(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107451126 B
(45)授权公告日 2020.07.28

(21)申请号 201710719167.7

(22)申请日 2017.08.21

(65)同一申请的已公布的文献号
申请公布号 CN 107451126 A

(43)申请公布日 2017.12.08

(73)专利权人 广州多益网络股份有限公司
地址 510530 广东省广州市萝岗区伴河路90号1号楼
专利权人 多益网络有限公司
广东利为网络科技有限公司

(72)发明人 徐波

(74)专利代理机构 广州骏思知识产权代理有限公司 44425
代理人 吴静芝

(51)Int.Cl.
*G06F 40/284*(2020.01)
*G06F 40/289*(2020.01)
*G06F 40/247*(2020.01)

(56)对比文件
CN 106649783 A,2017.05.10,说明书第[0020]-[0080]段.
CN 106844571 A,2017.06.13,说明书第[0024]-[0087]段,图1-6.
CN 107066497 A,2017.08.18,说明书第[0063]-[0136]段.
CN 105868236 A,2016.08.17,说明书第[0036]-[078]段.
WO 2014002775 A1,2014.01.03,说明书第6-21页.

审查员 郭燕

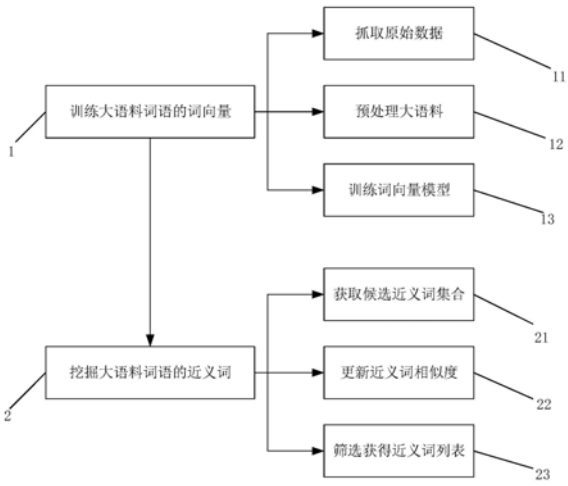
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种近义词筛选方法及系统

(57)摘要

本发明提供一种近义词筛选方法,包括以下步骤:训练大语料词语的词向量;挖掘大语料词语的近义词,具体包括:获取候选近义词集合;更新近义词相似度;筛选获得近义词列表。相比于现有技术,本发明的近义词筛选方法中,经过大语料训练得到的近义词覆盖面广,增添较新的大语料则能找到时效性好的近义词,经过近义词相互之间需要近义的原则筛选得到的近义词质量更高,为自然语言处理的语义理解增添非常有力的工具。将本发明应用于聊天机器人中,能够更好的识别用户用不同词语表达相同意思的句子,提高了机器人理解句子的水平。




CN 107451126 B

1.一种近义词筛选方法,其特征在于:包括以下步骤:

训练大语料词语的词向量,具体包括:

抓取原始数据;

预处理大语料:去除非中文字符,通过jieba分词的搜索引擎分词模式进行分词;

训练词向量模型:使用预处理后的大语料训练神经网络语言模型的词向量,设置参数,并获取大语料中每个词的词向量;

挖掘大语料词语的近义词,具体包括:

获取候选近义词集合;

更新近义词相似度:分别计算每个候选近义词集合的词语和其他候选近义词集合里全部词语的余弦相似度,取余弦相似度的平均值来更新目标词与该候选近义词的相似度;

筛选获得近义词列表。

2.根据权利要求1所述近义词筛选方法,其特征在于:所述

抓取原始数据具体为抓取各种题材文本数据作为大语料,包括各个领域的各种类型的数据。

3.根据权利要求1所述近义词筛选方法,其特征在于:所述步骤:获取候选近义词集合中,具体包括:

计算目标词的词向量和词向量模型里的其他词的词向量的余弦相似度,将余弦相似度降序排序,并输出余弦相似度在前N个词语组成候选近义词集合,所述N为正整数;

对所述候选近义词集合进行相似度阈值过滤和词性过滤,保留跟输入的目标词词性相同的词,作为候选近义词集。

4.根据权利要求1所述近义词筛选方法,其特征在于:所述步骤:筛选获得近义词列表,具体为:对候选近义词集合以更新后的余弦相似度降序排序,取余弦相似度在前N个词语或达到设定最小阈值的词语组成近义词列表,所述N为正整数。

5.一种近义词筛选系统,其特征在于:包括:

词向量训练模块,用于训练大语料词语的词向量;

所述词向量训练模块具体包括:

抓取模块;

预处理模块,用于去除非中文字符,通过jieba分词的搜索引擎分词模式进行分词;

训练模块,用于根据预处理后的大语料训练神经网络语言模型的词向量,设置参数,并获取大语料中每个词的词向量;

近义词挖掘模块,用于挖掘大语料词语的近义词;所述近义词挖掘模块具体包括:

候选集合获取模块,用于获取候选近义词集合;

更新模块,通过分别计算每个候选近义词集合的词语和其他候选近义词集合里全部词语的余弦相似度,取余弦相似度的平均值来更新目标词与该候选近义词的相似度;

筛选模块,用于筛选获得近义词列表。

6.根据权利要求5所述近义词筛选系统,其特征在于:所述

抓取模块,用于通过抓取各种题材文本数据作为大语料,包括各个领域的各种类型的数据。

7.根据权利要求5所述近义词筛选系统,其特征在于:所述候选集合获取模块具体包

括：

计算模块，用于计算目标词的词向量和词向量模型里的其他词的词向量的余弦相似度，将余弦相似度降序排序，并输出余弦相似度在前N个词语组成候选近义词集合，所述N为正整数；

过滤模块，用于对所述候选近义词集合进行相似度阈值过滤和词性过滤，保留跟输入的目标词词性相同的词，作为候选近义词集。

8.根据权利要求5所述近义词筛选系统，其特征在于：所述筛选模块具体通过对候选近义词集合以更新后的余弦相似度降序排序，取余弦相似度在前N个词语或达到设定最小阈值的词语组成近义词列表，所述N为正整数。

# 一种近义词筛选方法及系统

## 技术领域

[0001] 本发明涉及人工智能领域,特别是一种近义词筛选方法及系统。

## 背景技术

[0002] 在聊天机器人设计中,经常需要让计算机理解用户的同一句话,用不同的表达形式,以提高机器人对句子的识别水平,其中近义词的变换是最常用办法。近义词在信息抽取、问答系统、数据挖掘等基础应用中发挥重要的作用。现有的近义词挖掘方法要么词语的覆盖面窄,要么获取的近义词较陈旧,要么近义词的质量不高,这些问题都影响近义词在自然语言处理领域的应用。

[0003] 现有技术在进行近义词挖掘时所采用的方法主要包括:

[0004] 1、依靠本体词典或知识库的规则方法。例如用同义词词林,查找同义词来获取。

[0005] 2、基于搜索日志对用户行为的同义词自动挖掘的方法。例如,根据大量用户的不同输入词和相同页面的点击操作,及网页开发者对页面的关键词描述等。来挖掘用户之间用不同输入词表达出来的同义关系。

[0006] 3、利用神经网络语言模型学习词向量化表示,通过计算词向量的余弦相似度来衡量词汇语义上相似的方法。

[0007] 然而,现有技术中仍然存在以下的缺点和不足:

[0008] 1、对于依靠本体词典或知识库的规则方法,由于词典和知识库大多依赖人工构建,其时效性和覆盖面都比较差。

[0009] 2、基于搜索日志行为的方法需要利用同义词集的结构模板,可拓展性和覆盖面都不好。

[0010] 3、通过神经网络语言模型的词向量化表示的余弦相似度来衡量词汇语义上的相似度,这类方法有一定效果,但是现有的方法不能获取较高质量的近义词。神经网络语言模型的词向量能一定程度反映语义的相似性,但是获取的相似词中有一些词语在语义上并不相近,这些方法都不能把非近义词有效去除从而得到质量较高的近义词。

[0011] 综上,现有技术的近义词获取方法在获取的近义词时,不能同时达到覆盖面广,时效性好,质量较高的要求,还不能满足自然语言处理的需求,也难以提高聊天机器人理解句子的水平。

## 发明内容

[0012] 本发明的目的在于克服现有技术的缺点与不足,提供了一种近义词筛选方法及系统。

[0013] 本发明通过以下的方案实现:一种近义词筛选方法,包括以下步骤:

[0014] 训练大语料词语的词向量具体包括:

[0015] 抓取原始数据;

[0016] 预处理大语料:去除非中文字符,通过jieba分词的搜索引擎分词模式进行分词;

[0017] 训练词向量模型:使用预处理后的大语料训练神经网络语言模型的词向量,设置参数,并获取大语料中每个词的词向量;挖掘大语料词语的近义词,具体包括:

[0018] 获取候选近义词集合;

[0019] 更新近义词相似度:分别计算每个候选近义词集合的词语和其他候选近义词集合里全部词语的余弦相似度,取余弦相似度的平均值来更新目标词与该候选近义词的相似度;

[0020] 筛选获得近义词列表。

[0021] 作为本发明的进一步改进,所述抓取原始数据具体为:抓取各种题材文本数据作为大语料,包括各个领域的各种类型的数据。

[0022] 作为本发明的进一步改进,所述步骤:获取候选近义词集合中,具体包括:

[0023] 计算目标词的词向量和词向量模型里的其他词的词向量的余弦相似度,将余弦相似度降序排序,并输出余弦相似度在前N个词语组成候选近义词集合,所述N为正整数;

[0024] 对所述候选近义词集合进行相似度阈值过滤和词性过滤,保留跟输入的目标词词性相同的词,作为候选近义词集。

[0025] 作为本发明的进一步改进,所述步骤:筛选获得近义词列表,具体为:对候选近义词集合以更新后的余弦相似度降序排序,取余弦相似度在前N个词语或达到设定最小阈值的词语组成近义词列表,所述N为正整数。

[0026] 本发明还提供了一种近义词筛选系统,其包括:

[0027] 词向量训练模块,用于训练大语料词语的词向量;

[0028] 所述词向量训练模块具体包括:

[0029] 抓取模块;

[0030] 预处理模块,用于去除非中文字符,通过jieba分词的搜索引擎分词模式进行分词;

[0031] 训练模块,用于根据预处理后的大语料训练神经网络语言模型的词向量,设置参数,并获取大语料中每个词的词向量;

[0032] 近义词挖掘模块,用于挖掘大语料词语的近义词;所述近义词挖掘模块具体包括:

[0033] 候选集合获取模块,用于获取候选近义词集合;

[0034] 更新模块,通过分别计算每个候选近义词集合的词语和其他候选近义词集合里全部词语的余弦相似度,取余弦相似度的平均值来更新目标词与该候选近义词的相似度;

[0035] 筛选模块,用于筛选获得近义词列表。

[0036] 作为本发明的进一步改进,所述词向量训练模块具体包括:

[0037] 抓取模块,用于通过抓取各种题材文本数据作为大语料,包括各个领域的各种类型的数据。

[0038] 作为本发明的进一步改进,所述候选集合获取模块具体包括:

[0039] 计算模块,用于计算目标词的词向量和词向量模型里的其他词的词向量的余弦相似度,将余弦相似度降序排序,并输出余弦相似度在前N个词语组成候选近义词集合,所述N为正整数;

[0040] 过滤模块,用于对所述候选近义词集合进行相似度阈值过滤和词性过滤,保留跟输入的目标词词性相同的词,作为候选近义词集。

[0041] 作为本发明的进一步改进，所述更新模块具体通过分别计算每个候选近义词集合的词语和其他候选近义词集合里全部词语的余弦相似度，取余弦相似度的平均值来更新目标词与该候选近义词的相似度。

[0042] 作为本发明的进一步改进，所述筛选模块具体通过对候选近义词集合以更新后的余弦相似度降序排序，取余弦相似度在前N个词语或达到设定最小阈值的词语组成近义词列表，所述N为正整数。

[0043] 相比于现有技术，本发明的近义词筛选方法中，经过大语料训练得到的近义词覆盖面广，增添较新的大语料则能找到时效性好的近义词，经过近义词相互之间需要近义的原则筛选得到的近义词质量更高，为自然语言处理的语义理解增添非常有力的工具。将本发明应用于聊天机器人中，能够更好的识别用户用不同词语表达相同意思的句子，提高了机器人理解句子的水平。

[0044] 为了更好地理解和实施，下面结合附图详细说明本发明。

## 附图说明

[0045] 图1是本发明的近义词筛选方法的步骤流程图。

[0046] 图2是本发明的近义词筛选系统的模块框图。

## 具体实施方式

[0047] 以下结合实施例及附图对本发明作进一步详细的描述，但本发明的实施方式不限于此。

[0048] 请同时参阅图1，其为本发明的近义词筛选方法的步骤流程图。本发明提供了一种近义词筛选方法，包括以下步骤：

[0049] S1：训练大语料词语的词向量。

[0050] 进一步，所述步骤S1中具体包括：

[0051] S11：抓取原始数据。具体的，S11具体为：抓取各种题材文本数据作为大语料，包括各个领域的各种类型的数据，例如：各种类型的新闻文本，各种题材的小说文本，全部条目的百科文本。

[0052] S12：预处理大语料。所述步骤S12中具体为：去除非中文字符，通过jieba分词的搜索引擎分词模式进行分词，从而避免错过同一个语义的词语的不同表述。

[0053] S13：训练词向量模型。所述步骤S13中，具体为：使用预处理后的大语料训练神经网络语言模型的词向量，设置参数，并获取大语料中每个词的词向量。

[0054] S2：挖掘大语料词语的近义词。所述步骤S2中，具体包括：

[0055] S21：获取候选近义词集合。具体的所述步骤S21中包括：

[0056] S211：计算目标词的词向量和词向量模型里的其他词的词向量的余弦相似度，将余弦相似度降序排序，并输出余弦相似度在前N个词语组成候选近义词集合，所述N为正整数。比如，可以将余弦相似度前10个作为候选词。

[0057] S212：对所述候选近义词集合进行相似度阈值过滤和词性过滤，保留跟输入的目标词词性相同的词，作为候选近义词集。比如：对前10个候选词进行过滤，将相似度低于0.5的过滤等等，具体可以根据实际情况设置不同的阈值。

[0058] S22:更新近义词相似度。

[0059] 具体的,所述步骤S22具体为:分别计算每个候选近义词集合的词语和其他候选近义词集合里全部词语的余弦相似度,取余弦相似度的平均值来更新目标词与该候选近义词的相似度。

[0060] S23:筛选获得近义词列表。

[0061] 具体的,所述步骤S23具体为:对候选近义词集合以更新后的余弦相似度降序排序,取余弦相似度在前N个词语或达到设定最小阈值的词语组成近义词列表,所述N为正整数。

[0062] 以下结合具体的例子,说明本发明的近义词筛选方法的应用:

[0063] 第一、语料训练。具体的,对每个中文单词,找到和它语义接近的其它汉语单词,可以采用Word Embedding技术来实现这个语义的关联。采用的工具是Word2Vec,采用整个百度百科作为训练数据,这样就能得到每个中文单词对应的词向量,这是一种低维度向量形式的单词表示,能够表征单词的部分语义及语法含义。

[0064] 第二、近义词挖掘。对于任意两个已经用WordEmbedding形式表示的单词,我们可以简单通过计算两个向量之间的Cosine相似性,得出两个单词语义接近程度。

[0065] 于是,某个单词,我们可以从所有其它单词中找出和这个单词语义最接近的一部分单词,也就是Cosine得分最高的一批单词。例如:对于目标词,“歌曲”可以得出挖掘的结果:歌词:0.87,首歌:0.91,颂歌:0.93,曲调:0.69,进行曲:0.75,唱歌:0.58

[0066] 对某个单词W找出语义最接近的单词列表后对其进行过滤,过滤规则是:先抽取超过一定阈值的词,例如相似度超过0.5的所有词。接着,根据词性过滤,把这些单词中词性和W相同的留下来,不同的过滤掉。例如,上面的‘唱歌’是动词,因此可以被过滤。这步其实是很关键的,对于后面最终产生的句子语义一致性及可读性有很大影响。主要原因是,尽管理论上通过Word Embedding可以找到语义相似的其它单词,但是其实还是有不少看上去不合理的内容,这是Word Embedding本身产生方式决定的,增加合理的过滤措施能够极大改善句子生成质量,而根据词性过滤就是一个简单易行的方法。

[0067] 最后,对挖掘到的关键词,进行两两的关键词相似度计算,并把计算结果进行求和与平均,得到一个最终的得分。例如‘首歌’得跟其他的除了目标词之外的,‘歌词’、‘进行曲’等各个词,进行相似度计算,然后取平均值。接着对‘颂歌’进行计算。并对结果按分数高低排序。其中,排序最高的为最适合的近义词。

[0068] 请同时参阅图2,其为本发明的近义词筛选系统的模块框图。为了实现上述的方法,本发明还提供了一种近义词筛选系统,其包括:词向量训练模块1和近义词挖掘模块2。

[0069] 所述词向量训练模块1,用于训练大语料词语的词向量。

[0070] 所述近义词挖掘模块2,用于挖掘大语料词语的近义词。

[0071] 进一步,所述词向量训练模块1具体包括:抓取模块11、预处理模块12和训练模块13。

[0072] 所述抓取模块11,用于抓取原始数据,具体通过抓取各种题材文本数据作为大语料,包括各个领域的各种类型的数据。

[0073] 所述预处理模块12,用于预处理大语料,具体包括:去除非中文字符,通过jieba分词的搜索引擎分词模式进行分词。

[0074] 所述训练模块13，用于训练词向量模型，具体为：使用预处理后的大语料训练神经网络语言模型的词向量，设置参数，并获取大语料中每个词的词向量。

[0075] 具体的，所述近义词挖掘模块2具体包括：候选集合获取模块21、更新模块22和筛选模块23。

[0076] 所述候选集合获取模块21，用于获取候选近义词集合。

[0077] 进一步，所述候选集合获取模块具体包括：计算模块211和过滤模块212。

[0078] 所述计算模块211，用于计算目标词的词向量和词向量模型里的其他词的词向量的余弦相似度，将余弦相似度降序排序，并输出余弦相似度在前N个词语组成候选近义词集合，所述N为正整数。

[0079] 所述过滤模块212，用于对所述候选近义词集合进行相似度阈值过滤和词性过滤，保留跟输入的目标词词性相同的词，作为候选近义词集。

[0080] 所述更新模块22，用于更新近义词相似度，具体通过分别计算每个候选近义词集合的词语和其他候选近义词集合里全部词语的余弦相似度，取余弦相似度的平均值来更新目标词与该候选近义词的相似度。

[0081] 所述筛选模块23，用于筛选获得近义词列表，具体通过对候选近义词集合以更新后的余弦相似度降序排序，取余弦相似度在前N个词语或达到设定最小阈值的词语组成近义词列表，所述N为正整数。

[0082] 相比于现有技术，本发明的近义词筛选方法中，经过大语料训练得到的近义词覆盖面广，增添较新的大语料则能找到时效性好的近义词，经过近义词相互之间需要近义的原则筛选得到的近义词质量更高，为自然语言处理的语义理解增添非常有力的工具。将本发明应用于聊天机器人中，能够更好的识别用户用不同词语表达相同意思的句子，提高了机器人理解句子的水平。

[0083] 上述实施例为本发明较佳的实施方式，但本发明的实施方式并不受上述实施例的限制，其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化，均应为等效的置换方式，都包含在本发明的保护范围之内。

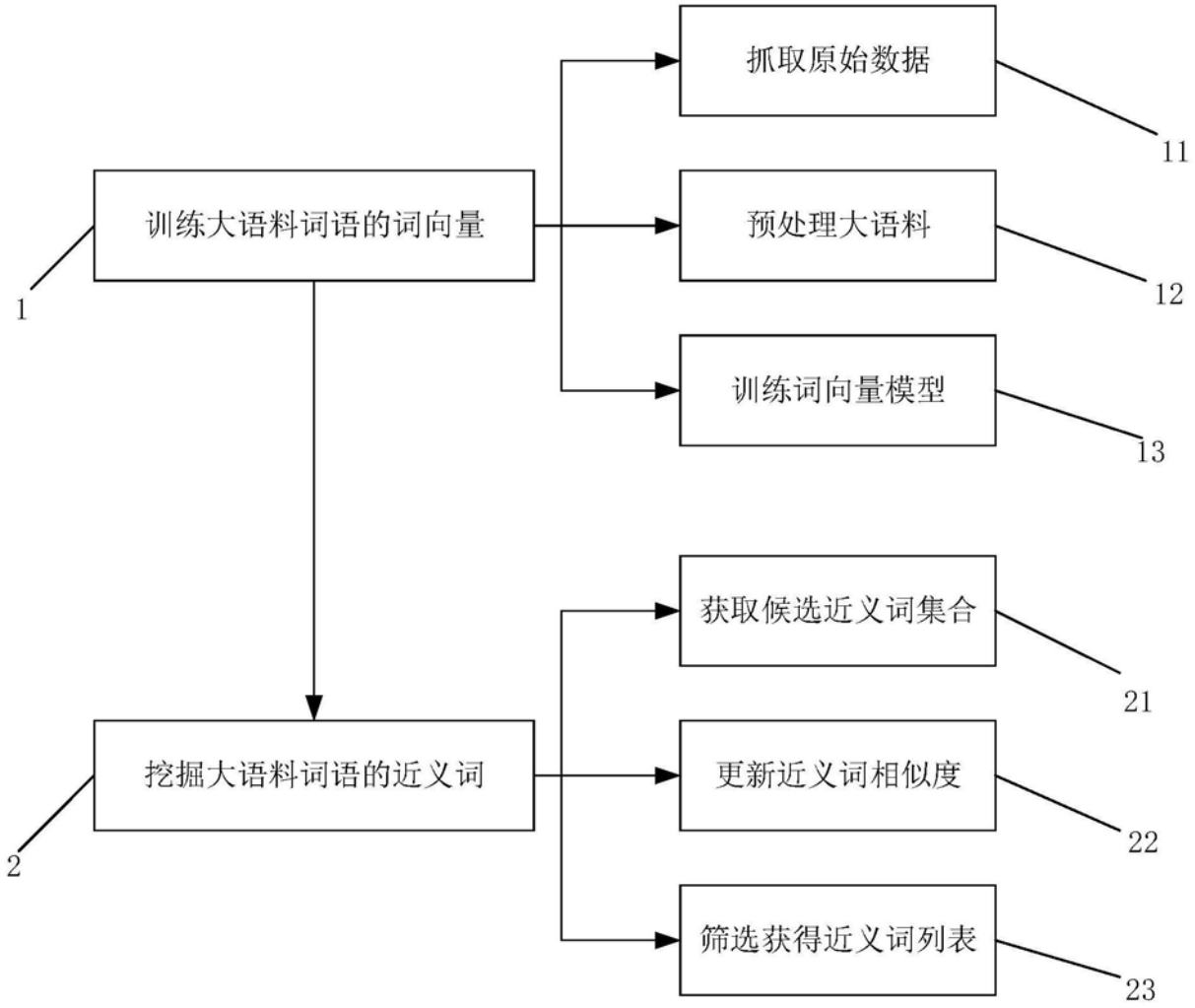

训练大语料词语的词向量
1
抓取原始数据
11
预处理大语料
12
训练词向量模型
13
挖掘大语料词语的近义词
2
获取候选近义词集合
21
更新近义词相似度
22
筛选获得近义词列表
23

图1

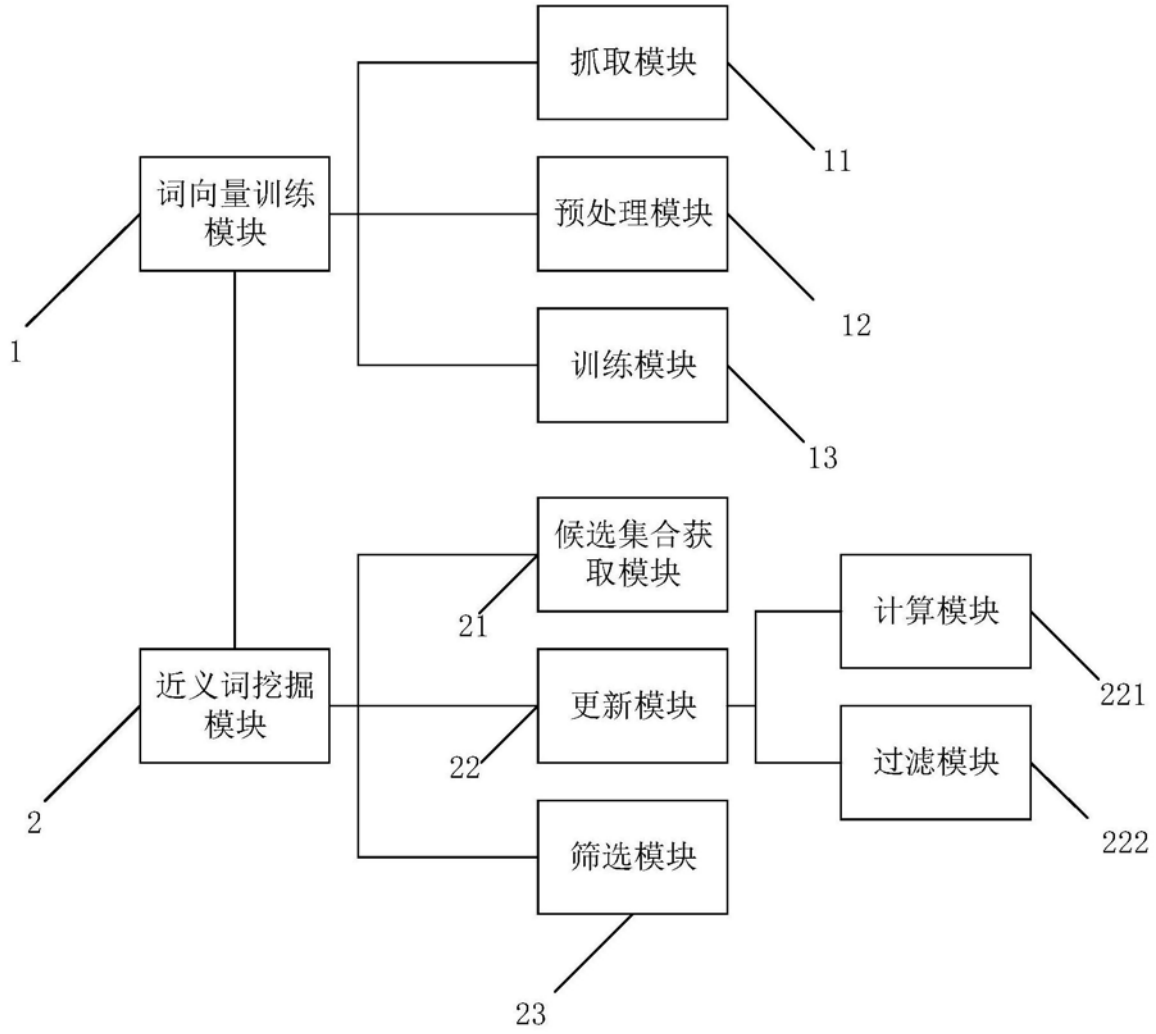

词向量训练模块
1
抓取模块
11
预处理模块
12
训练模块
13
近义词挖掘模块
2
候选集合获取模块
21
更新模块
22
筛选模块
23
计算模块
221
过滤模块
222

图2