

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 7 月 2 日 (02.07.2020)



(10) 国际公布号
WO 2020/133291 A1

- (51) 国际专利分类号:
G06F 17/27 (2006.01)
- (21) 国际申请号: PCT/CN2018/125107
- (22) 国际申请日: 2018 年 12 月 28 日 (28.12.2018)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 深圳市优必选科技有限公司 (UBTECH ROBOTICS CORP) [CN/CN]; 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。
- (72) 发明人: 熊友军 (XIONG, Youjun); 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。 罗沛鹏

(LUO, Peipeng); 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。 廖洪涛 (LIAO, Hongtao); 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳中细软知识产权代理有限公司 (SHENZHEN CIPRUN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深圳市南山区粤海街道粤兴四道 1 号中山大学深圳产学研大楼 11 楼, Guangdong 518000 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

(54) Title: TEXT ENTITY RECOGNITION METHOD AND APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 文本实体识别方法、装置、计算机设备及存储介质

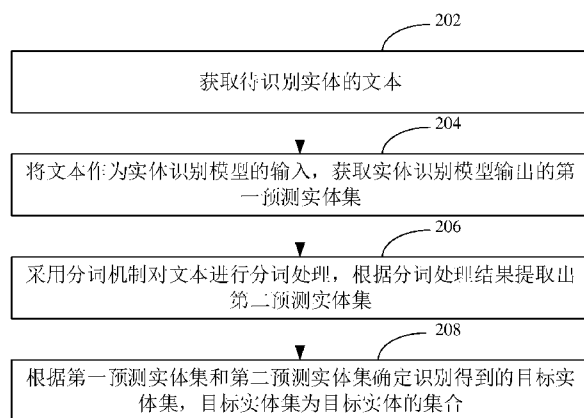


图 2

- 202 Obtain a text of an entity to be recognized
- 204 Use the text as an input of an entity recognition model to obtain a first predicted entity set output by the entity recognition model
- 206 Use a word segmentation mechanism to perform word segmentation on the text, and extract, according to the word segmentation processing result, a second predicted entity set
- 208 Determine, according to the first predicted entity set and the second predicted entity set, a target entity set obtained through recognition, the target entity set being a set of target entities

(57) Abstract: A text entity recognition method, comprising: obtaining a text of an entity to be recognized (S202); using the text as an input of an entity recognition model to obtain a first predicted entity set output by the entity recognition model (S204); using a word segmentation mechanism to perform word segmentation on the text, and extracting, according to the word segmentation processing result, a second predicted entity set (S206); and determining, according to the first predicted entity set and the second predicted entity set, a target entity set obtained through recognition, the target entity set being a set of target entities (S208). By using the word segmentation mechanism as a supplementary recognition, the accuracy of entity recognition is improved. In addition, also provided are a text entity recognition apparatus, a computer device, and a storage medium.

(57) 摘要: 一种文本实体识别方法, 该方法包括: 获取待识别实体的文本 (S202); 将所述文本作为实体识别模型的输入, 获取所述实体识别模型输出的第一预测实体集 (S204); 采用分词机制对所述文本进行分词处理, 根据分词处理结果提取出第二预测实体集 (S206); 根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集, 所述目标实体集为目标实体的集合 (S208)。通过采用分词机制作为补充识别, 提高了实体识别的准确率。此外, 还提出了一种文本实体识别装置、计算机设备及存储介质。

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

说明书

发明名称：文本实体识别方法、装置、计算机设备及存储介质

技术领域

[0001] 本发明涉及计算机处理领域，尤其是涉及一种文本实体识别方法、装置、计算机设备及存储介质。

背景技术

[0002] 随着深度学习技术的成熟以及计算机性能的提升，深度学习技术广泛运用于文本实体识别任务中，尤其是机器人对话语料，需要获取实体才能进行相关的会话编排。

发明概述

技术问题

[0003] 然而，深度学习实体识别模型预测的准确率具有一定的瓶颈（只能达到80%到90%之间），所以传统的语料中的实体识别的准确率偏低。

问题的解决方案

技术解决方案

[0004] 基于此，有必要针对上述问题，提供了一种实体识别准确率高的文本实体识别方法、装置、计算机设备及存储介质。

[0005] 第一方面，本发明实施例提供一种文本实体识别方法，所述方法包括：

[0006] 获取待识别实体的文本；

[0007] 将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；

[0008] 采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；

[0009] 根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。

[0010] 第二方面，本发明实施例提供一种文本实体识别装置，所述装置包括：

[0011] 文本获取模块，用于获取待识别实体的文本；

[0012] 输入输出模块，用于将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；

[0013] 分词提取模块，用于采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；

[0014] 确定模块，用于根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。

[0015] 第三方面，本发明实施例提供一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述计算机程序被所述处理器执行时，使得所述处理器执行如下步骤：

[0016] 获取待识别实体的文本；

[0017] 将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；

[0018] 采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；

[0019] 根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。

[0020] 第四方面，本发明实施例提供一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行如下步骤：

[0021] 获取待识别实体的文本；

[0022] 将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；

[0023] 采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；

[0024] 根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。

发明的有益效果

有益效果

[0025] 上述文本实体识别方法，通过获取待识别实体的文本，将文本作为实体识别模

型的输入，获取实体识别模型输出的第一预测实体集，同时采用分词机制对文本进行分词处理，根据分词处理结果提取出第二预测实体集，然后根据第一预测实体集和第二预测实体集确定识别得到的目标实体集。上述文本实体识别方法在得到第一预测实体集后，通过分词机制提取出第二预测实体集，然后采用第二预测实体集对第一预测实体集进行补充识别，即采用实体识别模型和分词机制共同来确定目标实体集，能够得到更多、更可信、更准确的实体，从而提高了实体识别的准确率。

对附图的简要说明

附图说明

- [0026] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术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示出的结构获得其他的附图。
- [0027] 图1为一个实施例中文本实体识别方法的应用环境图；
- [0028] 图2为一个实施例中文本实体识别方法的流程图；
- [0029] 图3为一个实施例中提取第二预测实体集的方法流程图；
- [0030] 图4为另一个实施例中文本实体识别方法的流程图；
- [0031] 图5为一个实施例中文本实体识别方法的流程示意图；
- [0032] 图6为一个实施例中文本实体识别装置的结构框图；
- [0033] 图7为另一个实施例中文本实体识别装置的结构框图；
- [0034] 图8为又一个实施例中文本实体识别装置的结构框图；
- [0035] 图9为一个实施例中计算机设备的内部结构图。

发明实施例

本发明的实施方式

- [0036] 图1为一个实施例中文本实体识别方法的应用环境图。参照图1，该文本实体识别应用于文本实体识别系统。该文本实体识别系统包括终端110和服务器120。终端110和服务器120通过网络连接，终端110具体可以是台式终端或移动终端，移动终端具体可以是手机、平板电脑、笔记本电脑等中的至少一种。服务器120

可以用独立的服务器或者是多个服务器组成的服务器集群来实现。终端110用于获取到的待识别实体的文本上传到服务器120，服务器120用于接收待识别实体的文本，将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，然后将得到的目标实体集发送给终端110。

[0037] 在另一个实施例中，上述文本实体识别方法可以直接应用于终端110，终端110用于获取待识别实体的文本，将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集，采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集，根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集。

[0038] 如图2所示，提出了一种文本实体识别方法，该文本实体识别方法可以应用于终端，也可以应用于服务器，本实施例中以应用于终端为例说明，该文本实体识别方法具体包括以下步骤：

[0039] 步骤202，获取待识别实体的文本。

[0040] 其中，实体是指文本中的关键字。实体识别是指识别文本中的各种实体，如人名、地名、机构名或其他特有标识的关键字。文本实体识别是自然语言处理中非常重要的一项基础性技术。在一个音乐应用的场景中，通过识别文本中的实体来为用户查询想要听的音乐。比如，输入“我想听刘德华的歌”，通过识别实体“刘德华”就可以在相应的音乐曲库中查找到刘德华相应的歌曲列表。

[0041] 在一个实施例中，待识别实体的文本的获取方式可以是直接获取用户输入的文本，也可以是通过获取用户语音，然后通过语音转文字的方式得到相应的文本，当然也可以是其他方式。

[0042] 步骤204，将文本作为实体识别模型的输入，获取实体识别模型输出的第一预测实体集。

[0043] 其中，实体识别模型是指训练得到的用于对文本中的实体进行识别的模型。在一个实施例中，实体识别模型是基于深度学习方法训练得到的，比如，深度卷积神经网络。通过将文本作为实体识别模型的输入，然后得到输出的第一预测

实体集，第一预测实体集是指预测得到的实体的集合。

[0044] 步骤206，采用分词机制对文本进行分词处理，根据分词处理结果提取出第二预测实体集。

[0045] 其中，分词机制是指用于对文本进行分词处理的机制，比如，分词器。在一个实施例中，分词机制采用jieba（结巴）分词器，当然也可以采用其他分词器。分词是指将文本切分为一个个的词语。比如，假设文本为“今天温度高吗”，相应地分词处理结果为“今天 温度 高 吗”。在使用分词处理得到分词处理结果后，提取分词处理结果中的实体。在一个实施例中，根据词性来提取出实体，假设将名词作为实体，那么提取一句话中的名词作为实体，比如，“今天 温度 高 吗”中的“温度”为名词，那么将“温度”作为提取到的名词。在另一个实施例中，自定义设置实体类型，然后根据实体类型来提取出想要的实体，比如，假设“今天”是实体库中的实体，可以将今天的实体类型设置为time，并存储到实体库中，在提取实体时，根据实体库中自定义的实体类型来提取得到相应的实体。将根据分词处理结果提取出的实体加入第二预测实体集。第二预测实体集是指提取到的实体的集合。由于实体识别模型不一定在所有情况下都能准确地识别到实体，通过采用分词机制提取到文本的关键词，作为补充机制在弥补实体识别模型的不足。

[0046] 步骤208，根据第一预测实体集和第二预测实体集确定识别得到的目标实体集，目标实体集为目标实体的集合。

[0047] 其中，在确定了第一预测实体集和第二预测实体集后，就可以确定识别得到的目标实体集。目标实体集是指识别得到的目标实体的集合。在一个实施例中，将第一预测实体集和第二预测实体集的交集作为目标实体集，即将第一预测实体集和第二预测实际集中的重合的实体作为目标实体。在另一个实施例中，将第一预测实体集和第二预测实体集的并集作为目标实体集。通过将采用分词机制提取出的第二预测实体集作为后处理机制来对实体识别模型得到的第一预测实体集进行补充和确认，有利于得到更多、更可信、更准确的实体，从而提高了实体识别的准确度。

[0048] 上述文本实体识别方法，通过获取待识别实体的文本，将文本作为实体识别模

型的输入，获取实体识别模型输出的第一预测实体集，同时采用分词机制对文本进行分词处理，根据分词处理结果提取出第二预测实体集，然后根据第一预测实体集和第二预测实体集确定识别得到的目标实体集。上述文本实体识别方法在得到第一预测实体集后，通过分词机制提取出第二预测实体集，然后采用第二预测实体集对第一预测实体集进行补充识别，即采用实体识别模型和分词机制共同来确定目标实体集，能够得到更多、更可信、更准确的实体，从而有利于提高文本实体识别的准确率。

[0049] 如图3所示，在一个实施例中，采用分词机制对文本进行分词处理，根据分词处理结果提取出第二预测实体集，包括：

[0050] 步骤206A，通过分词机制对文本进行分词处理，得到多个词语和相应的词语类型；

[0051] 其中，分词机制对应的词典是基于实体库建立的，将实体库中的实体作为词典加入到分词机制中。实体库中包括实体和相应的实体类型。实体类型可以根据业务需求自定义设置，这样也便于后续根据实体类型进行实体的提取。

[0052] 对文本进行分词处理得到多个词语，每个词语都对应有相应的词语类型。在一个实施例中，将词性作为词语类型。对于实体库中的实体可以自定义相应的实体类型，并将实体类型作为该词的词语类型。比如，自定义“今天”这个词语的实体类型为“time”，将“time”作为今天的词语类型，对“今天温度高吗”进行分词后，得到“今天（time）温度（n）高（adj）吗（y）”，其中，括号里面的time、n、adj、y分别表示相应的词语类型。time为自定义的词语类型，n为名词，adj为形容词，y为语气助词。由于time为实体类型，所以后续根据词语类型就可以确定哪个词语是实体。

[0053] 步骤206B，根据词语类型从多个词语中提取出符合实体类型的词语，得到第二预测实体集。

[0054] 其中，通过预先对每个实体的类型进行自定义，获取到多个词语后，根据每个词语的词语类型来确定该词语是否为实体库中的实体，若是，则将该词语作为第二预测实体集中的实体。上述根据分词机制提取出实体，作为实体识别模型的补充机制，有利于提高实体识别的准确度。

- [0055] 在一个实施例中，根据第一预测实体集和第二预测实体集确定识别得到的目标实体集，目标实体集是目标实体的集合，包括：获取第一预测实体集和第二预测实体集的交集，将交集集中的实体作为目标实体。
- [0056] 其中，为了提高实体识别的准确率，通过对实体识别模型识别得到的实体与分词机制提取到的实体取交集，交集集中的实体为目标实体，该目标实体为确信实体。交集部分的实体比较具有代表性，可以确定一个到多个目标实体。如：“深圳 市委办公室”与“深圳 市委 办公室”的交集为“深圳”，可以确认“深圳”为确信的实体。
- [0057] 在一个实施例中，根据第一预测实体集和第二预测实体集确定识别得到的目标实体，还包括：将第一预测实体集中的第一实体与第二预测实体集中的第二实体进行匹配；当第一实体中包含有第二实体时，将包含有第二实体的第一实体作为目标实体；当第二实体中包含有第一实体时，将包含有第一实体的第二实体作为目标实体。
- [0058] 其中，为了区分，将第一预测实体集中的实体称为“第一实体”，将第二预测实体集中的实体称为“第二实体”。为了更多更好地保留实体信息，在得到第一预测实体集和第二预测实体集后，采用最长匹配原则将第一实体与第二实体进行匹配，然后保留较长的实体。具体地，当第一实体中包含有第二实体时，将包含有第二实体的第一实体作为目标实体。比如，如果第一实体为“深圳大学城”，第二实体为“深圳”，由于第一实体中包含有“深圳”，所以保留“深圳大学城”为目标实体。同样地，当第二实体中包含有第一实体时，将包含有第一实体的第二实体作为目标实体。
- [0059] 如图4所示，在一个实施例中，上述文本实体识别方法还包括：
- [0060] 步骤210，将目标实体集中的目标实体与预设实体库中的实体进行匹配，当目标实体在预设实体库中时，进入步骤212，当目标实体不在预设实体库中时，进入步骤214。
- [0061] 其中，实体库是指存储实体的数据库，里面包含有很多个实体。实体库一般是根据业务需求自定义建立的。在预测得到目标实体后，在预设实体库中查找预测得到的目标实体是否在实体库中，如果实体库中有该实体，则说明该目标实

体为确认实体。如果实体库中没有查找到该实体，那么该实体具有一定的不确定性，返回时需要明确告知该实体为预测到的实体，并不在实体库中。

[0062] 步骤212，判定目标实体为第一类实体。

[0063] 步骤214，判定目标实体为第二类实体。

[0064] 其中，为了将确认的实体和不确定的实体进行区分，将在实体库中存在的实体称为“第一类实体”，将不在实体库中存在的实体称为“第二类实体”。在一个实施例中，如果是第一类实体，返回“确认”，如果是第二类实体，返回“预测”。

[0065] 在一个实施例中，上述文本实体识别方法还包括：当目标实体集中包含有英文实体时，获取英文实体对应的多个变形英文实体；将多个变形英文实体与预设实体库中的实体进行匹配；当多个变形英文实体中的至少一个变形英文实体在预设实体库中时，则判定对应的英文实体为第一类实体，否则，判定对应的英文实体为第二类实体。

[0066] 其中，英文实体是指以英文形式存在的实体。当目标实体为英文实体时，由于英文分为大小写，为了能够匹配的完整，将英文实体改写为多个变形英文实体。比如，将英文实体中的每个首字母大写，将第一首字符大写，将所有字母全小写，单个英文词全大写等。在得到多个变形英文实体后，将多个变形英文实体与预设实体库中的实体进行匹配，然后当多个变形英文实体中的至少一个变形英文实体在预设实体库中时，则认为该英文实体是在预设实体库中，从而判定该英文实体为第一类实体，否则，判定为第二类实体。举个例子，假设识别得到的英文实体为computer，由于实体库中存储的可能为Computer或COMPUTER，如果直接进行查询很可能查询不到，所以通过将computer进行大小写变形，然后再查询，就可以在实体库中查询到。

[0067] 在一个实施例中，上述文本实体识别方法还包括：当第一预测实体集为空时，获取第二预测实体集中的实体和相应的实体类型；获取文本的意图，根据文本的意图确定目标实体类型；根据目标实体类型从第二预测实体集中筛选出符合目标实体类型的目标实体。

[0068] 其中，当通过实体识别模型没有预测到实体时，即第一预测实体集为空时，则需要依赖分词机制提取到的第二预测实体集。但是第二预测实体集中的实体很

可能包含有非目标实体，所以为了准确地得到目标实体，还需要进行进一步筛选。文本意图是指文本识别的意图，比如，是为了搜索音乐，还是为了搜索电影等。比如，刘德华这个实体具有两个实体类型，一个是歌手，一个是演员。如果文本意图为音乐场景，那么刘德华相应的目标实体类型应该歌手。如果第二预测实体集中的目标实体为刘德华，相应的实体类型为歌手，那么刘德华就是符合目标实体类型的目标实体。如果相应的实体类型为演员，那么刘德华就不符合目标实体类型，所以为非目标实体。

[0069] 如图5所示为一个实施例中，文本实体识别方法的流程示意图。首先，获取待识别实体的文本（比如，接收用户输入的文本），然后将该文本作为实体识别模型的输入，得到输出的第一预测实体集，同时采用分词机制提取第二预测实体集，接下来，判断第一预测实体集是否为空，若否，则获取第一预测实体集与第二预测实体集的交集，保留交集集中的实体，接下来，采用最长匹配原则检查第一预测实体集中和第二预测实体集中存在包含关系的实体，留下最长的。然后，检查保留下来实体（即目标实体）中是否有英文实体，如果有，则使用英文改写机制得到多个变形英文实体，然后判断提取到的实体是否在实体库中，若没有英文实体，则直接进入判断提取到的实体是否在实体库中，若在实体库中，则返回确认，若不在实体库中，则返回预测。若第一预测实体集为空，则直接将第二预测实体集中实体作为目标实体，进入检查留下来的实体是否有英文实体的步骤。

[0070] 如图6所示，在一个实施例中，提出了一种文本实体识别装置，该装置包括：

[0071] 文本获取模块602，用于获取待识别实体的文本；

[0072] 输入输出模块604，用于将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；

[0073] 分词提取模块606，用于采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；

[0074] 确定模块608，用于根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。

[0075] 在一个实施例中，所述分词提取模块还用于通过分词机制对所述文本进行分词

处理，得到多个词语和相应的词语类型；根据所述词语类型从所述多个词语中提取出符合实体类型的词语，得到第二预测实体集。

[0076] 在一个实施例中，所述确定模块还用于获取所述第一预测实体集和所述第二预测实体集的交集，将所述交集集中的实体作为目标实体。

[0077] 在一个实施例中，所述确定模块还用于将所述第一预测实体集中的第一实体与所述第二预测实体集中的第二实体进行匹配；当所述第一实体中包含有第二实体时，将包含有第二实体的第一实体作为目标实体；当所述第二实体中包含有第一实体时，将包含有第一实体的第二实体作为目标实体。

[0078] 如图7所示，在一个实施例中，所述装置还包括：

[0079] 匹配模块610，用于将所述目标实体集中的目标实体与预设实体库中的实体进行匹配，当所述目标实体在所述预设实体库中时，则判定所述目标实体为第一类实体，当所述目标实体不在所述预设实体库中时，则判定所述目标实体为第二类实体。

[0080] 在一个实施例中，所述匹配模块还用于当所述目标实体集中包含有英文实体时，获取所述英文实体对应的多个变形英文实体，将所述多个变形英文实体与所述预设实体库中的实体进行匹配，当所述多个变形英文实体中的至少一个变形英文实体在所述预设实体库中时，则判定对应的所述英文实体为第一类实体，否则，判定对应的所述英文实体为第二类实体。

[0081] 如图8所示，在一个实施例中，所述装置还包括：

[0082] 类型获取模块612，用于当所述第一预测实体集为空时，获取所述第二预测实体集中的实体和相应的实体类型；

[0083] 实体类型确定模块614，用于获取所述文本的意图，根据所述文本的意图确定目标实体类型；

[0084] 筛选模块616，用于根据所述目标实体类型从所述第二预测实体集中筛选出符合所述目标实体类型的目标实体。

[0085] 图9示出了一个实施例中计算机设备的内部结构图。该计算机设备可以是终端，也可以是服务器。如图9所示，该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中，存储器包括非易失性存储介质和内存。该计

计算机设备的非易失性存储介质存储有操作系统，还可存储有计算机程序，该计算机程序被处理器执行时，可使得处理器实现文本实体识别方法。该内存储器中也可储存有计算机程序，该计算机程序被处理器执行时，可使得处理器执行文本实体识别方法。网络接口用于与外界进行通信。本领域技术人员可以理解，图9中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

[0086] 在一个实施例中，本申请提供的文本实体识别方法可以实现为一种计算机程序的形式，计算机程序可在如图9所示的计算机设备上运行。计算机设备的存储器中可存储组成该文本实体识别装置的各个程序模板。比如，文本获取模块602、输入输出模块604、分词提取模块606和确定模块608。

[0087] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述计算机程序被所述处理器执行时，使得所述处理器执行如下步骤：获取待识别实体的文本；将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。

[0088] 在一个实施例中，所述采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集，包括：通过分词机制对所述文本进行分词处理，得到多个词语和相应的词语类型；根据所述词语类型从所述多个词语中提取出符合实体类型的词语，得到第二预测实体集。

[0089] 在一个实施例中，所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，包括：获取所述第一预测实体集和所述第二预测实体集的交集，将所述交集中的实体作为目标实体。

[0090] 在一个实施例中，所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体，还包括：将所述第一预测实体集中的第一实体与所述第二预测实体集中的第二实体进行匹配；当所述第一实体中包含有第二实体时，将包含有第二实体的第一实体作为目标实体；当所述第二实体中包含有第一实体

时，将包含有第一实体的第二实体作为目标实体。

[0091] 在一个实施例中，在所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集之后，所述计算机程序被所述处理器执行时，还用于执行以下步骤：将所述目标实体集中的目标实体与预设实体库中的实体进行匹配，当所述目标实体在所述预设实体库中时，则判定所述目标实体为第一类实体；当所述目标实体不在所述预设实体库中时，则判定所述目标实体为第二类实体。

[0092] 在一个实施例中，所述计算机程序被所述处理器执行时，还用于执行以下步骤：当所述目标实体集中包含有英文实体时，获取所述英文实体对应的多个变形英文实体；将所述多个变形英文实体与所述预设实体库中的实体进行匹配；当所述多个变形英文实体中的至少一个变形英文实体在所述预设实体库中时，则判定对应的所述英文实体为第一类实体，否则，判定对应的所述英文实体为第二类实体。

[0093] 在一个实施例中，所述计算机程序被所述处理器执行时，还用于执行以下步骤：当所述第一预测实体集为空时，获取所述第二预测实体集中的实体和相应的实体类型；获取所述文本的意图，根据所述文本的意图确定目标实体类型；根据所述目标实体类型从所述第二预测实体集中筛选出符合所述目标实体类型的目标实体。

[0094] 一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行如下步骤：获取待识别实体的文本；将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。

[0095] 在一个实施例中，所述采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集，包括：通过分词机制对所述文本进行分词处理，得到多个词语和相应的词语类型；根据所述词语类型从所述多个词语中提取出符合实体类型的词语，得到第二预测实体集。

- [0096] 在一个实施例中，所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，包括：获取所述第一预测实体集和所述第二预测实体集的交集，将所述交集集中的实体作为目标实体。
- [0097] 在一个实施例中，所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体，还包括：将所述第一预测实体集中的第一实体与所述第二预测实体集中的第二实体进行匹配；当所述第一实体中包含有第二实体时，将包含有第二实体的第一实体作为目标实体；当所述第二实体中包含有第一实体时，将包含有第一实体的第二实体作为目标实体。
- [0098] 在一个实施例中，在所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集之后，所述计算机程序被所述处理器执行时，还用于执行以下步骤：将所述目标实体集中的目标实体与预设实体库中的实体进行匹配，当所述目标实体在所述预设实体库中时，则判定所述目标实体为第一类实体；当所述目标实体不在所述预设实体库中时，则判定所述目标实体为第二类实体。
- [0099] 在一个实施例中，所述计算机程序被所述处理器执行时，还用于执行以下步骤：当所述目标实体集中包含有英文实体时，获取所述英文实体对应的多个变形英文实体；将所述多个变形英文实体与所述预设实体库中的实体进行匹配；当所述多个变形英文实体中的至少一个变形英文实体在所述预设实体库中时，则判定对应的所述英文实体为第一类实体，否则，判定对应的所述英文实体为第二类实体。
- [0100] 在一个实施例中，所述计算机程序被所述处理器执行时，还用于执行以下步骤：当所述第一预测实体集为空时，获取所述第二预测实体集中的实体和相应的实体类型；获取所述文本的意图，根据所述文本的意图确定目标实体类型；根据所述目标实体类型从所述第二预测实体集中筛选出符合所述目标实体类型的目标实体。
- [0101] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一非易失性计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的

流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink) DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0102] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0103] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本申请专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权利要求书

- [权利要求 1] 一种文本实体识别方法，其特征在于，所述方法包括：
- 获取待识别实体的文本；
- 将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；
- 采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；
- 根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，所述目标实体集为目标实体的集合。
- [权利要求 2] 根据权利要求1所述的方法，其特征在于，所述采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集，包括：
- 通过分词机制对所述文本进行分词处理，得到多个词语和相应的词语类型；
- 根据所述词语类型从所述多个词语中提取出符合实体类型的词语，得到第二预测实体集。
- [权利要求 3] 根据权利要求1所述的方法，其特征在于，所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集，包括：
- 获取所述第一预测实体集和所述第二预测实体集的交集，将所述交集中的实体作为目标实体。
- [权利要求 4] 根据权利要求3所述的方法，其特征在于，所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体，还包括：
- 将所述第一预测实体集中的第一实体与所述第二预测实体集中的第二实体进行匹配；
- 当所述第一实体中包含有第二实体时，将包含有第二实体的第一实体作为目标实体；
- 当所述第二实体中包含有第一实体时，将包含有第一实体的第二实体作为目标实体。

- [权利要求 5] 根据权利要求1所述的方法，其特征在于，在所述根据所述第一预测实体集和所述第二预测实体集确定识别得到的目标实体集之后，还包括：
- 将所述目标实体集中的目标实体与预设实体库中的实体进行匹配，当所述目标实体在所述预设实体库中时，则判定所述目标实体为第一类实体；
- 当所述目标实体不在所述预设实体库中时，则判定所述目标实体为第二类实体。
- [权利要求 6] 根据权利要求5所述的方法，其特征在于，所述方法还包括：
- 当所述目标实体集中包含有英文实体时，获取所述英文实体对应的多个变形英文实体；
- 将所述多个变形英文实体与所述预设实体库中的实体进行匹配；
- 当所述多个变形英文实体中的至少一个变形英文实体在所述预设实体库中时，则判定对应的所述英文实体为第一类实体，否则，判定对应的所述英文实体为第二类实体。
- [权利要求 7] 根据权利要求1所述的方法，其特征在于，所述方法还包括：
- 当所述第一预测实体集为空时，获取所述第二预测实体集中的实体和相应的实体类型；
- 获取所述文本的意图，根据所述文本的意图确定目标实体类型；
- 根据所述目标实体类型从所述第二预测实体集中筛选出符合所述目标实体类型的目标实体。
- [权利要求 8] 一种文本实体识别装置，其特征在于，所述装置包括：
- 文本获取模块，用于获取待识别实体的文本；
- 输入输出模块，用于将所述文本作为实体识别模型的输入，获取所述实体识别模型输出的第一预测实体集；
- 分词提取模块，用于采用分词机制对所述文本进行分词处理，根据分词处理结果提取出第二预测实体集；
- 确定模块，用于根据所述第一预测实体集和所述第二预测实体集确定

识别得到的目标实体集，所述目标实体集为目标实体的集合。

[权利要求 9] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述计算机程序被所述处理器执行时，使得所述处理器执行如权利要求1至7中任一项所述方法的步骤。

[权利要求 10] 一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行如权利要求1至7中任一项所述方法的步骤。

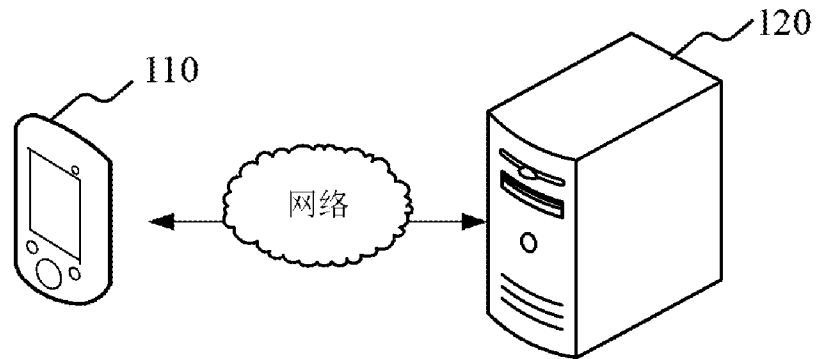


图 1

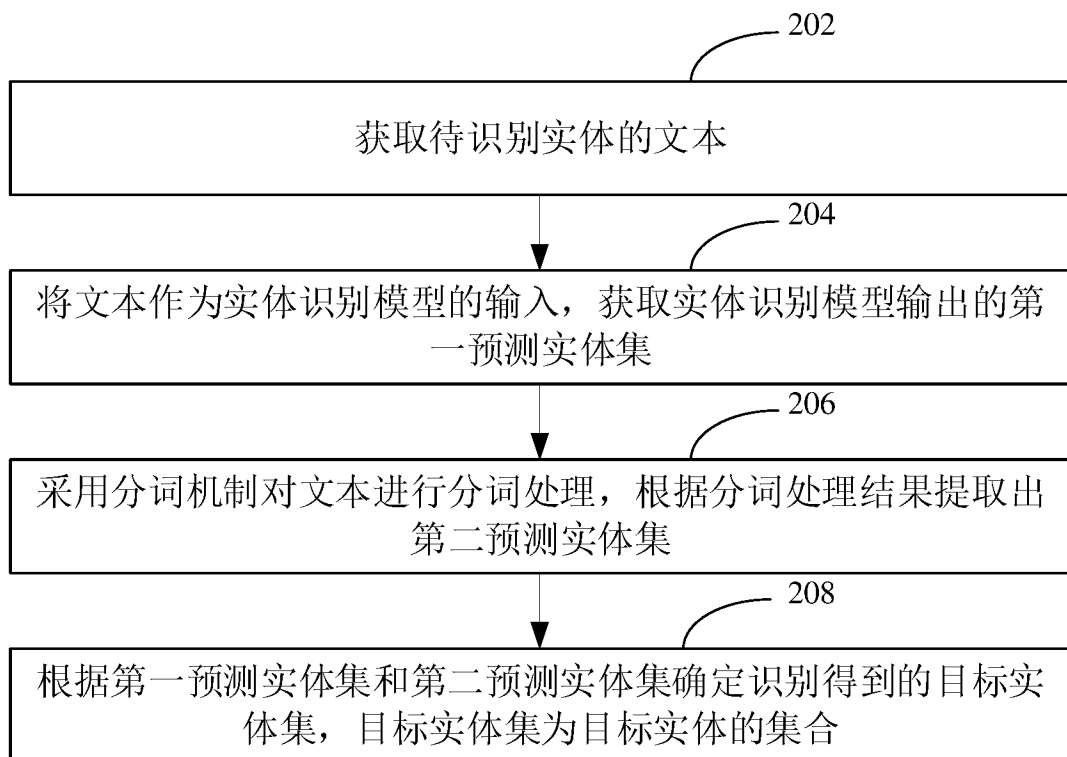


图 2

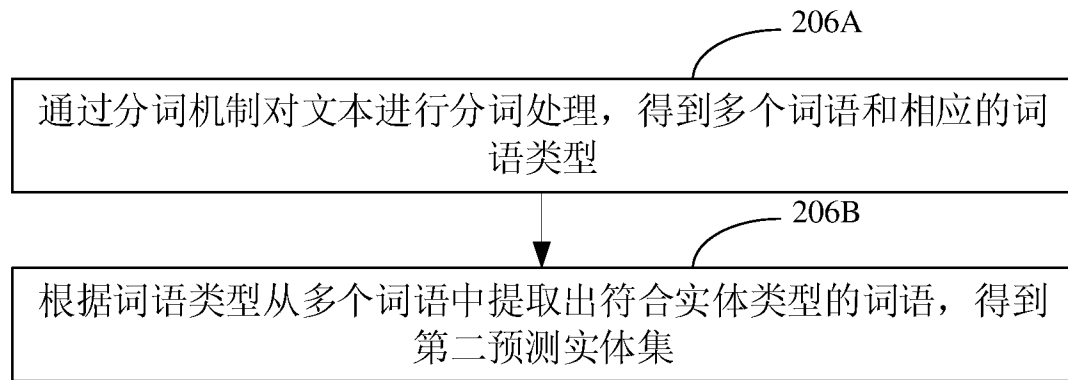


图 3

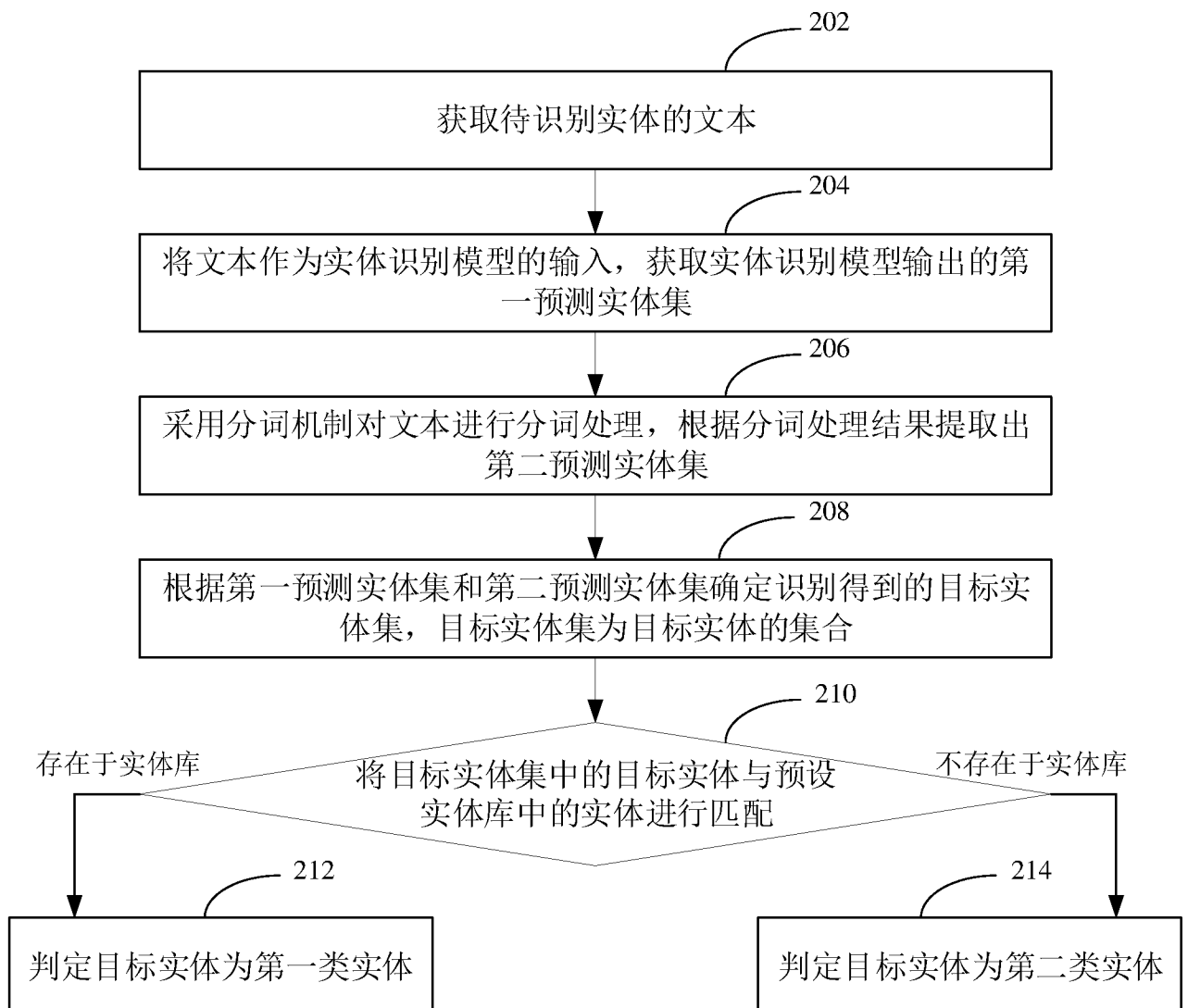


图 4

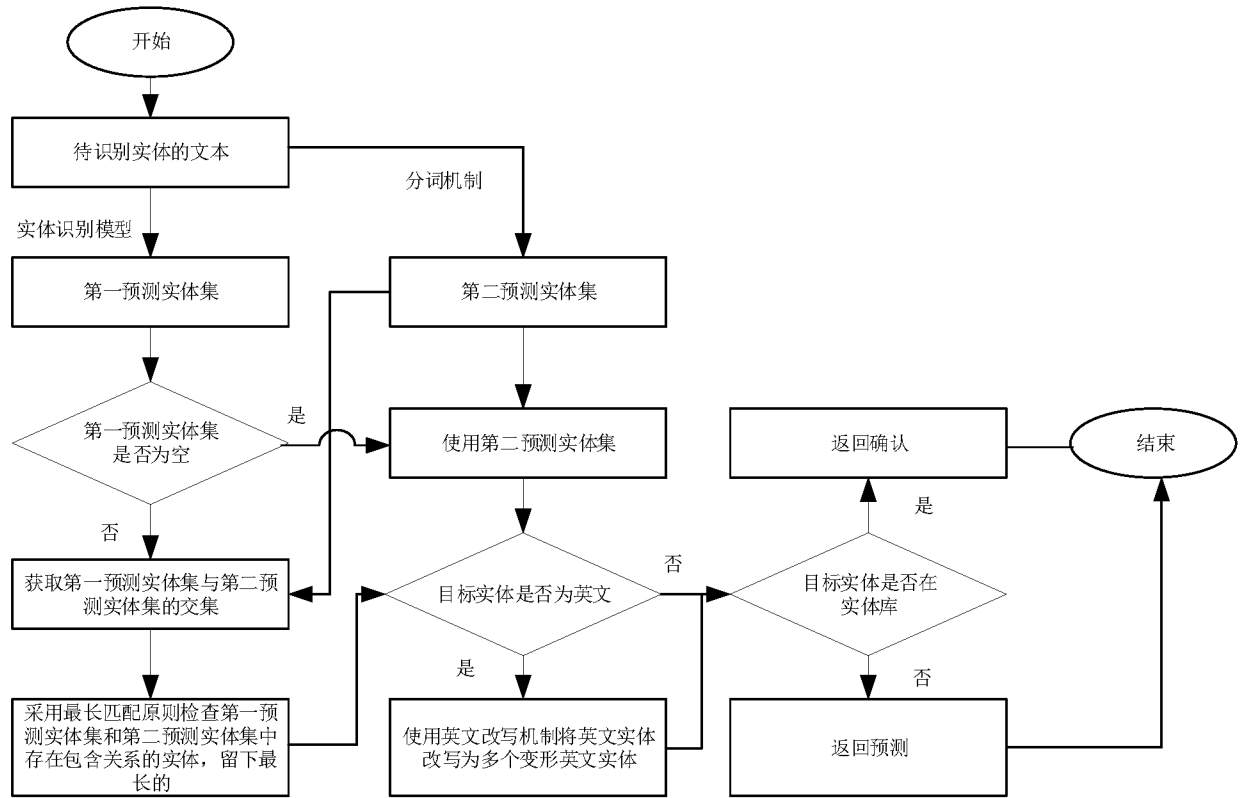


图 5

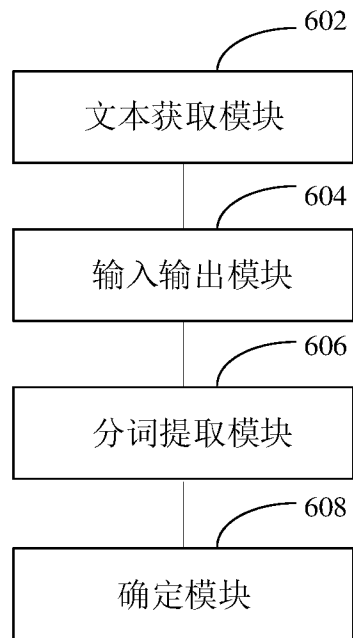


图 6

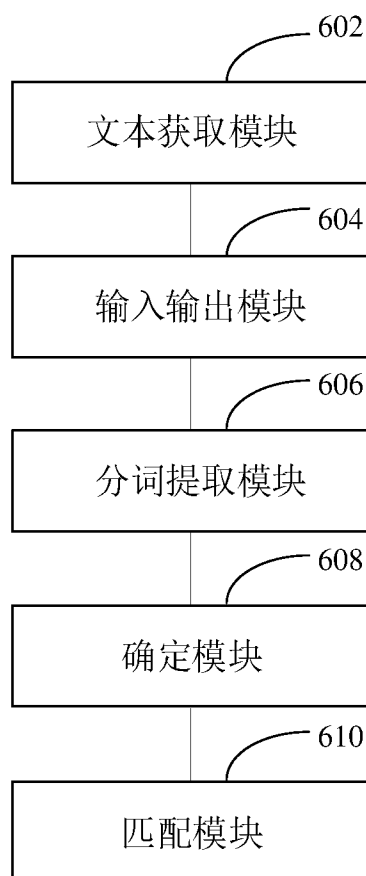


图 7

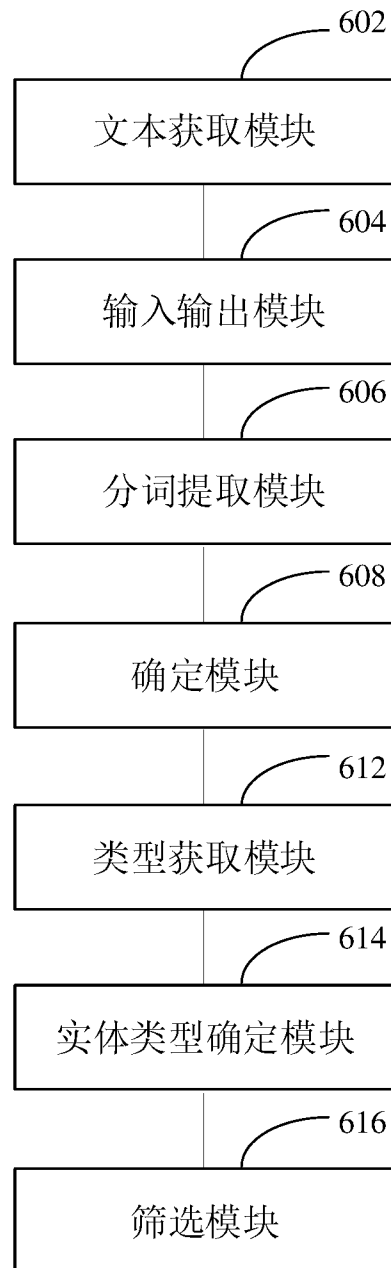


图 8

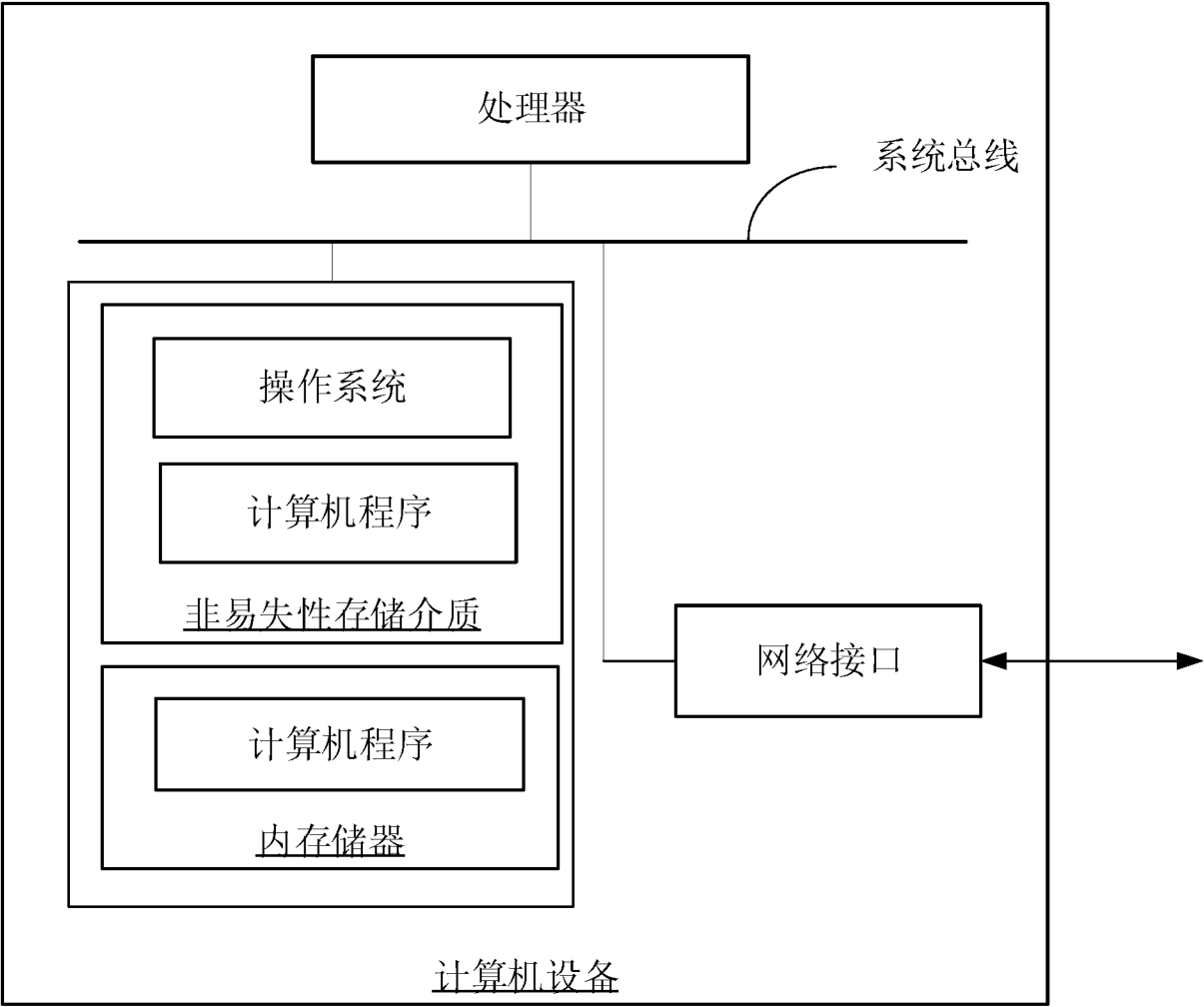


图 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/125107

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE: 文本, 实体, 识别, 预测, 模型, 分词, 属性, 词性, 交集, 并集, 匹配, text, entity, recognition, predict, model, segmentation, attribute, intersection, union, match

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 107943786 A (GUANGZHOU WANLONG SECURITIES CONSULTING CO., LTD.) 20 April 2018 (2018-04-20) description, paragraphs [0044]-[0062]	1-10
A	CN 108763201 A (NANJING UNIVERSITY) 06 November 2018 (2018-11-06) entire document	1-10
A	CN 108108344 A (BEIJING INTELLIGENT STEWARD CO., LTD.) 01 June 2018 (2018-06-01) entire document	1-10
A	CN 108460014 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 28 August 2018 (2018-08-28) entire document	1-10
A	US 2006136208 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 22 June 2006 (2006-06-22) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

21 August 2019

Date of mailing of the international search report

09 October 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/125107

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	107943786	A	20 April 2018	None			
CN	108763201	A	06 November 2018	None			
CN	108108344	A	01 June 2018	None			
CN	108460014	A	28 August 2018	None			
US	2006136208	A1	22 June 2006	KR	20060069616	A	21 June 2006

国际检索报告

国际申请号

PCT/CN2018/125107

A. 主题的分类 G06F 17/27 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																				
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNPAT, WPI, EPDOC, CNKI, IEEE: 文本, 实体, 识别, 预测, 模型, 分词, 属性, 词性, 交集, 并集, 匹配, text, entity, recognition, predict, model, segmentation, attribute, intersection, union, match																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 107943786 A (广州市万隆证券咨询顾问有限公司) 2018年 4月 20日 (2018 - 04 - 20) 说明书第[0044]-[0062]段</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 108763201 A (南京大学) 2018年 11月 6日 (2018 - 11 - 06) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 108108344 A (北京智能管家科技有限公司) 2018年 6月 1日 (2018 - 06 - 01) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 108460014 A (百度在线网络技术北京有限公司) 2018年 8月 28日 (2018 - 08 - 28) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>US 2006136208 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 2006年 6月 22日 (2006 - 06 - 22) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 107943786 A (广州市万隆证券咨询顾问有限公司) 2018年 4月 20日 (2018 - 04 - 20) 说明书第[0044]-[0062]段	1-10	A	CN 108763201 A (南京大学) 2018年 11月 6日 (2018 - 11 - 06) 全文	1-10	A	CN 108108344 A (北京智能管家科技有限公司) 2018年 6月 1日 (2018 - 06 - 01) 全文	1-10	A	CN 108460014 A (百度在线网络技术北京有限公司) 2018年 8月 28日 (2018 - 08 - 28) 全文	1-10	A	US 2006136208 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 2006年 6月 22日 (2006 - 06 - 22) 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	CN 107943786 A (广州市万隆证券咨询顾问有限公司) 2018年 4月 20日 (2018 - 04 - 20) 说明书第[0044]-[0062]段	1-10																		
A	CN 108763201 A (南京大学) 2018年 11月 6日 (2018 - 11 - 06) 全文	1-10																		
A	CN 108108344 A (北京智能管家科技有限公司) 2018年 6月 1日 (2018 - 06 - 01) 全文	1-10																		
A	CN 108460014 A (百度在线网络技术北京有限公司) 2018年 8月 28日 (2018 - 08 - 28) 全文	1-10																		
A	US 2006136208 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 2006年 6月 22日 (2006 - 06 - 22) 全文	1-10																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																				
国际检索实际完成的日期 2019年 8月 21日		国际检索报告邮寄日期 2019年 10月 9日																		
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 李玉坤 电话号码 86-(10)-53961358																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/125107

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	107943786	A	2018年 4月 20日	无	
CN	108763201	A	2018年 11月 6日	无	
CN	108108344	A	2018年 6月 1日	无	
CN	108460014	A	2018年 8月 28日	无	
US	2006136208	A1	2006年 6月 22日	KR 20060069616 A	2006年 6月 21日