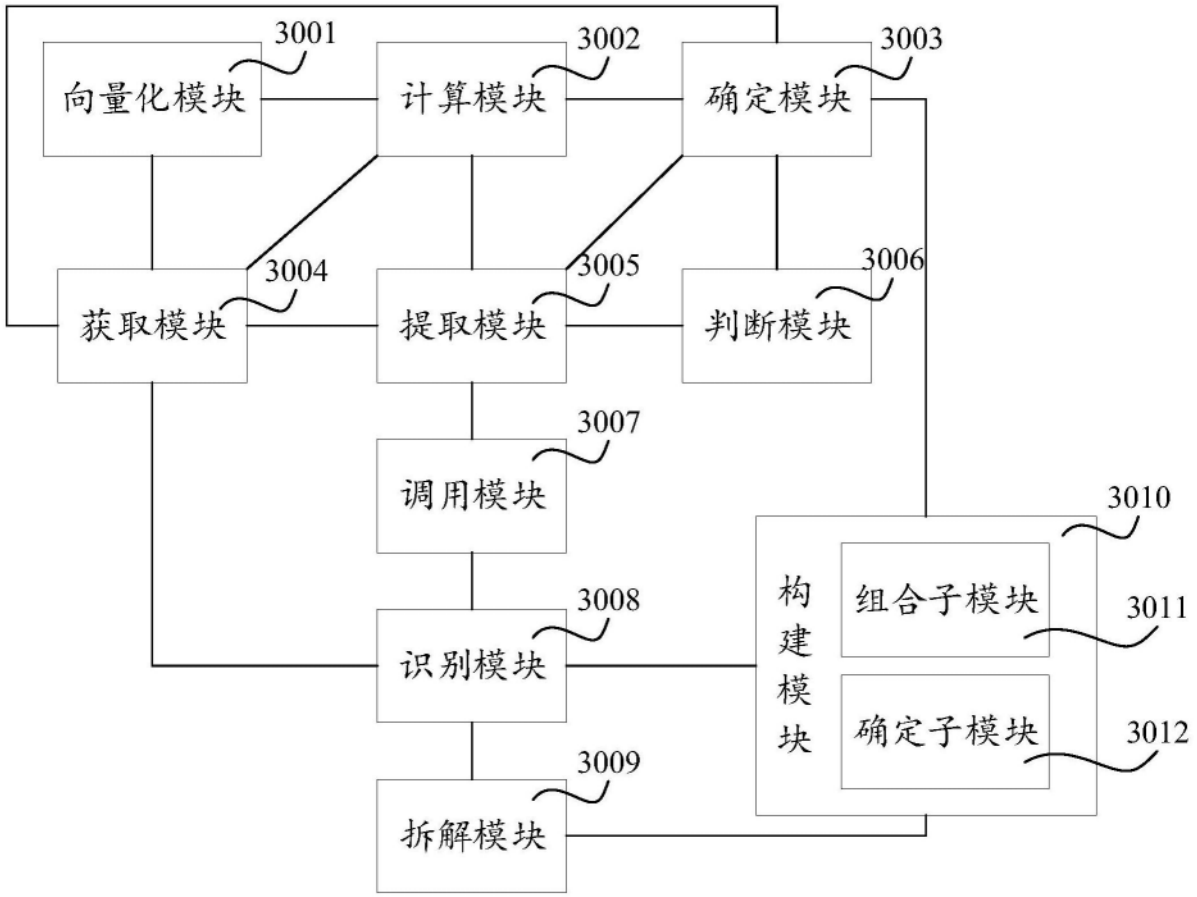

3001
向量化模块
3002
计算模块
3003
确定模块
3004
获取模块
3005
提取模块
3006
判断模块
3007
调用模块
3008
识别模块
3009
拆解模块
3010
构建模块
组合子模块
3011
确定子模块
3012

图14

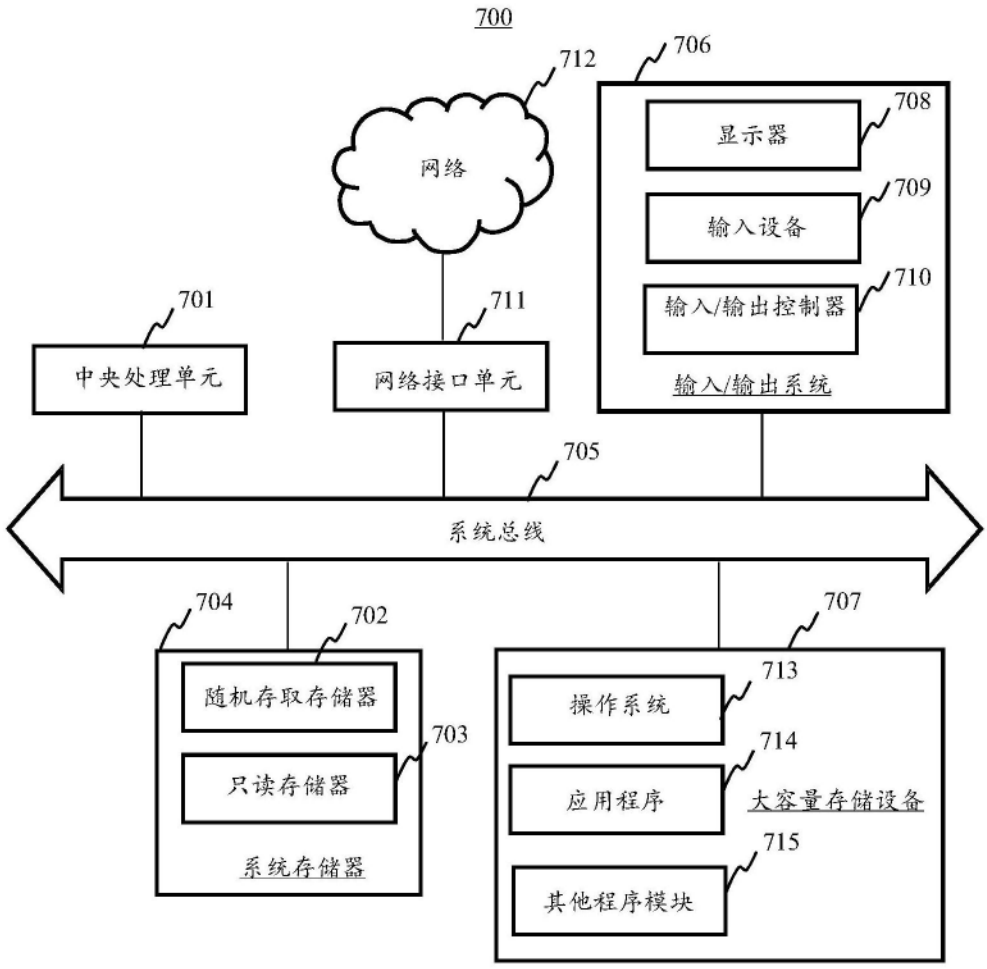

700
712
706
网络
显示器
708
709
输入设备
710
输入/输出控制器
701
711
中央处理单元
网络接口单元
输入/输出系统
705
系统总线
704
702
707
713
随机存取存储器
操作系统
703
714
只读存储器
应用程序
大容量存储设备
系统存储器
715
其他程序模块

图15