

(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 4 月 16 日 (16.04.2020)



(10) 国际公布号
WO 2020/073523 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2018/124797

(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

(30) 优先权:
201811191755.9 2018 年 10 月 12 日 (12.10.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 马骏(MA, Jun); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 王少军(WANG, Shaojun); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市精英专利事务所(SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路 6009 号绿景广场 B 栋 20 层 B, 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, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: NEW WORD RECOGNITION METHOD AND APPARATUS, COMPUTER DEVICE, AND COMPUTER READABLE STORAGE MEDIUM

(54) 发明名称: 新词识别方法、装置、计算机设备及计算机可读存储介质

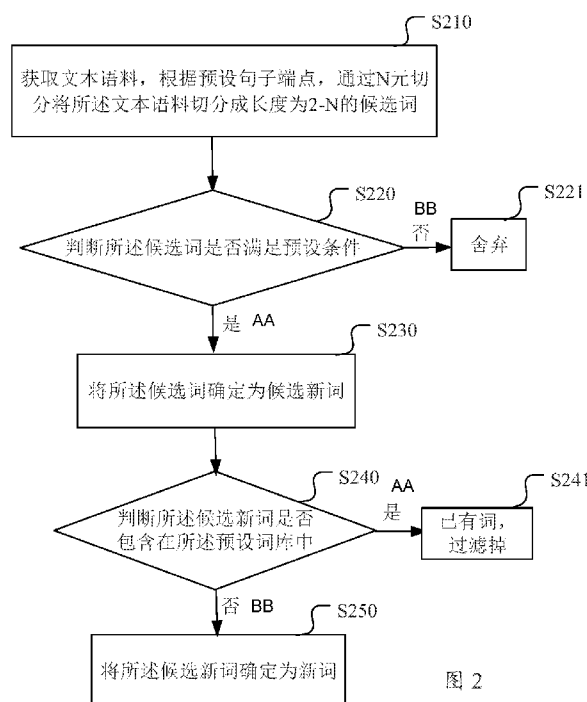


图 2

S210 Obtain a text corpus, segment the text corpus into candidate words having a length of 2-N by means of N-element segmentation according to a preset sentence endpoint
S220 Determine whether the candidate words satisfy a preset condition
S221 Discard
S230 Determine the candidate word as a candidate new word
S240 Determine whether the candidate new word is comprised in a preset vocabulary
S241 Be an existed word, filter same
S250 Determine the candidate new word as a new word
AA Yes
BB No

(57) Abstract: A new word recognition method and apparatus, a computer device, and a computer readable storage medium. The method comprises: obtaining a text corpus, segmenting the text corpus into candidate words having a length of 2-N by means of N-element segmentation according to a preset sentence endpoint, N being a natural number, $N \geq 2$ (S210); determining whether the candidate words satisfy a preset condition (S220); determining the candidate word as a candidate new word if the candidate word satisfies the preset condition (S230); determining whether the candidate new word is comprised in a preset vocabulary (S240); and determining the

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))。

candidate new word as a new word if the candidate new word is not comprised in the preset vocabulary (S250).

(57) 摘要: 一种新词识别方法、装置、计算机设备及计算机可读存储介质。方法包括: 获取文本语料, 根据预设句子端点, 通过N元切分将文本语料切分成长度为2-N的候选词, N为自然数, $N \geq 2$ (S210); 判断候选词是否满足预设条件 (S220); 若候选词满足预设条件, 将所述候选词确定为候选新词 (S230); 判断候选新词是否包含在预设词库中 (S240); 以及若候选新词不包含在预设词库中, 将所述候选新词确定为新词 (S250)。

新词识别方法、装置、计算机设备及计算机可读存储介质

本申请要求于2018年10月12日提交中国专利局、申请号为201811191755.9、申请名称为“新词识别方法、装置、计算机设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及自然语言处理技术领域，尤其涉及一种新词识别方法、装置、计算机设备及计算机可读存储介质。

背景技术

中文分词作为目前NLP（NLP，英文为 Natural Language Processing，自然语言处理）项目的基础技术，其准确率直接关系到NLP项目的最终表现。新词发现对分词系统的准确率有直接的影响。传统的新词发现技术中，通常先对文本进行分词，然后猜测未能成功匹配的剩余片段就是新词，但分词的准确性依赖于词库的完整性，因此导致新词发现的效果较差。

发明内容

本申请实施例提供了一种新词识别方法、装置、计算机设备及计算机可读存储介质，能够解决传统技术中新词发现的效果过低的问题。

第一方面，本申请实施例提供了一种新词识别方法，包括：获取文本语料，根据预设句子端点，通过N元切分将所述文本语料切分成长度为2-N的候选词，其中，N为自然数，且 $N \geq 2$ ，所述候选词是指切分所述文本语料获取的文本片段；判断所述候选词是否满足预设条件；若所述候选词满足所述预设条件，将所述候选词确定为候选新词；判断所述候选新词是否包含在所述预设词库中；以及若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词。

第二方面，本申请实施例还提供了一种新词识别装置，包括：切分单元，用于获取文本语料，根据预设句子端点，通过N元切分将所述文本语料切分成

长度为 $2-N$ 的候选词，其中， N 为自然数，且 $N \geq 2$ ，所述候选词是指切分所述文本语料获取的文本片段；判断单元，用于判断所述候选词是否满足预设条件；第一识别单元，用于若所述候选词满足所述预设条件，将所述候选词确定为候选新词；过滤单元，用于判断所述候选新词是否包含在所述预设词库中；以及第二识别单元，用于若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词。

第三方面，本申请实施例还提供了一种计算机设备，其包括存储器及处理器，所述存储器上存储有计算机程序，所述处理器执行所述计算机程序时实现所述新词识别方法。

第四方面，本申请实施例还提供了一种计算机可读存储介质，所述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所述处理器执行所述新词识别方法。

附图说明

为了更清楚地说明本申请实施例技术方案，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。

图1为本申请实施例提供的新词识别方法的应用场景示意图；

图2为本申请实施例提供的新词识别方法的流程示意图；

图3为本申请另一个实施例提供的新词识别方法的流程示意图；

图4为本申请实施例提供的新词识别装置的示意性框图；

图5为本申请另一个实施例提供的新词识别装置的示意性框图；以及

图6为本申请实施例提供的计算机设备的示意性框图。

具体实施方式

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，所描述的实施例是本申请一部分实施例，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，都属于本申请保护的范围。

请参阅图 1，图 1 为本申请实施例提供的新词识别方法的应用场景示意图。所述应用场景包括：

(1) 计算机设备。图 1 所示计算机设备是进行新词识别的设备，其上安装有进行新词识别的应用，所述计算机设备由人工进行操作。所述计算机设备可以为笔记本电脑、平板电脑、台式电脑或者服务器等电子设备。

图 1 中的各个主体工作过程如下：人工使用计算机设备进行新词识别，计算机设备上安装有新词识别的应用，计算机设备获取文本语料，根据预设句子端点，通过 N 元切分将所述文本语料切分成长度为 2-N 的候选词，判断所述候选词是否满足预设条件，若所述候选词满足所述预设条件，将所述候选词确定为候选新词，判断所述候选新词是否包含在所述预设词库中，若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词，计算机设备将识别结果显示给人工，以完成对所述文本语料的新词识别。

需要说明的是，图 1 中仅仅示意出台式电脑作为计算机设备，在实际操作过程中，计算机设备的类型不限于图 1 中所示，所述计算机设备还可以为笔记本电脑或者平板电脑等电子设备，上述新词识别方法的应用场景仅仅用于说明本申请技术方案，并不用于限定本申请技术方案。

图 2 为本申请实施例提供的新词识别方法的示意性流程图。该新词识别方法应用于图 1 中的终端中，以完成新词识别方法的全部或者部分功能。

图 2 是本申请实施例提供的新词识别方法的流程示意图。如图 2 所示，该方法包括以下步骤 S210-S250：

S210、获取文本语料，根据预设句子端点，通过 N 元切分将所述文本语料切分成长度为 2-N 的候选词，其中，N 为自然数，且 $N \geq 2$ ，其中，所述候选词是指切分所述文本语料获取的文本片段。

其中，新词是指给定一段文本，随机取一个片段，若该片段具有独立含义，且未包含在现有词库或者词典中，不为已知的词语，判断该片段为新词。判断文本中的一个片段是否为新词，可以通过给定一段文本，随机取一个片段，若

该片段左右的搭配很丰富，即该片段的左端和右端可以与不同的字或者词进行搭配以完成完整含义的表述，并且片段内部成分搭配很固定，也就是该片段经常以固定整体出现，可以判断该片段是一个词汇，如果这个词不存在现有词典中，判断该片段是一个新词。比如，在一个文本中，若“普惠金融”的左端或者右端分别和不同的字或者词搭配，以进行语义的描述，并且“普惠金融”未包含在现有词典中，判断“普惠金融”为一个新词。

N 元切分，是指对文本语料依次进行相邻 N 个汉字的分割或者划分，获取包含 N 个汉字的文本片段，比如，2 元切分，是指对文本语料依次进行相邻 2 个汉字的分割或者划分，获取包含 2 个汉字的文本片段，3 元切分，是指对文本语料依次进行相邻 3 个汉字的分割或者划分，获取包含 3 个汉字的文本片段等。比如，对文本语料“我是一个人”进行 2 元切分，获取的文本片段为“我是”、“是一”、“一个”及“个人”，对文本语料“我是一个人”进行 3 元切分，获取的文本片段为“我是一”、“是一个”及“一个人”。

文本语料，是指进行新词识别的文本所属的语言材料。所述文本语料可以为一段文字、一篇文章、一个网站的网页或者一本书籍等。所述文本语料可以是存储在移动存储器、计算机设备或者互联网中的电子版书籍或者文本，比如，以 Word 格式保存的文本，也可以是指定网站的网页等。

所述候选词是指切分所述文本语料获取的文本片段。根据预设句子端点将文本语料进行切分后，会获取多个文本片段，所述文本片段有可能是词语，也有可能不是词语，需要根据预设条件进行筛选以判断是否是词语，若所述文本片段满足预设条件，判断是词语，若所述文本片段不满足预设条件，判断为不是词语。由于所述文本片段处于成为词语的候选状态，称为候选词。2-N 的候选词，是指候选词的长度为 2 至 N，也就是候选词包含的汉字个数分别为 2、3、4...N，比如，2-5 的候选词，是指候选词的长度分别为 2、3、4 和 5，也就是候选词分别包含 2 个汉字、3 个汉字、4 个汉字及 5 个汉字。

进一步地，所述预设句子端点是指预先设置所述候选词的词边界，以这些词边界为端点将文本语料进行切分，以获取候选词。所述预设句子端点包括标点符号和预设分割端点，所述预设分割端点是指所述文本语料中除标点符号外

预先被设置为句子端点的成分，相对于标点符号，是人为将文本语料中具有独立意义的固定成分作为切分所述文本语料的词边界，属于人工句子端点。其中，文本语料一般包括文字、标点符号、回车、空格等成分。

标点符号，是指文本语料中完成完整含义描述后、起停顿作用以形成断句的句子标点符号，比如逗号、分号、双引号及句号等一般用于句子断句的符号。

所述预设分割端点包括所述文本语料中除标点符号外具有停顿或者断句作用的符号，比如，空格符和回车符等，以及具有独立语义的停用字和停用词。停用字是指文本语料中用于停顿用的字，停用词是指文本语料中用于停顿用的词，停用字和停用词一般具有独立的含义，比如，常用的停用字包括“你”、“我”、“她”及“的”等字，停用词包括“我们”、“根据”、“所述”及“作为”等词语。预设分割端点可以看作是标点符号的延伸，标点符号一般是用在句子与句子之间形成断句，预设分割端点可以看作是句子内部之间对句子成分之间形成停顿或者断句作用，可以像标点符号一样作为句子端点识别词边界。所述常用字和常用词在切分文本中作为独立特征使用，把停用字和停用词作为切分文本语料以获取候选词的左边界和右边界，所述停用字比如“在”、“我”、“的”等，停用词包括“我们”“你们”“这些”等本身已经不可拆分并且具有独立语义的固定词语，把它们作为候选词的词边界的左边界或者右边界，通过预设分割端点，可以有效地提高新词发现的准确率，提高新词识别的效率。

在切分文本语料以获得候选词时，加入属于人工句子端点的预设分割端点作为独立特征，结合标点符号作为切分所述文本语料的预设句子端点，不依赖于任何已有的词库，仅仅根据词的共同特征，将一段大规模语料中可能成词的文本片段全部提取出来，作为候选词，再对所述候选词进行满足预设条件的识别，识别出具有独立语义的候选新词，再把所有抽出来的候选新词和已有词库进行比较，筛选出现有词库中未包含的候选新词为识别出的新词，可以有效提高新词发现的准确率和召回率。

具体地，计算机设备在进行新词识别时，获取要进行新词识别的文本语料，所述文本语料可以为一段文字、一篇文章、一个网站的网页或者一本书籍等。根据包括标点符号、空格、回车、停用字和停用词等预设句子端点切分所述文

本语料，通过 N 元切分将所述文本语料切分成长度为 $2-N$ 的候选词，其中， N 为自然数，且 $N \geq 2$ ，获取切分后的候选词。比如，若 N 为 5，将获得的文本语料切分为长度分别 2、3、4 及 5 的候选词，也就是以候选词为两个字、三个字、四个字及五个字等分别进行切分，获取切分后的候选词。比如，将文本语料“我非常热衷于用统计的方法去分析汉语资料。”，进行 N 元切分（假设 $N=3$ ），预设句子端点为标点符号及“我”“的”及“非常”，切分后得到候选词包括：“热衷”、“热衷于”、“衷于”“衷于用”“于用”、“于用统”、“用统”、“用统计”、“统计”、“统计的”、“计的”、“的方”、“的方法”、“方法”、“方法去”、“法去”、“法去分”、“去分”、“去分析”、“分析”、“分析汉”、“析汉”、“析汉语”、“汉语”、“汉语资”、“语资”、“语资料”、“资料”、“资料的”、“料的”。

进一步地， N 需要根据具体文本进行设置，比如“君子之交淡如水”为七个字的词语，“百尺竿头更进一步”为八个字的词语，而一些公司名字可以有更多的字，需要根据不同的文本语料设置具体的数字 N 。进一步地，针对同一文本语料，可以设置不同的数字 N ，比较不同的识别结果，将相同的识别结果去掉，从而筛选出长粒度新词，根据长粒度新词的识别结果，出结果比较理想的新词识别。

S220、判断所述候选词是否满足预设条件。

其中，所述预设条件是指识别所述候选词为所述候选新词的条件，若所述候选词满足所述预设条件，识别所述候选词为词语，判断所述候选词为候选新词，若所述候选词不满足预设条件，识别所述候选词不为词语，判断所述候选词不为候选新词。所述预设条件所述候选词分别满足所述词频、互信息和左右信息熵各自的第一预设阈值，或者所述候选词分别满足所述词频、互信息和句子端点各自的第二预设阈值。

具体地，计算机设备将获取的文本语料进行切分，获取作为候选词的文本片段，有的候选词是词语，有的不是词语，因此需要使用预设条件对所述候选词进行筛选，将不是词语的文本片段过滤掉，将是词语的文本片段保留，以做进一步识别。因此，判断所述候选词是否满足预设条件，以识别所述候选词是否成为词语，成为词语的候选词，识别为候选新词，进入步骤 S230，若所述候

选词不满足预设条件，表明所述候选词不能成为词语，进入步骤 S221，将所述候选词过滤点，进行舍弃，以进一步缩小新词识别的范围，提高新词识别的效率和准确率。

S230、若所述候选词满足所述预设条件，将所述候选词确定为候选新词。

其中，所述候选新词是指识别为词语的候选词。通过 N 元切分，将所述文本语料切分成长度为 2-N 的候选词，其中有的候选词不能成为词语。比如，将文本语料“我非常热衷于用统计的方法去分析汉语资料。”，进行 N 元切分（假设 N=3），预设句子端点为标点符号及“我”“的”及“非常”，切分后得到候选词中“于用”、“于用统”、“用统”、“计的”、“的方”及“的方法”等候选词，由人工根据经验判断明显不能成为词语，因此，对获得的候选词需要根据预设条件进行筛选，识别所述候选词为候选新词，且所述候选新词是指识别为词语的候选词，以过滤掉不能成为词语的文本片段，缩小新词识别的范围。

具体地，计算机设备对于获得的候选词进行筛选，对于不能成为词语的候选词，需要通过预设条件进行过滤，移除不能成为词语的候选词，只保留能成为词语的候选词，以进一步进行新词识别。若所述候选词满足所述预设条件，识别所述候选词为词语，判断所述候选词为候选新词，若所述候选词不满足预设条件，识别所述候选词不为词语，判断所述候选词不为候选新词，从而进一步缩小新词识别的范围，提高新词识别的准确率、效率和召回率。

比如，请参阅表格 1。若设置候选词的最低左右信息熵的预设阈值为 1，最低互信息的预设阈值为 1，句子端点的预设阈值为 3。其中，最低左右信息熵的预设阈值为 1 是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值为 1。句子端点的预设阈值是指所述候选词的左边界或者右边界的句子端点出现的次数。请参见表格 1，在对一语料识别后，得到的结果如表格 1 所示，根据上述判断标准可知，“南山区”、“南山”、“普惠”被识别为候选新词，“去南山”不成词，被排除在候选新词外。

表格 1

候选词	词频	互信息	左右信息熵	端点数	是否成词
南山区	175	5.7548	2.2881	8	是

去南山	23	0.8256	3.3751	3	否
南山	2774	9.6310	5.7200	28	是
普惠	18	2.3811	0.8332	3	是

S240、判断所述候选新词是否包含在所述预设词库中。

其中，预设词库，也可以称为已有词库，是指包含有已经确定为词语的已知词的集合，可以为预设词典。

具体地，计算机设备通过 N 元切分将所述文本语料切分成长度为 2-N 的候选词，候选词也就是对文本语料经过切分获取的文本片段。若所述候选词满足所述预设条件，将所述候选词确定为候选新词，识别出候选新词，只是从候选词中筛选出了能够成为词语的文本片段。候选新词中包含有自然语言处理技术中已经确认为词语的词和新识别出来的词语，因此需要将已经确认为词语的词过滤掉，筛选出来为识别出来的新词。预设词库中包含有自然语言处理技术中已经确认为词语的词，所述预设词库为传统技术中已有的各种现有词典，也可以为由人工设置的词库，比如设置传统技术中已有的几本词典的集合为词库，所述预设词库中还可以包含以往识别出来的新词。使用预设词库过滤所述候选新词，以判断所述候选新词是否包含在所述预设词库中，也就是检测所述候选新词是否包含在所述预设词库中，可以采用匹配的方式，若所述候选新词能在所述预设词库中匹配得到，判断所述候选新词包含在所述预设词库中，进入步骤 S241，所述候选新词为已知的词语，过滤掉已有词；若所述候选新词未能在所述预设词库中匹配得到，判断所述候选新词未包含在所述预设词库中，所述候选新词为未知的词语，为识别出来的新词，进入步骤 S250。比如，请参见表格 1，在对一语料识别后，得到的结果如表格 1 所示，若“南山区”、“南山”及“普惠”被识别为候选新词，则用预设词典过滤获得的候选新词“南山区”、“南山”及“普惠”，若候选新词“南山区”、“南山”在所述预设词库中，“南山区”及“南山”为已知的旧词，若候选新词“普惠”未包含在所述预设词库中，判定“普惠”为识别出来的新词。

S250、若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词。

具体地，若所述候选新词未包含在所述预设词库中，认为所述候选新词不为已知的词语，所述候选新词为未知的词语，为识别出来的新词，识别出来的新词一般是在以往未见过的词语，从而完成对所给文本语料的新词识别。比如，请参见表格 1，在对一语料识别后，得到的结果如表格 1 所示，若识别出来的候选新词“普惠”未包含在所述预设词库中，判断“普惠”为识别出来的新词。

本申请实施例基于语音语义中的自然语言处理，在切分文本语料以获得候选词时，通过 N 元切分结合预设句子端点将文本语料进行准确切分，以获取长度为 2-N 候选词，不依赖于任何已有的词库，仅仅根据词的共同特征，将一段大规模语料中可能成词的文本片段全部提取出来，作为候选词，通过预设句子端点作为独立特征，作为切分所述文本语料的词边界，减少候选词的数量，提高切分的准确率和效率，再对所述候选词进行是否满足预设条件的识别，若候选词满足预设条件，识别为候选新词，作为具有独立语义的候选新词，从而缩小新词识别的范围，再把所有抽出来的候选新词和已有词库进行比较，若候选新词不包含在预设词库，识别为新词，筛选出现有词库中未包含的候选新词为识别出的新词，可以有效提高新词发现的准确率、效率和召回率。

在一个实施例中，所述判断所述候选词是否满足预设条件的步骤包括：

获取所述候选词的互信息和左右信息熵，并且获取所述候选词的词频，其中，所述左右信息熵是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值；

判断所述候选词的词频、互信息和左右信息熵是否分别大于或等于词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值；

若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，判定所述候选词满足预设条件。

其中，互信息是指候选词的内部凝聚力，也可以称为候选词的内部凝固程度或者凝合程度。互信息的公式为：

$$MI(W) = \log \frac{p(lr)}{p(l)p(r)}, \text{ 公式 (1)}$$

其中 w 表示候选词, $p(x)$ 为候选词 x 在整个语料中出现的概率, l 表示组成候选词的左字符串, r 表示组成候选词的右字符串。

比如, 在一个包含候选词“南山区”的语料中, 若“南山区”由“南山”和“区”组成, “南山”为候选词“南山区”的左字符串, “区”为候选词“南山区”的右字符串。如果“南山”和“区”是各自独立地在文本语料中随机出现, 它俩正好拼到一起的概率是多少呢? 若在所述文本语料整个 2400 万字的数据中, “南山”一共出现了 2774 次, 出现的概率约为 0.000113, “区”出现了 4797 次, 出现的概率约为 0.0001969。如果两者之间毫无关系, 它们恰好拼在了一起的概率就应该是 $0.000113 \times 0.0001969$, 约为 2.223×10^{-8} 次方。但“南山区”在语料中一共出现了 175 次, 出现概率约为 7.183×10^{-6} 次方, 是预测值的 300 多倍。同样地, 在所述文本语料中, 统计可得“去”字的出现概率约为 0.0166, 因而“去”和“南山”随机组合到了一起的理论概率值为 0.0166×0.000113 , 约为 1.875×10^{-6} , 这与“去南山”出现的真实概率很接近, 真实概率约为 1.6×10^{-5} 次方, 是预测值的 8.5 倍。结果表明, “南山区”更可能是一个有意义的搭配, 而“去南山”则更像是“去”和“南山”这两个成分偶然拼到一起的。但是, 在新词识别中, 无法判断“南山区”是“南山”加“区”得来的, 也无法判断“去南山”是“去”加上“南山”得来的。错误的切分方法会过高地估计该片段的凝合程度。如果把“南山区”看作是“南”加“山区”所得, 由此得到的凝合程度会更高一些。因此, 为了算出一个候选词的凝合程度, 采取枚举它的凝合方式, 这个候选词是由哪两部分组合而来的。令 $p(x)$ 为候选词 x 在整个语料中出现的概率, 那么定义“南山区”的凝合程度就是 $p(\text{南山区})$ 与 $p(\text{南}) * p(\text{山区})$ 比值和 $p(\text{南山区})$ 与 $p(\text{南山}) * p(\text{区})$ 的比值中的较小值, “去南山”的凝合程度则是 $p(\text{去南山})$ 分别除以 $p(\text{去}) * p(\text{南山})$ 和 $p(\text{去南}) * p(\text{山})$ 所得的互信息的较小值。可以得到, 凝合程度最高的候选词就是诸如“蝙蝠”、“蜘蛛”、“彷徨”、“忐忑”、“玫瑰”之类的词, 这些词里的每一个字几乎总是会和另一个字同时出现, 从不在其他场合中使用。

信息熵是指候选词的自由程度, 也就是所述候选词的左邻字或者右邻字的丰富程度, 所述候选词的信息熵与该候选词的左邻字或者右邻字的数量成正比, 若所述候选词能够搭配更多的左邻字或者右邻字, 所述候选词的对应信息熵越

大，若所述候选词能够搭配更少的左邻字或者右邻字，所述候选词的对应信息熵越小。候选词的信息熵又可以称为左右信息熵，一个候选词的左右信息熵，也就是该候选词的自由程度定义为它的左邻字信息熵和右邻字信息熵中的较小值。

进一步地，左邻字信息熵，又称为左信息熵，是指所述候选词的左邻字丰富程度，也就是所述候选词左侧能够搭配的字词的数量，左信息熵的公式为：

$$HL(W) = -\sum_a p(a|W) \log p(a|W), \text{ 公式 (2)}$$

其中 W 表示候选词， a 表示候选词左边的字， $p(a|W)$ 表示候选词左边出现字 a 的概率，其中， $p(a|W)$ 是条件概率。条件概率是指事件 A 在另外一个事件 B 已经发生条件下的发生概率。条件概率表示为： $p(A|B)$ ，读作“在 B 的条件下 A 的概率”。

右邻字信息熵，又称为右左信息熵，是指所述候选词的右邻字丰富程度，也就是所述候选词右侧能够搭配的字词的数量，右信息熵的公式为：

$$HR(W) = -\sum_b p(b|W) \log p(b|W), \text{ 公式 (3)}$$

其中 W 表示候选词， b 表示候选词右边的字， $p(b|W)$ 表示候选词右边出现字 b 的概率。

一个词能否出现在各种不同的环境中，具有非常丰富的左邻字集合和右邻字集合，自由程度用信息熵来表述，通过信息熵能够反映，知道一个事件的结果后平均会给你带来多大的信息量。如果某个结果的发生概率为 p ，当你知道它确实发生了，你得到的信息量就被定义为 $-\log(p)$ 。用信息熵来衡量一个候选词的左邻字集合和右邻字集合有多随机，比如，“吃葡萄不吐葡萄皮不吃葡萄倒吐葡萄皮”，“葡萄”一词出现了四次，其中左邻字分别为 {吃, 吐, 吃, 吐}，右邻字分别为 {不, 皮, 倒, 皮}。根据公式 (2) 和 (3)， “葡萄”一词的左邻字的信息熵为 $-(1/2) * \log(1/2) - (1/2) * \log(1/2) \approx 0.693$ ，它的右邻字的信息熵则为 $-(1/2) * \log(1/2) - (1/4) * \log(1/4) - (1/4) * \log(1/4) \approx 1.04$ 。可见，在这个句子中，“葡萄”一词的右邻字更加丰富一些。

词频，英文为 Term Frequency，简写为 TF，是指在一份给定的文本语料中，一个给定的词语在该文本语料中出现的次数，字词的重要性随着它在文件中出

现的次数成正比增加。获得识别新词的文本语料后，统计各个候选词出现的次数。比如，在一段 2400 万字的数据中，“南山”一共出现了 2774 次，“南山”的词频为 2774 次，“区”字则出现了 4797 次，“区”的词频为 4797 次。

具体地，能够反映候选词的词边界信息的参数包括候选词的句子端点和左右信息熵。由于句子端点和左右信息熵反映的均为候选词的词边界信息，因此句子端点和左右信息熵在候选词的识别中所起的作用是一样的，在新词识别过程中，两者满足其中一个条件即可。在本实施例中，若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，判定所述候选词满足预设条件，将所述候选词确定为候选新词为例，也就是所述候选词的词频满足词频的第一预设阈值，互信息满足互信息的第一预设阈值，左右信息熵满足左右信息熵的第一预设阈值。比如，若候选词的词频的第一预设阈值为 10，所述候选词的互信息的第一预设阈值为 1，所述候选词的左右信息熵的第一预设阈值为 1，切分的候选词分别满足词频、互信息和左右信息熵各自的第一预设阈值，是指所述候选词的词频大于或者等于 10，互信息大于或者等于 1 和左右信息熵大于或者等于 1。比如，若切分的候选词的词频大于 10、互信息大于 1 和左信息熵大于 1，则判断该候选词为候选新词，或者若切分的候选词的互信息大于 1 和右信息熵大于 1，则也可以判断该候选词为候选新词。请继续参见表格 1，在表格 1 中，由于“南山区”和“南山”的互信息和左右信息熵均大于 1，识别“南山区”和“南山”为候选新词。而由于“去南山”的互信息小于 1，“普惠”的左右信息熵小于 1，通过互信息和左右信息熵识别新词时，“去南山”和“普惠”不为候选新词。

综上所述，在表格 1 涉及的语料进行的新词识别中，“南山区”和“南山”为候选新词，“去南山”和“普惠”不为候选新词。本申请实施例中，通过将包含停用词、停用字、空格、回车的预设分割端点及标点符号作为候选词的左右词边界，通过词边界的统计，统计候选词的左右信息熵，由于细化了新词识别的粒度，通过候选词左右信息熵的统计，可有效发现低频及长粒度新词，可以有效提高新词识别的效率和准确率。

进一步地，若所述候选词的词频第一预设阈值、互信息第一预设阈值、左

右信息熵第一预设阈值设置的分别越大，识别出来的候选新词越准确，所述候选词的词频第一预设阈值、互信息第一预设阈值、左右信息熵第一预设阈值设置的分别越小，识别出来的候选新词越多。

在一个实施例中，所述判断所述候选词是否满足预设条件的步骤包括：

获取所述候选词的互信息，并且获取所述候选词的词频及所述候选词的句子端点数，所述句子端点数是指所述候选词的左端点数或者所述候选词的右端点数，所述左端点数是指所述候选词的左端点出现的次数，所述右端点数是指所述候选词的右端点出现的次数；

判断所述候选词的词频、互信息和句子端点数是否分别大于或等于词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值；

若所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值，判定所述候选词满足预设条件。

其中，所述候选词的端点是指所述候选词的左邻词边界和右邻词边界，其中，词边界是指词语的分界边缘，也就是词语的分界线，通过分界线将文本语料进行划分为不同的候选词。所述候选词的左端点是指所述候选词的左邻词边界，候选词的右端点是指所述候选词的右邻词边界。左右端点数，是指候选词的左邻词边界和右邻词边界各自出现的次数，词边界包括标点符号及预设分割端点包括的空格、回车、停用词及停用字等。为了便于统计所述候选词的左右端点数，可以用统一标识符替换所述文本语料中作为词边界的所述预设句子端点。如果词边界用统一标识符进行了替换，则左右端点数为候选词左边出现的标识符数和候选词右边出现的标识符数。比如，若统一标识符为“*”的文本，一语料为：“电影院是为观众放映电影的场所。随着电影的进步与发展，出现了专门为放映电影而建造的电影院。电影的发展使电影院的形体、尺寸、比例和声学技术都发生了很大变化。电影院必须满足电影放映的工艺要求。”用统一标识符替换句子端点后为：“电影院*为观众放映电影*场所*随着电影的进步与发展*出现了专门为放映电影而建造*电影院*电影的发展*电影院*形体、尺寸、比例和声学技术都发生了很大变化*电影院*满足电影放映*工艺要求*”，从中，可知，

候选词“电影院”的左端点出现了3次，右端点出现了4次。将包含停用词、停用字的预设分割端点、空格、回车及标点符号作为候选词的左右词边界，通过词边界的统计，统计候选词的左右端点的出现次数，由于细化了新词识别的粒度，通过候选词端点的统计，可有效发现低频及长粒度新词，可以有效提高新词识别的效率和准确率。

具体地，由于候选词的句子端点和左右信息熵反映的均为候选词的词边界信息，因此句子端点和左右信息熵在候选词的识别中所起的作用是一样的，在新词识别过程中，两者满足其中一个条件即可。在本实施例中，以所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值，判定所述候选词满足预设条件，将所述候选词确定为候选新词为例。比如，若候选词的词频第二预设阈值为10，设置所述候选词的最低左右信息熵的第二预设阈值为1，句子端点的第一预设阈值为3。其中，最低左右信息熵的第二预设阈值为1是指候选词的左邻字信息熵和右邻字信息熵中的较小值为1。句子端点第一预设阈值是指候选词的左边界或者右边界的句子端点出现的次数。请参见表格1，在对一语料识别后，得到的结果如表格1所示，若切分的候选词分别满足互信息第二预设阈值和句子端点数第一预设阈值，是指候选词的互信息大于1和句子端点出现的次数大于3，若切分的候选词的互信息大于1和句子端点出现的次数大于3，则判断该候选词为候选新词。请继续参见表格1，在表格1中，由于“南山区”、“南山”和“普惠”的互信息和句子端点数均大于3，判断“南山区”、“南山”和“普惠”为候选新词。而由于“去南山”的互信息小于1，并且句子端点数等于3，不满足大于3的条件，通过互信息和句子端点数识别新词时，“去南山”不为候选新词。根据上述判断标准可知，“南山区”、“南山”、“普惠”被识别为候选新词，“去南山”不成词，被排除在候选新词外。综上所述，在表格1涉及的语料进行的新词识别中，“南山区”、“南山”和“普惠”为候选新词，“去南山”不为候选新词。

请参阅图3，图3为本申请另一个实施例提供的新词识别方法的流程示意图。如图3所示，在该实施例中，所述判断所述候选词是否满足预设条件的步骤包括：

S211、获取所述候选词的互信息和左右信息熵，并且获取所述候选词的词频及所述候选词的句子端点数，其中，所述左右信息熵是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值；

S212、判断所述候选词的词频、互信息和左右信息熵是否分别大于或等于词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，或者所述候选词的词频、互信息和句子端点数是否分别大于或等于词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值；

S213、若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，或者所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值，判定所述候选词满足预设条件。

其中，所述词频第一预设阈值与所述词频第二预设阈值可以对应相同，所述互信息第一预设阈值与所述互信息第二预设阈值可以对应相同。

具体地，计算机设备根据所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，判定所述候选词满足预设条件，识别出来的第一候选新词，结合所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值，判定所述候选词满足预设条件，识别所述候选词为第二候选新词，取所述第一候选新词和所述第二候选新词的并集，作为识别的最终候选新词，可以提高候选新词的准确率，进入步骤 S230，以进一步识别候选新词，否则，若所述候选词不满足上述条件，进入步骤 S221，所述候选词不能成为词，舍弃掉所述候选词。请继续参阅表格 1，根据若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，判定所述候选词满足预设条件，将所述候选词确定为候选新词，识别“南山区”和“南山”为候选新词，“去南山”和“普惠”不为候选新词，而根据若所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈

值以及句子端点数第一预设阈值,判定所述候选词满足预设条件,将所述候选词确定为候选新词,“南山区”、“南山”、“普惠”被识别为候选新词,“去南山”不成词,被排除在候选新词外,将两者结合起来,识别“南山区”、“南山”、“普惠”被识别为候选新词,从而避免识别“普惠”不为候选新词,识别出“普惠”为候选新词,提高了新词识别的准确率。

本申请实施例中,通过将包含停用词、停用字、空格、回车的预设分割端点及标点符号作为候选词的左右词边界,通过词边界的统计,统计候选词的左右端点的出现次数,由于细化了新词识别的粒度,通过候选词端点的统计,可有效发现低频及长粒度新词,可以有效提高新词识别的效率和准确率。

进一步地,若所述词频的预设阈值、互信息的预设阈值、左右信息熵的预设阈值和句子端点信息的预设阈值设置的分别越大,候选新词的识别越准确,若所述词频的预设阈值、互信息的预设阈值、左右信息熵的预设阈值和句子端点信息的预设阈值设置的分别越小,识别出来的候选新词越多。

在一个实施例中,所述根据预设句子端点,通过N元切分将所述文本语料切分成长度为 $2-N$ 的候选词,其中,N为自然数,且 $N \geq 2$ 的步骤之前还包括:使用统一标识符替换所述文本语料中的所述预设句子端点。

具体地,使用统一标识符替换所述文本语料中的所述预设句子端点,是指将所述预设句子端点包括的标点符号及包括停用字、停用词、回车和空格的预设分割端点用统一标识符进行置换。比如,设置统一标识符为“*”、“#”或者“△”等标识符号,将标点符号和预设分割端点在文本中用统一标识符替代,可以方便后续句子端点数统计,从而提高所述文本语料的切分效率,提高所述文本语料的新词识别效率。比如,用“*”作为统一标识符替换所述预设句子端点中的标点符号、停用字及停用词,一段文本语料替换前为“我非常热衷于用统计的方法去分析汉语资料。”,用“*”作为统一标识符替换所述预设句子端点中的“我”、“的”及“。”后的文本语料为“*热衷于用统计*方法去分析汉语资料*”。

把停用字、停用词空格及回车作为预设分割端点,将空格、回车、停用字、停用词及标点符号替换为“*”,置换所述预设句子端点后,将文本通过N元切分成长度为 $2-N$ 的候选词,并统计各个候选词出现的次数,比如,在一段2400万

字的数据中,“南山”一共出现了 2774 次,“区”字则出现了 4797 次。

请继续参阅图 3,在该实施例中,所述若所述候选新词不包含在所述预设词库中,将所述候选新词确定为新词的步骤之后还包括:

S260、获取所述新词的长度,判断所述新词的长度是否大于或者等于预设长度阈值;

S261 若所述新词的长度大于或者等于所述预设长度阈值,识别所述新词为长粒度新词;

S262、若所述新词的长度小于所述预设长度阈值,识别所述新词为非长粒度新词。

其中,新词的长度是指新词包含的字符个数,比如词语“电影院”包含三个文字,“电影院”词语的长度为 3。

预设长度阈值是指预先设置的词语的长度临界值。预设长度阈值可以由人工进行设置。

具体地,长粒度新词是指识别的新词包含的字符数大于或者等于预设长度阈值的词。比如,若设置的预设长度阈值为 5,则识别出的新词若包含超过五个字符或者等于五个字符,识别该新词为长粒度新词。针对长粒度新词,可以根据长粒度新词的属性,进行对应处理。

请继续参阅图 3,在该实施例中,所述若所述候选新词不包含在所述预设词库中,将所述候选新词确定为新词的步骤之后还包括:

S270、获取所述新词的词频,判断所述新词的词频是否低于预设词频阈值;

S271、若所述新词出现的词频低于所述预设词频阈值,识别所述新词为低频新词;

S272、若所述新词出现的词频大于或者等于所述预设词频阈值,识别所述新词为非低频新词。

其中,低频新词是指识别出的新词的在所述文本语料中词频低于预设词频阈值。

具体地,若预设词频阈值为 10,计算机设备识别出的新词中,若识别出的新词的词频小于 10,则该新词为低频新词。由于低频新词为不常用新词,在进

行新词识别时，根据不同的文本语料，可以选择将低频新词包含或者不包含进预设词库。若选择将低频新词不包含进预设词库，可以减小预设词库的数量，提高新词识别过程中与预设词库的匹配效率，提高新词识别的效率。

需要说明的是，上述各个实施例所述的新词识别方法，可以根据需要将不同实施例中包含的技术特征重新进行组合，以获取组合后的实施方案，但都在本申请要求的保护范围之内。

请参阅图 4，图 4 为本申请实施例提供的新词识别装置的示意性框图。对应于上述新词识别方法，本申请实施例还提供一种新词识别装置。请参阅图 4，该新词识别装置包括用于执行上述新词识别方法的单元，该装置可以被配置于台式机电脑等计算机设备中。具体地，请参阅图 4，该新词识别装置 400 包括切分单元 401、判断单元 402、第一识别单元 403、过滤单元 404 及第二识别单元 405。

其中，切分单元 401，用于获取文本语料，根据预设句子端点，通过 N 元切分将所述文本语料切分成长度为 $2-N$ 的候选词，其中， N 为自然数，且 $N \geq 2$ ，所述候选词是指切分所述文本语料获取的文本片段；

判断单元 402，用于判断所述候选词是否满足预设条件；

第一识别单元 403，用于若所述候选词满足所述预设条件，将所述候选词确定为候选新词；

过滤单元 404，用于判断所述候选新词是否包含在所述预设词库中；以及

第二识别单元 405，用于若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词。

在一个实施例中，所述预设句子端点包括标点符号和预设分割端点，所述预设分割端点是指所述文本语料中除标点符号外预先被设置为句子端点的成分。

请参阅图 5，图 5 为本申请另一个实施例提供的新词识别装置的示意性框图。如图 5 所示，在该实施例中，所述判断单元 402 包括：

第一获取子单元 4021，用于获取所述候选词的互信息和左右信息熵，并且获取所述候选词的词频，其中，所述左右信息熵是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值；

第一判断子单元 4022，用于判断所述候选词的词频、互信息和左右信息熵

是否分别大于或等于词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值；

第一判定子单元 4023，用于若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，判定所述候选词满足预设条件。

在一个实施例中，所述判断单元 402 包括：

第二获取子单元，用于获取所述候选词的互信息，并且获取所述候选词的词频及所述候选词的句子端点数，所述句子端点数是指所述候选词的左端点数或者所述候选词的右端点数，所述左端点数是指所述候选词的左端点出现的次数，所述右端点数是指所述候选词的右端点出现的次数；

第二判断子单元，用于判断所述候选词的词频、互信息和句子端点数是否分别大于或等于词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值；

第二判定子单元，用于若所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值，判定所述候选词满足预设条件。

请继续参阅图 5，在该实施例中，所述装置 400 还包括：

替换单元 406，用于使用统一标识符替换所述文本语料中的所述预设句子端点。

请继续参阅图 5，在该实施例中，所述装置 400 还包括：

第三获取单元 407，用于获取所述新词的长度，判断所述新词的长度是否大于或者等于预设长度阈值；

第三识别单元 408，用于若所述新词的长度大于或者等于所述预设长度阈值，识别所述新词为长粒度新词。

请继续参阅图 5，在该实施例中，所述装置 400 还包括：

第四获取单元 409，用于获取所述新词的词频，判断所述新词的词频是否低于预设词频阈值；

第四识别单元 410，用于若所述新词的词频低于所述预设词频阈值，识别所

述新词为低频新词。

需要说明的是，所属领域的技术人员可以清楚地了解到，上述新词识别装置和各单元的具体实现过程，可以参考前述方法实施例中的相应描述，为了描述的方便和简洁，在此不再赘述。

同时，上述新词识别装置中各个单元的划分和连接方式仅用于举例说明，在其他实施例中，可将新词识别装置按照需要划分为不同的单元，也可将新词识别装置中各单元采取不同的连接顺序和方式，以完成上述新词识别装置的全部或部分功能。

上述新词识别装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 6 所示的计算机设备上运行。

请参阅图 6，图 6 是本申请实施例提供的一种计算机设备的示意性框图。该计算机设备 600 可以是台式机电脑或者平板电脑等电子设备，也可以是其他设备中的组件或者部件。

参阅图 6，该计算机设备 600 包括通过系统总线 601 连接的处理器 602、存储器和网络接口 605，其中，存储器可以包括非易失性存储介质 603 和内存储器 604。

该非易失性存储介质 603 可存储操作系统 6031 和计算机程序 6032。该计算机程序 6032 被执行时，可使得处理器 602 执行一种上述新词识别方法。

该处理器 602 用于提供计算和控制能力，以支撑整个计算机设备 600 的运行。

该内存储器 604 为非易失性存储介质 603 中的计算机程序 6032 的运行提供环境，该计算机程序 6032 被处理器 602 执行时，可使得处理器 602 执行一种上述新词识别方法。

该网络接口 605 用于与其它设备进行网络通信。本领域技术人员可以理解，图 6 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 600 的限定，具体的计算机设备 600 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。例如，在一些实施例中，计算机设备可以仅包括存储器及处理器，在

这样的实施例中，存储器及处理器的结构及功能与图 6 所示实施例一致，在此不再赘述。

其中，所述处理器 602 用于运行存储在存储器中的计算机程序 6032，以实现本申请实施例的新词识别方法。

应当理解，在本申请实施例中，处理器 602 可以是中央处理单元 (Central Processing Unit, CPU)，该处理器 602 还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中，通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

本领域普通技术人员可以理解的是实现上述实施例的方法中的全部或部分流程，是可以通过计算机程序来完成，该计算机程序可存储于一计算机可读存储介质。该计算机程序被该计算机系统至少一个处理器执行，以实现上述方法的实施例的流程步骤。

因此，本申请还提供一种计算机可读存储介质。该计算机可读存储介质可以为非易失性的计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序被处理器执行时使处理器执行以上各实施例中所描述的新词识别方法的步骤。

所述计算机可读存储介质可以是 U 盘、移动硬盘、只读存储器 (Read-Only Memory, ROM)、磁碟或者光盘等各种可以存储计算机程序的存储介质。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限

于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权利要求书

1. 一种新词识别方法，包括：

获取文本语料，根据预设句子端点，通过 N 元切分将所述文本语料切分成长度为 2-N 的候选词，其中，N 为自然数，且 $N \geq 2$ ，所述候选词是指通过切分所述文本语料获取的文本片段；

判断所述候选词是否满足预设条件；

若所述候选词满足所述预设条件，将所述候选词确定为候选新词；

判断所述候选新词是否包含在所述预设词库中；以及

若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词。

2. 根据权利要求 1 所述新词识别方法，其中，所述预设句子端点包括标点符号和预设分割端点，所述预设分割端点是指所述文本语料中除标点符号外预先被设置为句子端点的成分。

3. 根据权利要求 1 所述新词识别方法，其中，所述判断所述候选词是否满足预设条件的步骤包括：

获取所述候选词的互信息和左右信息熵，并且获取所述候选词的词频，其中，所述左右信息熵是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值；

判断所述候选词的词频、互信息和左右信息熵是否分别大于或等于词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值；

若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，判定所述候选词满足预设条件。

4. 根据权利要求 1 所述新词识别方法，其中，所述判断所述候选词是否满足预设条件的步骤包括：

获取所述候选词的互信息，并且获取所述候选词的词频及所述候选词的句子端点数，所述句子端点数是指所述候选词的左端点数或者所述候选词的右端点数，所述左端点数是指所述候选词的左端点出现的次数，所述右端点数是指所述候选词的右端点出现的次数；

判断所述候选词的词频、互信息和句子端点数是否分别大于或等于词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值；

若所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值，判定所述候选词满足预设条件。

5.根据权利要求1所述新词识别方法，其中，所述根据预设句子端点，通过N元切分将所述文本语料切分成长度为2-N的候选词的步骤之前还包括：

使用统一标识符替换所述文本语料中的所述预设句子端点。

6.根据权利要求1所述新词识别方法，其中，所述若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词的步骤之后还包括：

获取所述新词的长度，判断所述新词的长度是否大于或者等于预设长度阈值；

若所述新词的长度大于或者等于所述预设长度阈值，识别所述新词为长粒度新词。

7.根据权利要求1所述新词识别方法，其中，所述若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词的步骤之后还包括：

获取所述新词的词频，判断所述新词的词频是否低于预设词频阈值；

若所述新词的词频低于所述预设词频阈值，识别所述新词为低频新词。

8.一种新词识别装置，包括：

切分单元，用于获取文本语料，根据预设句子端点，通过N元切分将所述文本语料切分成长度为2-N的候选词，其中，N为自然数，且 $N \geq 2$ ，所述候选词是指切分所述文本语料获取的文本片段；

判断单元，用于判断所述候选词是否满足预设条件；

第一识别单元，用于若所述候选词满足所述预设条件，将所述候选词确定为候选新词；

过滤单元，用于判断所述候选新词是否包含在所述预设词库中；以及

第二识别单元，用于若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词。

9.根据权利要求8所述新词识别装置,其中,所述预设句子端点包括标点符号和预设分割端点,所述预设分割端点是指所述文本语料中除标点符号外预先被设置为句子端点的成分。

10.根据权利要求8所述新词识别装置,其中,所述判断单元包括:

第一获取子单元,用于获取所述候选词的互信息和左右信息熵,并且获取所述候选词的词频,其中,所述左右信息熵是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值;

第一判断子单元,用于判断所述候选词的词频、互信息和左右信息熵是否分别大于或等于词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值;

第一判定子单元,用于若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值,判定所述候选词满足预设条件。

11.一种计算机设备,所述计算机设备包括存储器以及与所述存储器相连的处理器;所述存储器用于存储计算机程序;所述处理器用于运行所述存储器中存储的计算机程序,所述处理器执行所述计算机程序时实现以下步骤:

获取文本语料,根据预设句子端点,通过N元切分将所述文本语料切分成长度为2-N的候选词,其中,N为自然数,且 $N \geq 2$,所述候选词是指通过切分所述文本语料获取的文本片段;

判断所述候选词是否满足预设条件;

若所述候选词满足所述预设条件,将所述候选词确定为候选新词;

判断所述候选新词是否包含在所述预设词库中;以及

若所述候选新词不包含在所述预设词库中,将所述候选新词确定为新词。

12.根据权利要求11所述计算机设备,其中,所述预设句子端点包括标点符号和预设分割端点,所述预设分割端点是指所述文本语料中除标点符号外预先被设置为句子端点的成分。

13.根据权利要求11所述计算机设备,其中,所述判断所述候选词是否满足预设条件的步骤包括:

获取所述候选词的互信息和左右信息熵，并且获取所述候选词的词频，其中，所述左右信息熵是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值；

判断所述候选词的词频、互信息和左右信息熵是否分别大于或等于词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值；

若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值，判定所述候选词满足预设条件。

14.根据权利要求 11 所述新词识别方法，其中，所述判断所述候选词是否满足预设条件的步骤包括：

获取所述候选词的互信息，并且获取所述候选词的词频及所述候选词的句子端点数，所述句子端点数是指所述候选词的左端点数或者所述候选词的右端点数，所述左端点数是指所述候选词的左端点出现的次数，所述右端点数是指所述候选词的右端点出现的次数；

判断所述候选词的词频、互信息和句子端点数是否分别大于或等于词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值；

若所述候选词的词频、互信息和句子端点数分别大于或等于所述词频第二预设阈值、互信息第二预设阈值以及句子端点数第一预设阈值，判定所述候选词满足预设条件。

15.根据权利要求 11 所述新词识别方法，其中，所述根据预设句子端点，通过 N 元切分将所述文本语料切分成长度为 2-N 的候选词的步骤之前还包括：

使用统一标识符替换所述文本语料中的所述预设句子端点。

16. 根据权利要求 11 所述新词识别方法，其中，所述若所述候选新词不包含在所述预设词库中，将所述候选新词确定为新词的步骤之后还包括：

获取所述新词的长度，判断所述新词的长度是否大于或者等于预设长度阈值；

若所述新词的长度大于或者等于所述预设长度阈值，识别所述新词为长粒度新词。

17. 根据权利要求 11 所述新词识别方法, 其中, 所述若所述候选新词不包含在所述预设词库中, 将所述候选新词确定为新词的步骤之后还包括:

获取所述新词的词频, 判断所述新词的词频是否低于预设词频阈值;

若所述新词的词频低于所述预设词频阈值, 识别所述新词为低频新词。

18. 一种计算机可读存储介质, 所述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所述处理器执行如下步骤:

获取文本语料, 根据预设句子端点, 通过 N 元切分将所述文本语料切分成长度为 $2-N$ 的候选词, 其中, N 为自然数, 且 $N \geq 2$, 所述候选词是指通过切分所述文本语料获取的文本片段;

判断所述候选词是否满足预设条件;

若所述候选词满足所述预设条件, 将所述候选词确定为候选新词;

判断所述候选新词是否包含在所述预设词库中; 以及

若所述候选新词不包含在所述预设词库中, 将所述候选新词确定为新词。

19. 根据权利要求 18 所述计算机可读存储介质, 其中, 所述预设句子端点包括标点符号和预设分割端点, 所述预设分割端点是指所述文本语料中除标点符号外预先被设置为句子端点的成分。

20. 根据权利要求 18 所述计算机可读存储介质, 其中, 所述判断所述候选词是否满足预设条件的步骤包括:

获取所述候选词的互信息和左右信息熵, 并且获取所述候选词的词频, 其中, 所述左右信息熵是指所述候选词的左邻字信息熵和右邻字信息熵中的较小值;

判断所述候选词的词频、互信息和左右信息熵是否分别大于或等于词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值;

若所述候选词的词频、互信息和左右信息熵分别大于或者等于所述词频第一预设阈值、互信息第一预设阈值以及左右信息熵第一预设阈值, 判定所述候选词满足预设条件。

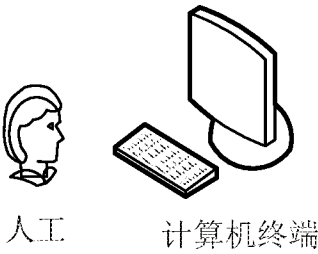


图 1

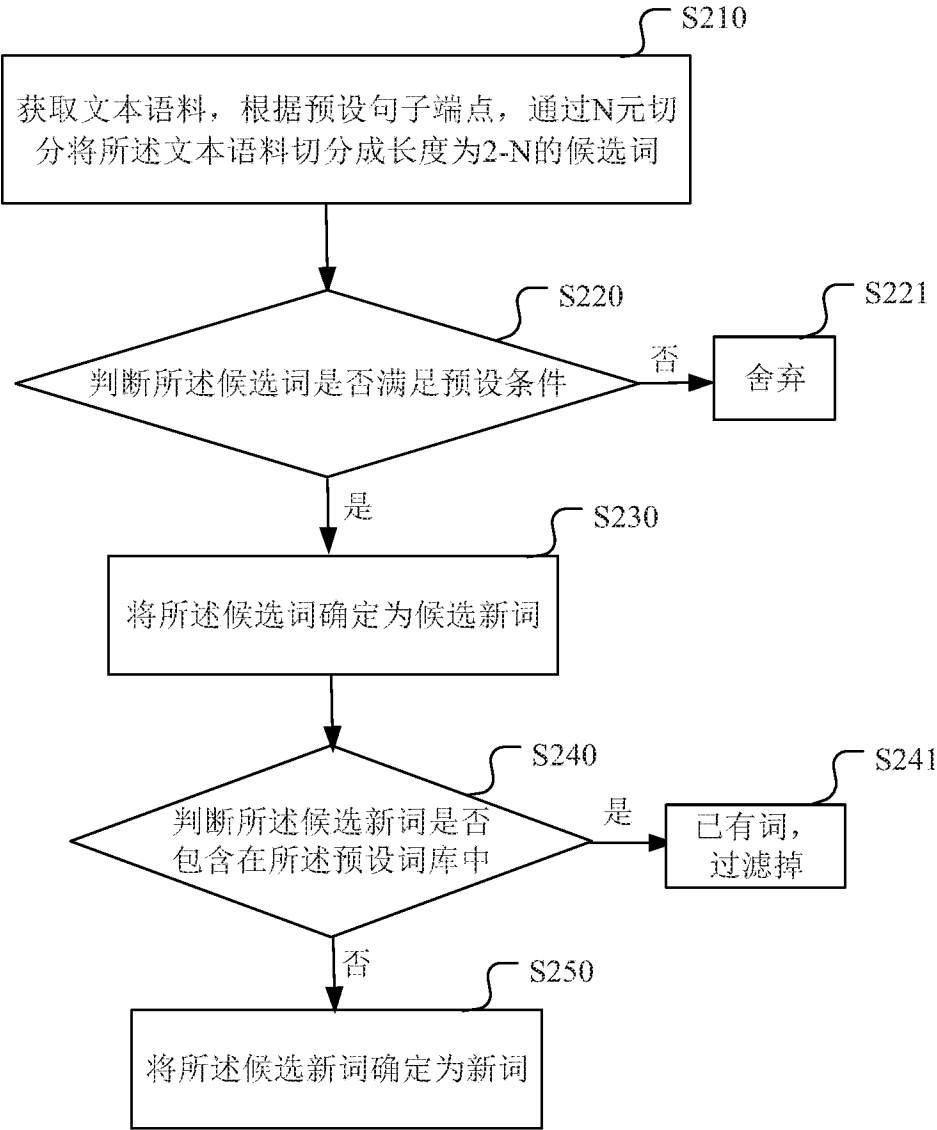


图 2

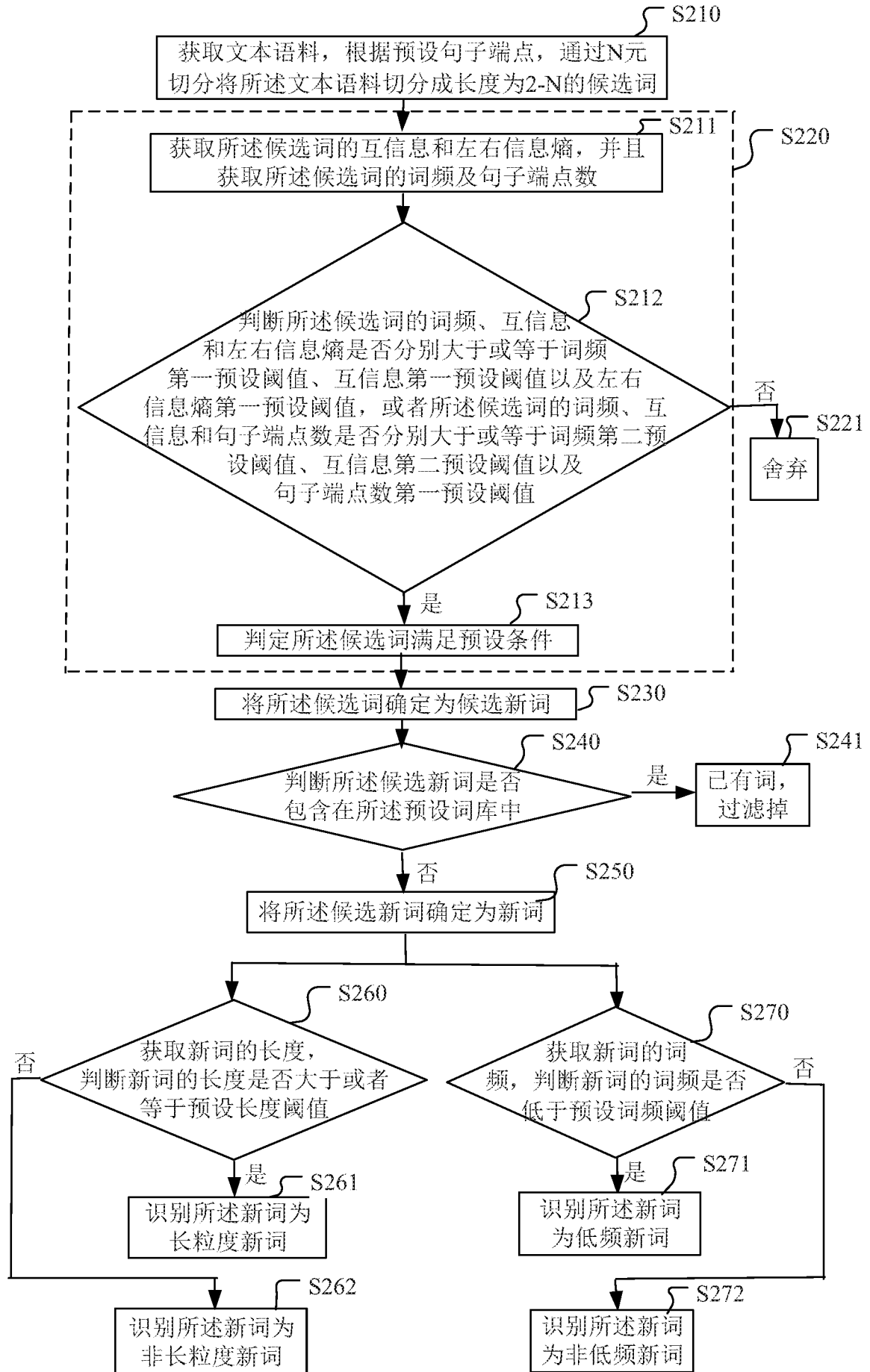


图 3

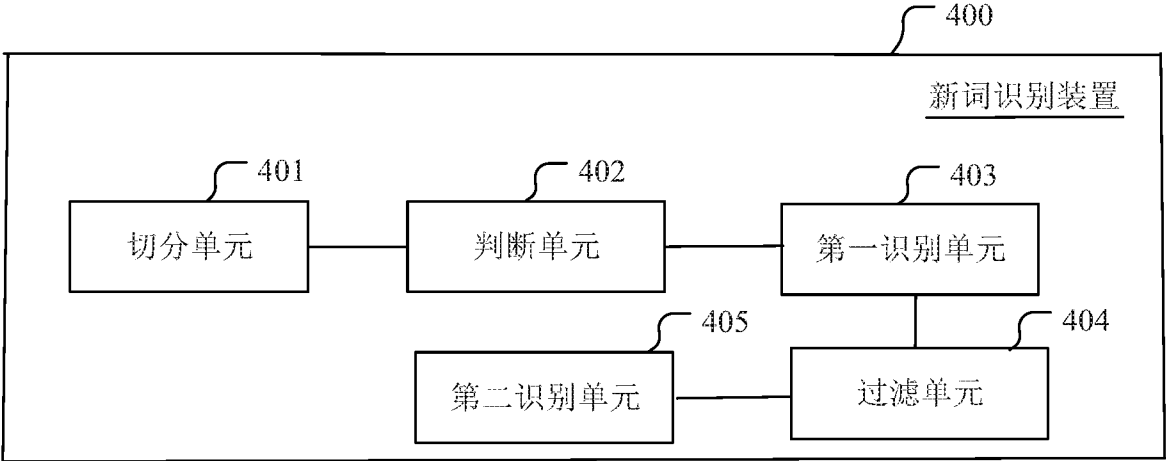


图 4

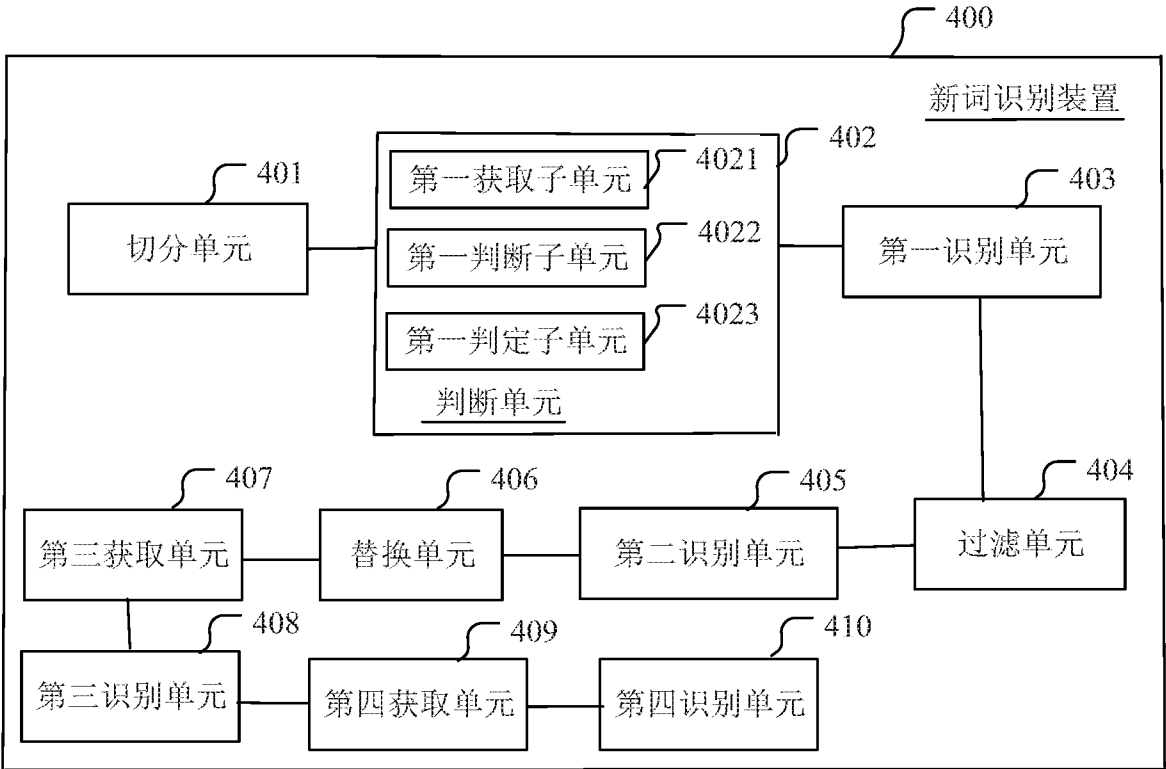


图 5

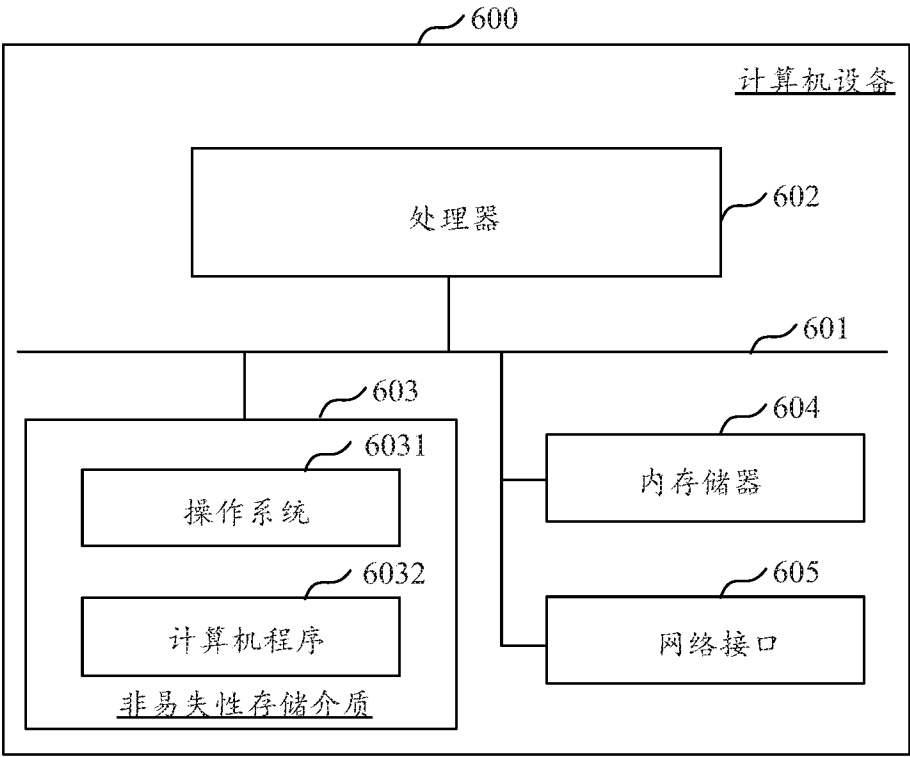


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/124797

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, EPODOC, WPI, CNKI, IEEE: 新词, 识别, 切分, 候选, 阈值, 词库, 词频, 频率, new, fresh, word, recognition, identify, segmentation, candidate, threshold, library, frequency

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	周清 (ZHOU, Qing). "网络新词发现算法研究 (Non-official translation: Research on Network New Word Discovery Algorithm)" 中国优秀硕士学位论文全文数据库(电子期刊)工程科技I辑 (Chinese Master's Theses Full-text Database (Electronic Journal), Information Science & Technology), 15 August 2017 (2017-08-15), ISSN: 1674-0246, abstract, and main body, sections 4.1-4.6	1, 2, 5-9, 11, 12, 15-19
A	CN 106970919 A (BEIJING GRIDSUM TECHNOLOGY CO., LTD.) 21 July 2017 (2017-07-21) entire document	1-20
A	CN 107180025 A (BEIJING QIYI CENTURY SCIENCE & TECHNOLOGY CO., LTD.) 19 September 2017 (2017-09-19) entire document	1-20
A	CN 108536667 A (CHINA MOBILE GROUP GUANGDONG CO., LTD. ET AL.) 14 September 2018 (2018-09-14) entire document	1-20
A	CN 107291684 A (HUAWEI TECHNOLOGIES CO., LTD.) 24 October 2017 (2017-10-24) entire document	1-20



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

28 June 2019

Date of mailing of the international search report

10 July 2019

Name and mailing address of the ISA/CN

National Intellectual Property Administration, PRC (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Authorized officer

Facsimile No. (86-10)62019451

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/124797

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106970919	A	21 July 2017	None			
CN	107180025	A	19 September 2017	None			
CN	108536667	A	14 September 2018	None			
CN	107291684	A	24 October 2017	WO	2017177809	A1	19 October 2017
				US	2019018836	A1	17 January 2019
				EP	3416064	A1	19 December 2018

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, EPDOC, WPI, CNKI, IEEE: 新词, 识别, 切分, 候选, 阈值, 词库, 词频, 频率, new, fresh, word, recognition, identify, segmentation, candidate, threshold, library, frequency		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	周清. “网络新词发现算法研究” 中国优秀硕士学位论文全文数据库 (电子期刊) 信息科技辑, 2017年 8月 15日 (2017 - 08 - 15), ISSN: 1674-0246, 摘要, 正文第4.1-4.6节	1, 2, 5-9, 11, 12, 15-19
A	CN 106970919 A (北京国双科技有限公司) 2017年 7月 21日 (2017 - 07 - 21) 全文	1-20
A	CN 107180025 A (北京奇艺世纪科技有限公司) 2017年 9月 19日 (2017 - 09 - 19) 全文	1-20
A	CN 108536667 A (中国移动通信集团广东有限公司 等) 2018年 9月 14日 (2018 - 09 - 14) 全文	1-20
A	CN 107291684 A (华为技术有限公司) 2017年 10月 24日 (2017 - 10 - 24) 全文	1-20
☐ 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		国际检索报告邮寄日期
2019年 6月 28日		2019年 7月 10日
ISA/CN的名称和邮寄地址		授权官员
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		王莹 电话号码 86-(10)-53961411

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/124797

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	106970919	A	2017年 7月 21日	无			
CN	107180025	A	2017年 9月 19日	无			
CN	108536667	A	2018年 9月 14日	无			
CN	107291684	A	2017年 10月 24日	WO	2017177809	A1	2017年 10月 19日
				US	2019018836	A1	2019年 1月 17日
				EP	3416064	A1	2018年 12月 19日