

(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
2019 年 8 月 8 日 (08.08.2019)



(10) 国际公布号
WO 2019/149076 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2019/072081

(22) 国际申请日: 2019 年 1 月 17 日 (17.01.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810113710.3 2018 年 2 月 5 日 (05.02.2018) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 曹绍升 (CAO, Shaosheng); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。周俊 (ZHOU, Jun); 中国浙江省杭州市余杭区

文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 409, Beijing 100085 (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, 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。

(54) Title: WORD VECTOR GENERATION METHOD, APPARATUS AND DEVICE

(54) 发明名称: 词向量生成方法、装置以及设备

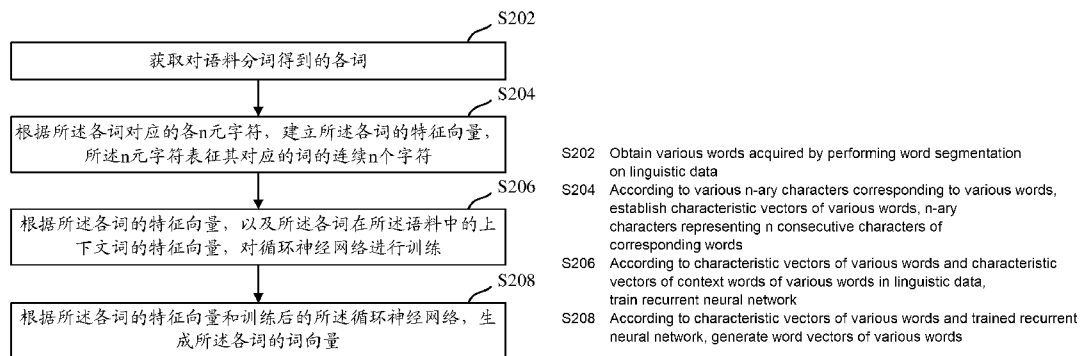


图 2

(57) Abstract: A word vector generation method, apparatus and device. The method comprises: obtaining various words acquired by performing word segmentation on linguistic data (S202); according to various n-ary characters corresponding to the various words, establishing characteristic vectors of the various words, the n-ary characters representing n consecutive characters of the corresponding words (S204); according to the characteristic vectors of the various words and characteristic vectors of context words of the various words in the linguistic data, training a recurrent neural network (S206); according to the characteristic vectors of the various words and the trained recurrent neural network, generating word vectors of the various words (S208).

(57) 摘要: 一种词向量生成方法、装置以及设备。所述方法包括: 获取对语料分词得到的各词 (S202), 根据所述各词对应的各n元字符, 建立所述各词的特征向量, 所述n元字符表征其对应的词的连续n个字符 (S204), 根据所述各词的特征向量, 以及所述各词在语料中的上下文词的特征向量, 对循环神经网络进行训练 (S206), 根据所述各词的特征向量和训练后的所述循环神经网络, 生成所述各词的词向量 (S208)。

(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))。

词向量生成方法、装置以及设备

相关申请的交叉引用

5 [01]本专利申请要求于2018年2月5日提交的、申请号为201810113710.3、发明名称为“词向量生成方法、装置以及设备”的中国专利申请的优先权，该申请的全文以引用的方式并入本文中。

技术领域

10 [02]本说明书涉及计算机软件技术领域，尤其涉及词向量生成方法、装置以及设备。

背景技术

[03]如今的自然语言处理的解决方案，大都采用基于神经网络的架构，而在这种架构下一个重要的基础技术就是词向量。词向量是将词映射到一个固定维度的向量，该向量表征了该词的语义信息。

15 [04]在现有技术中，常见的用于生成词向量的算法比如包括：谷歌公司的词向量算法、微软公司的深度神经网络算法等。

[05]基于现有技术，需要一种更准确的词向量生成方案。

发明内容

20 [06]本说明书实施例提供词向量生成方法、装置以及设备，用以解决如下技术问题：需要一种更准确的词向量生成方案。

[07]为解决上述技术问题，本说明书实施例是这样实现的：

[08]本说明书实施例提供一种词向量生成方法，包括：

获取对语料分词得到的各词；

根据所述各词对应的各 n 元字符，建立所述各词的特征向量，所述 n 元字符表征其对应的词的连续 n 个字符；

根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练；

5 根据所述各词的特征向量和训练后的所述循环神经网络，生成所述各词的词向量。

[09] 本说明书实施例提供一种词向量生成装置，包括：

获取模块，获取对语料分词得到的各词；

建立模块，根据所述各词对应的各 n 元字符，建立所述各词的特征向量，所述 n 元字符表征其对应的词的连续 n 个字符；

10 训练模块，根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练；

生成模块，根据所述各词的特征向量和训练后的所述循环神经网络，生成所述各词的词向量。

[10] 本说明书实施例提供的另一种词向量生成方法，包括：

15 步骤 1，建立通过对语料分词得到的各词构成的词汇表，所述各词不包括在所述语料中出现次数少于设定次数的词；跳转步骤 2；

步骤 2，确定各词对应的各 n 元字符的总数量，相同的 n 元字符只计一次，所述 n 元字符表征其对应的词的连续 n 个字符；跳转步骤 3；

20 步骤 3，根据所述各词对应的各 n 元字符，为各词分别建立维度为所述总数量的特征向量，所述特征向量的每维分别对应一个不同的 n 元字符，所述每维的取值表明其对应的 n 元字符是否对应于所述特征向量对应的词；跳转步骤 4；

步骤 4，遍历分词后的所述语料，对遍历到的当前词执行步骤 5，若遍历完成则执行步骤 6，否则继续遍历；

25 步骤 5，以当前词为中心，向两侧分别滑动至多 k 个词建立窗口，将窗口中除当前词以外的词作为上下文词，并将所有上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行循环计算，得到第一向量；将当前词以及在所述语料中选择的负样例词的特征向量输入所述循环神经网络的全连接层进行计算，分别得到第二向量和第三向量；根据所述第一向量、所述第二向量、所述第三向量，以及指定的损失函数，更新所

述循环神经网络的参数;

所述循环计算按照如下公式进行:

$$s_t = \sigma(Ux_t + Ws_{t-1})$$

$$o_t = \text{softmax}(Vs_t)$$

5 所述损失函数包括:

$$l(w, c; U, W, V, \varsigma, \tau) = \log \left(1 + \sum_{m=1}^{\lambda} \exp \left(-\gamma \cdot (s(w, c) - s(w'_m, c)) \right) \right)$$

10 其中, x_t 表示序列表示层在 t 时刻的输入单元, 即当前词的第 $t+1$ 个上下文词的特征向量, s_t 表示序列表示层在 t 时刻的隐藏单元, o_t 表示对当前词的前 $t+1$ 个上下文词的特征向量进行循环计算得到的向量, U 、 W 、 V 表示序列表示层的权重参数, σ 表示激励函数, c 表示所述第一向量, w 表示所述第二向量, w'_m 表示第 m 个负样例词对应的所述第三向量, ς 表示全连接层的权重参数, τ 表示全连接层的偏置参数, γ 表示超参数, s 表示相似度计算函数, λ 表示负样例词的数量;

步骤 6, 将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算, 得到对应的词向量。

15 [11] 本说明书实施例提供的一种词向量生成设备, 包括:

至少一个处理器; 以及,

与所述至少一个处理器通信连接的存储器; 其中,

所述存储器存储有可被所述至少一个处理器执行的指令, 所述指令被所述至少一个处理器执行, 以使所述至少一个处理器能够:

20 获取对语料分词得到的各词;

根据所述各词对应的各 n 元字符, 建立所述各词的特征向量, 所述 n 元字符表征其对应的词的连续 n 个字符;

根据所述各词的特征向量, 以及所述各词在所述语料中的上下文词的特征向量, 对循环神经网络进行训练;

25 根据所述各词的特征向量和训练后的所述循环神经网络, 生成所述各词的词向量。

[12]本说明书实施例采用的上述至少一个技术方案能够达到以下有益效果：循环神经网络可以通过循环计算，对词的上下文整体语义信息进行刻画，提取更多的上下文语义信息，而且 n 元字符能够更精细地对词进行表达，因此，有助于更准确地生成词向量。

5 附图说明

[13]为了更清楚地说明本说明书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说明书中记载的一些实施例，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。

10 [14]图 1 为本说明书的方案在一种实际应用场景下涉及的一种整体架构示意图；

[15]图 2 为本说明书实施例提供的一种词向量生成方法的流程示意图；

[16]图 3 为本说明书实施例提供的实际应用场景下，一个英文词的特征向量示意图；

[17]图 4 为本说明书实施例提供的实际应用场景下，基于循环神经网络的词向量生成方法的原理示意图；

15 [18]图 5 为本说明书实施例提供的循环神经网络的序列表示层的结构示意图；

[19]图 6 为本说明书实施例提供的另一种词向量生成方法的流程示意图；

[20]图 7 为本说明书实施例提供的对应于图 2 的一种词向量生成装置的结构示意图。

具体实施方式

20 [21]本说明书实施例提供词向量生成方法、装置以及设备。

[22]为了使本技术领域的人员更好地理解本说明书中的技术方案，下面将结合本说明书实施例中的附图，对本说明书实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本说明书实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

25 [23]图 1 为本说明书的方案在一种实际应用场景下涉及的一种整体架构示意图。该整体

架构中，主要涉及训练循环神经网络用以生成词向量的服务器。可以基于 n 元字符建立各词的特征向量，并利用特征向量和词的上下文关系训练循环神经网络。其中，特征向量可以由该服务器或者另一设备来建立。

[24]本说明书的方案适用于由字母构成的语言，比如英文、法文、德文、西班牙文等，也适用于由非字母元素构成但是能够便利地映射为字母的语言，比如中文（可以映射为拼音字母）、日文（可以映射为罗马字母）等。为了便于描述，以下各实施例主要针对英文的场景，对本说明书的方案进行说明。

[25]图 2 为本说明书实施例提供的一种词向量生成方法的流程示意图。从设备角度而言，该流程的执行主体比如包括以下至少一种设备：个人计算机、大中型计算机、计算机集群、手机、平板电脑、智能可穿戴设备、车机等。

[26]图 2 中的流程可以包括以下步骤：

[27]S202：获取对语料分词得到的各词。

[28]在本说明书实施例中，所述各词具体可以是：语料中至少出现过一次的词中的至少部分词。为了便于后续处理，可以将各词保存在词汇表中，需要使用时从词汇表中读取词即可。

[29]需要说明的是，考虑到若某词在语料中出现的次数太少，则后续处理时相应的迭代次数也少，训练结果可信度相对低，因此，可以将这种词筛除，使其不包含在所述各词中。在这种情况下，所述各词具体是：语料中至少出现过一次的词中的部分词。

[30]S204：根据所述各词对应的各 n 元字符，建立所述各词的特征向量，所述 n 元字符表征其对应的词的连续 n 个字符。

[31]在本说明书实施例中，词的字符可以包括构成该词的字符，或者构成该词的字符所映射的其他字符。以词 “boy” 为例，“b”、“o”、“y” 均为构成词 “boy” 的字符。

[32]为了表现词序，可以按照一定的规则在原始的词中添加一些标记字符，可以将这些标记字符也视为词的字符，比如，可以在原始的词的开始位置和/或结束位置等位置添加标注字符，在标注后，词 “boy” 比如记作 “#boy#”，可以将这两个 “#” 也视为词 “boy” 的字符。

[33]进一步地， n 为不小于 1 的整数。以 “#boy#” 为例，其 3 元字符包括 “#bo”（第 1~3 个字符）、“boy”（第 2~4 个字符）、“oy#”（第 3~5 个字符），其 4 元字符包

括“#boy”（第1~4个字符）、“boy#”（第2~5个字符）”。

[34]在本说明书实施例中， n 的取值可以是动态可调的。对于同一个词，在确定该词对应的 n 元字符时， n 的取值可以只取1个（比如，只确定该词对应的3元字符），也可以取多个（比如，确定该词对应的3元字符和4元字符等）。

5 [35]为了便于计算机处理， n 元字符可以基于指定的代码（比如，数字等）进行表示。比如，可以将不同的字符或者不同的 n 元字符分别用一个不同的代码或者代码串表示。

[36]在本说明书实施例中，可以使词的特征向量通过各维的取值，指示该词对应的 n 元字符。更精确地，还可以使词的特征向量指示该词对应的 n 元字符的顺序。

10 [37]S206：根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练。

[38]在本说明书实施例中，循环神经网络的序列表示层用于处理序列，以得到序列的全局信息，全局信息的内容也会受到序列中各元素顺序的影响。具体到本说明书的场景，序列由当前词（各词可以分别作为当前词）的各上下文词的特征向量作为元素构成，全局信息可以指当前词的全部上下文词的整体语义。

15 [39]S208：根据所述各词的特征向量和训练后的所述循环神经网络，生成所述各词的词向量。

[40]通过训练循环神经网络，能够为循环神经网络确定合理的参数，使得循环神经网络能够较为准确地刻画上下文词的整体语义，以及对应的当前词的语义。所述参数比如包括权重参数和偏置参数等。

20 [41]利用训练后的循环神经网络对特征向量进行推理，能够获得词向量。

[42]通过图2的方法，循环神经网络可以通过循环计算，对词的上下文整体语义信息进行刻画，提取更多的上下文语义信息，而且 n 元字符能够更精细地对词进行表达，因此，有助于更准确地生成词向量。

25 [43]基于图2的方法，本说明书实施例还提供了该方法的一些具体实施方案，以及扩展方案，下面进行说明。

[44]在本说明书实施例中，对于步骤S204，所述根据所述各词对应的各 n 元字符，建立所述各词的特征向量，具体可以包括：

[45]确定所述各词对应的各 n 元字符中不同 n 元字符的总数量；分别为所述各词建立根

据所述总数量确定维度的特征向量，所述特征向量通过各维的取值指示了其对应的词所对应的各 n 元字符。

[46] 例如，对各词对应的全部 n 元字符逐一进行编号，编号从 0 开始，依次加 1，相应的 n 元字符编号相同，假定总数量为 N_c ，则最后一个 n 元字符的编号为 $N_c - 1$ 。分别为
5 各词建立一个维度为 N_c 特征向量，具体地，假定 $n=3$ ，某词对应的全部 3 元字符编号分别为 2、34、127，则为其建立的特征向量中第 2、34、127 个元素可以为 1，其余元素为 0。

[47] 更直观地，基于上例，本说明书实施例提供实际应用场景下，一个英文词的特征向量示意图，如图 3 所示。该英文词为 “boy”，起始位置和结束位置分别添加有一个标注字符 “#”， f 表示根据词建立特征向量的处理过程，特征向量比如是一个列向量，
10 根据 “boy” 的各 3 元字符建立，可以看到，特征向量中有 3 个元素的取值为 1，分别指示了 3 元字符 “#bo”、“boy”、“oy#”，其他元素的取值为 0，表示 “boy” 未对应于其他 3 元字符。

[48] 在本说明书实施例中，在对循环神经网络进行训练时，目标是使得当前词与上下文
15 词的特征向量在经过训练后的循环神经网络推理后，相似度能够相对变高。

[49] 进一步地，将上下文词视为正样例词，作为对照，还可以按照一定的规则选择当前词的一个或者多个负样例词也参与训练，如此有利于训练快速收敛以及获得更为准确的训练结果。这种情况下，所述目标还可以包括使得当前词与负样例词的特征向量在经过
20 训练后的循环神经网络推理后，相似度能够相对变低。负样例词比如可以在语料中随机选择得到，或者在非上下文词中选择得到，等等。本说明书对计算相似度的具体方式并不做限定，比如，可以基于向量的夹角余弦运算计算相似度，可以基于向量的平方和运算计算相似度，等等。

[50] 根据上一段的分析，对于步骤 S206，所述根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练。具体可以包括：

[51] 根据所述各词的特征向量，以及所述各词在所述语料中的上下文词和负样例词的特征向量，对循环神经网络进行训练。
25

[52] 在本说明书实施例中，循环神经网络的训练过程可以是迭代进行的，比较简单的一种方式是对分词后的语料进行遍历，每遍历到上述各词中的一个词即进行一次迭代，直到遍历完毕，可以视为已经利用该语料训练过循环神经网络了。

[53]具体地，所述根据所述各词的特征向量，以及所述各词在所述语料中的上下文词和负样例词的特征向量，对循环神经网络进行训练，可以包括：

[54]对分词后的所述语料进行遍历，对遍历到的当前词执行（执行内容即为一次迭代过程）：

5 [55]确定当前词在分词后的所述语料中的一个或多个上下文词以及负样例词；将当前词的上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行循环计算，得到第一向量；将当前词的特征向量输入所述循环神经网络的全连接层进行计算，得到第二向量，以及将当前词的负样例词的特征向量输入所述循环神经网络的全连接层进行计算，得到第三向量；根据所述第一向量、所述第二向量、所述第三向量，以及指定的损失函数，更新所述循环神经网络的参数。

10

[56]更直观地，结合图 4 进行说明。图 4 为本说明书实施例提供的实际应用场景下，基于循环神经网络的词向量生成方法的原理示意图。

[57]图 4 的循环神经网络主要包括序列表示层、全连接层，以及 Softmax 层。在训练循环神经网络的过程中，上下文词的特征向量由序列表示层进行处理，以提取上下文词整体的词义信息，而当前词及其负样例词的特征向量则可以由全连接层进行处理。下面分别详细说明。

15

[58]在本说明书实施例中，假定采用滑动窗口来确定上下文词，滑动窗口的中心为遍历到的当前词，滑动窗口中除当前词以外的其他词为上下文词。将全部上下文词的特征向量依次（每个上下文词的特征向量分别是上述序列中的一个元素）输入序列表示层，进而可以按照如下公式，进行循环计算：

20

$$s_t = \sigma(Ux_t + Ws_{t-1})$$

$$o_t = \text{softmax}(Vs_t)$$

其中， x_t 表示序列表示层在 t 时刻的输入单元，即当前词的第 $t+1$ 个上下文词的特征向量， s_t 表示序列表示层在 t 时刻的隐藏单元， o_t 表示对当前词的前 $t+1$ 个上下文词的特征向量进行循环计算得到的向量， U 、 W 、 V 表示序列表示层的权重参数， σ 表示激励函数，比如， $\tanh$ 函数或者 ReLU 函数等。公式中参数的下标可以从 0 开始。

25

[59]为了便于理解循环计算所使用的公式，本说明书实施例还提供了循环神经网络的序列表示层的结构示意图，如图 5 所示。

[60]在图 5 左侧, x 表示序列表示层的输入单元, s 表示序列输入层的隐藏单元, o 表示序列输入层的输出单元, 隐藏单元会利用激活函数处理输入的数据, 输出单元会用 $softmax$ 函数处理输入的数据。隐藏单元前一时刻计算输出的数据会以一定的权重向下一时刻隐藏单元的输入进行反馈, 从而, 使得依次输入的这个序列的内容都会反映在序列表示层最终输出的数据中, U 表示从输入单元至隐藏单元的权重参数, W 表示从隐藏单元反馈至下一时刻的隐藏单元的权重参数, V 表示从隐藏单元至输出单元的权重参数。

[61]在图 5 右侧, 对左侧的结构进行了展开, 示出了连续三个时刻序列输入层的结构, 以及对序列依次输入的三个元素进行处理的原理。可以看到, 在 $t-1$ 时刻, 隐藏单元计算输出的数据表示为 s_{t-1} ; 在 t 时刻, 输入的是当前词的第 t 个上下文词的特征向量 x_t , 隐藏单元计算输出的数据表示为 $s_t = \sigma(Ux_t + Ws_{t-1})$, 输出单元计算输出的数据表示为 $o_t = softmax(Vs_t)$; 在 $t+1$ 时刻, 输入的是当前词的第 $t+1$ 个上下文词的特征向量 x_{t+1} , 隐藏单元计算输出的数据表示为 $s_{t+1} = \sigma(Ux_{t+1} + Ws_t)$, 输出单元计算输出的数据表示为 $o_{t+1} = softmax(Vs_{t+1})$ 。以此类推, 当前词的最后一个上下文词的特征向量输入后, 能够输出对应的第一向量。

[62]图 4 还示例性地示出了某语料中的某个当前词 “liquid”、该当前词在该语料中的 6 个上下文词 “as”、“the”、“vegan”、“gelatin”、“substitute”、“absorbs”, 以及该当前词在该语料中的两个负样例词 “year”、“make”。图 4 中示出了序列表示层对应于当前词 “liquid” 的最后一个上下文词 “absorbs” 的输出单元, 该输出单元输出当前词 “liquid” 对应的第一向量。

[63]对于当前词, 其特征向量可以输入全连接层, 比如按照以下公式进行计算:

$$w = \sigma(\zeta \cdot q + \tau)$$

其中, w 表示全连接层对当前词的特征向量处理后输出的所述第二向量, ζ 表示全连接层的权重参数, q 表示当前词的特征向量, τ 表示全连接层的偏置参数。

[64]类似地, 对于每个负样例词, 其特征向量可以分别输入全连接层, 参照当前词的方式进行处理, 得到所述第三向量, 将第 m 个负样例词对应的所述第三向量表示为 w'_m 。

[65]进一步地, 所述根据所述第一向量、所述第二向量、所述第三向量, 以及指定的损失函数, 更新所述循环神经网络的参数, 比如可以包括: 计算所述第二向量与所述第一

向量的第一相似度,以及所述第三向量与所述第一向量的第二相似度;根据所述第一相似度、所述第二相似度,以及指定的损失函数,更新所述循环神经网络的参数。

[66]列举一种损失函数作为示例。所述损失函数比如可以是:

$$l(w, c; U, W, V, \varsigma, \tau) = \log \left(1 + \sum_{m=1}^{\lambda} \exp \left(-\gamma \cdot (s(w, c) - s(w'_m, c)) \right) \right)$$

5 其中, c 表示所述第一向量, w 表示所述第二向量, w'_m 表示第 m 个负样例词对应的所述第三向量, U 、 W 、 V 表示序列表示层的权重参数, ς 表示全连接层的权重参数, τ 表示全连接层的偏置参数, γ 表示超参数, s 表示相似度计算函数, λ 表示负样例词的数量。

10 [67]在实际应用中,若未采用负样例词,则采用的损失函数中可以相应地去掉计算第一向量与第三向量的相似度的项。

[68]在本说明书实施例中,在循环神经网络训练后,可以通过对特征向量进行推理,以生成词向量。具体地,对于步骤 S208,所述根据所述各词的特征向量和训练后的所述循环神经网络,生成所述各词的词向量,具体可以包括:

15 [69]将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算,获得计算后输出的向量,作为对应的词向量。

[70]基于同样的思路,本说明书实施例提供了另一种词向量生成方法,其为图 2 中的词向量生成方法示例性的一种具体实施方案。图 6 为该另一种词向量生成方法的流程示意图。

[71]图 6 中的流程可以包括以下步骤:

20 步骤 1,建立通过对语料分词得到的各词构成的词汇表,所述各词不包括在所述语料中出现次数少于设定次数的词;跳转步骤 2;

步骤 2,确定各词对应的各 n 元字符的总数量,相同的 n 元字符只计一次,所述 n 元字符表征其对应的词的连续 n 个字符;跳转步骤 3;

25 步骤 3,根据所述各词对应的各 n 元字符,为各词分别建立维度为所述总数量的特征向量,所述特征向量的每维分别对应一个不同的 n 元字符,所述每维的取值表明其对应的 n 元字符是否对应于所述特征向量对应的词;跳转步骤 4;

步骤 4,遍历分词后的所述语料,对遍历到的当前词执行步骤 5,若遍历完成则执行

步骤 6, 否则继续遍历;

步骤 5, 以当前词为中心, 向两侧分别滑动至多 k 个词建立窗口, 将窗口中除当前词以外的词作为上下文词, 并将所有上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行循环计算, 得到第一向量; 将当前词以及在所述语料中选择的负样例词的特征向量输入所述循环神经网络的全连接层进行计算, 分别得到第二向量和第三向量; 根据所述第一向量、所述第二向量、所述第三向量, 以及指定的损失函数, 更新所述循环神经网络的参数;

所述循环计算按照如下公式进行:

$$s_t = \sigma(Ux_t + Ws_{t-1})$$

$$o_t = \text{softmax}(Vs_t)$$

所述损失函数包括:

$$l(w, c; U, W, V, \varsigma, \tau) = \log \left(1 + \sum_{m=1}^{\lambda} \exp \left(-\gamma \cdot (s(w, c) - s(w'_m, c)) \right) \right)$$

其中, x_t 表示序列表示层在 t 时刻的输入单元, 即当前词的第 $t+1$ 个上下文词的特征向量, s_t 表示序列表示层在 t 时刻的隐藏单元, o_t 表示对当前词的前 $t+1$ 个上下文词的特征向量进行循环计算得到的向量, U 、 W 、 V 表示序列表示层的权重参数, σ 表示激励函数, c 表示所述第一向量, w 表示所述第二向量, w'_m 表示第 m 个负样例词对应的所述第三向量, ς 表示全连接层的权重参数, τ 表示全连接层的偏置参数, γ 表示超参数, s 表示相似度计算函数, λ 表示负样例词的数量;

步骤 6, 将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算, 得到对应的词向量。

[72] 该另一种词向量生成方法中各步骤可以由相同或者不同的模块执行, 本说明书对此并不做具体限定。

[73] 上面对本说明书实施例提供的词向量生成方法进行了说明, 基于同样的思路, 本说明书实施例还提供了对应的装置, 如图 7 所示。

[74] 图 7 为本说明书实施例提供的对应于图 2 的一种词向量生成装置的结构示意图, 该装置可以位于图 2 中流程的执行主体, 包括:

获取模块 701, 获取对语料分词得到的各词;

建立模块 702, 根据所述各词对应的各 n 元字符, 建立所述各词的特征向量, 所述 n 元字符表征其对应的词的连续 n 个字符;

5 训练模块 703, 根据所述各词的特征向量, 以及所述各词在所述语料中的上下文词的特征向量, 对循环神经网络进行训练;

生成模块 704, 根据所述各词的特征向量和训练后的所述循环神经网络, 生成所述各词的词向量。

[75]可选地, 所述词的字符包括构成所述词的各字符, 以及添加于所述词的起始位置和/或结束位置的标注字符。

10 [76]可选地, 所述建立模块 702 根据所述各词对应的各 n 元字符, 建立所述各词的特征向量, 具体包括:

所述建立模块 702 确定所述各词对应的各 n 元字符中不同 n 元字符的总数量;

分别为所述各词建立根据所述总数量确定维度的特征向量, 所述特征向量通过各维的取值指示了其对应的词所对应的各 n 元字符。

15 [77]可选地, 所述训练模块 703 根据所述各词的特征向量, 以及所述各词在所述语料中的上下文词的特征向量, 对循环神经网络进行训练, 具体包括:

[78]所述训练模块 703 根据所述各词的特征向量, 以及所述各词在所述语料中的上下文词和负样例词的特征向量, 对循环神经网络进行训练。

20 [79]可选地, 所述训练模块 703 根据所述各词的特征向量, 以及所述各词在所述语料中的上下文词和负样例词的特征向量, 对循环神经网络进行训练, 具体包括:

[80]所述训练模块 703 对分词后的所述语料进行遍历, 对遍历到的当前词执行:

确定当前词在分词后的所述语料中的一个或多个上下文词以及负样例词;

将当前词的上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行循环计算, 得到第一向量;

25 将当前词的特征向量输入所述循环神经网络的全连接层进行计算, 得到第二向量, 以及将当前词的负样例词的特征向量输入所述循环神经网络的全连接层进行计算, 得到第三向量;

根据所述第一向量、所述第二向量、所述第三向量，以及指定的损失函数，更新所述循环神经网络的参数。

[81]可选地，所述训练模块 703 进行循环计算，具体包括：

[82]所述训练模块 703 按照如下公式，进行循环计算：

$$s_t = \sigma(Ux_t + Ws_{t-1})$$

$$o_t = \text{softmax}(Vs_t)$$

其中， x_t 表示序列表示层在 t 时刻的输入单元，即当前词的第 $t+1$ 个上下文词的特征向量， s_t 表示序列表示层在 t 时刻的隐藏单元， o_t 表示对当前词的前 $t+1$ 个上下文词的特征向量进行循环计算得到的向量， U 、 W 、 V 表示序列表示层的权重参数， σ 表示激励函数。

[83]可选地，所述训练模块 703 根据所述第一向量、所述第二向量、所述第三向量，以及指定的损失函数，更新所述循环神经网络的参数，具体包括：

所述训练模块 703 计算所述第二向量与所述第一向量的第一相似度，以及所述第三向量与所述第一向量的第二相似度；

根据所述第一相似度、所述第二相似度，以及指定的损失函数，更新所述循环神经网络的参数。

[84]可选地，所述损失函数具体包括：

$$l(w, c; U, W, V, \varsigma, \tau) = \log \left(1 + \sum_{m=1}^{\lambda} \exp \left(-\gamma \cdot (s(w, c) - s(w'_m, c)) \right) \right)$$

其中， c 表示所述第一向量， w 表示所述第二向量， w'_m 表示第 m 个负样例词对应的所述第三向量， U 、 W 、 V 表示序列表示层的权重参数， ς 表示全连接层的权重参数， τ 表示全连接层的偏置参数， γ 表示超参数， s 表示相似度计算函数， λ 表示负样例词的数量。

[85]可选地，所述生成模块 704 根据所述各词的特征向量和训练后的所述循环神经网络，生成所述各词的词向量，具体包括：

[86]所述生成模块 704 将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算，获得计算后输出的向量，作为对应的词向量。

[87]基于同样的思路,本说明书实施例还提供了对应的一种词向量生成设备,包括:

至少一个处理器;以及,

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个
5 处理器执行,以使所述至少一个处理器能够:

获取对语料分词得到的各词;

根据所述各词对应的各 n 元字符,建立所述各词的特征向量,所述 n 元字符表征其
对应的词的连续 n 个字符;

根据所述各词的特征向量,以及所述各词在所述语料中的上下文词的特征向量,对
10 循环神经网络进行训练;

根据所述各词的特征向量和训练后的所述循环神经网络,生成所述各词的词向量。

[88]基于同样的思路,本说明书实施例还提供了对应的一种非易失性计算机存储介质,
存储有计算机可执行指令,所述计算机可执行指令设置为:

获取对语料分词得到的各词;

15 根据所述各词对应的各 n 元字符,建立所述各词的特征向量,所述 n 元字符表征其
对应的词的连续 n 个字符;

根据所述各词的特征向量,以及所述各词在所述语料中的上下文词的特征向量,对
循环神经网络进行训练;

根据所述各词的特征向量和训练后的所述循环神经网络,生成所述各词的词向量。

20 [89]上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。
在一些情况下,在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执
行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要示出的特定顺
序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是
可以的或者可能是有利的。

25 [90]本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分
互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于装置、
设备、非易失性计算机存储介质实施例而言,由于其基本相似于方法实施例,所以描述

的比较简单，相关之处参见方法实施例的部分说明即可。

[91]本说明书实施例提供的装置、设备、非易失性计算机存储介质与方法是对应的，因此，装置、设备、非易失性计算机存储介质也具有与对应方法类似的有益技术效果，由于上面已经对方法的有益技术效果进行了详细说明，因此，这里不再赘述对应装置、设备、非易失性计算机存储介质的有益技术效果。

[92]在 20 世纪 90 年代，对于一个技术的改进可以很明显地区分是硬件上的改进（例如，对二极管、晶体管、开关等电路结构的改进）还是软件上的改进（对于方法流程的改进）。然而，随着技术的发展，当今的很多方法流程的改进已经可以视为硬件电路结构的直接改进。设计人员几乎都通过将改进的方法流程编程到硬件电路中来得到相应的硬件电路结构。因此，不能说一个方法流程的改进就不能用硬件实体模块来实现。例如，可编程逻辑器件（Programmable Logic Device, PLD）（例如现场可编程门阵列（Field Programmable Gate Array, FPGA））就是这样一种集成电路，其逻辑功能由用户对器件编程来确定。由设计人员自行编程来把一个数字系统“集成”在一片 PLD 上，而不需要请芯片制造厂商来设计和制作专用的集成电路芯片。而且，如今，取代手工地制作集成电路芯片，这种编程也多半改用“逻辑编译器（logic compiler）”软件来实现，它与程序开发撰写时所用的软件编译器相类似，而要编译之前的原始代码也得用特定的编程语言来撰写，此称之为硬件描述语言（Hardware Description Language, HDL），而 HDL 也并非仅有一种，而是有许多种，如 ABEL（Advanced Boolean Expression Language）、AHDL（Altera Hardware Description Language）、Confluence、CUPL（Cornell University Programming Language）、HDCal、JHDL（Java Hardware Description Language）、Lava、Lola、MyHDL、PALASM、RHDL（Ruby Hardware Description Language）等，目前最普遍使用的是 VHDL（Very-High-Speed Integrated Circuit Hardware Description Language）与 Verilog。本领域技术人员也应该清楚，只需要将方法流程用上述几种硬件描述语言稍作逻辑编程并编程到集成电路中，就可以很容易得到实现该逻辑方法流程的硬件电路。

[93]控制器可以按任何适当的方式实现，例如，控制器可以采取例如微处理器或处理器以及存储可由该（微）处理器执行的计算机可读程序代码（例如软件或固件）的计算机可读介质、逻辑门、开关、专用集成电路（Application Specific Integrated Circuit, ASIC）、可编程逻辑控制器和嵌入微控制器的形式，控制器的例子包括但不限于以下微控制器：ARC 625D、Atmel AT91SAM、Microchip PIC18F26K20 以及 Silicone Labs C8051F320，

存储器控制器还可以被实现为存储器的控制逻辑的一部分。本领域技术人员也知道，除了以纯计算机可读程序代码方式实现控制器以外，完全可以通过将方法步骤进行逻辑编程来使得控制器以逻辑门、开关、专用集成电路、可编程逻辑控制器和嵌入微控制器等的形式来实现相同功能。因此这种控制器可以被认为是一种硬件部件，而对其内包括的用于实现各种功能的装置也可以视为硬件部件内的结构。或者甚至，可以将用于实现各种功能的装置视为既可以是实现方法的软件模块又可以是硬件部件内的结构。

[94]上述实施例阐明的系统、装置、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为计算机。具体的，计算机例如可以为个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任何设备的组合。

[95]为了描述的方便，描述以上装置时以功能分为各种单元分别描述。当然，在实施本说明书时可以把各单元的功能在同一个或多个软件和/或硬件中实现。

[96]本领域内的技术人员应明白，本说明书实施例可提供为方法、系统、或计算机程序产品。因此，本说明书实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本说明书实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

[97]本说明书是参照根据本说明书实施例的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[98]这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[99] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上, 使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理, 从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

5 [100] 在一个典型的配置中, 计算设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[101] 内存可能包括计算机可读介质中的非永久性存储器, 随机存取存储器(RAM)和/或非易失性内存等形式, 如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。

10 [102] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括, 但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带, 磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质, 可用于存储可以被计算设备访问的信息。按照本文中的界定, 计算机可读介质不包括暂存电脑可读媒体(transitory media), 如调制的数据信号和载波。

15 [103] 还需要说明的是, 术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含, 从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素, 而且还包括没有明确列出的其他要素, 或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下, 由语句“包括一个……”限定的要素, 并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

20 [104] 本领域技术人员应明白, 本说明书实施例可提供为方法、系统或计算机程序产品。因此, 本说明书可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且, 本说明书可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

25 [105] 本说明书可以在由计算机执行的计算机可执行指令的一般上下文中描述, 例如

程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本说明书，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

5 [106] 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

10 [107] 以上所述仅为本说明书实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理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权利要求范围之内。

权利要求书

1、一种词向量生成方法，包括：

获取对语料分词得到的各词；

5 根据所述各词对应的各 n 元字符，建立所述各词的特征向量，所述 n 元字符表征其对应的词的连续 n 个字符；

根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练；

根据所述各词的特征向量和训练后的所述循环神经网络，生成所述各词的词向量。

10 2、如权利要求 1 所述的方法，所述词的字符包括构成所述词的各字符，以及添加于所述词的起始位置和/或结束位置的标注字符。

3、如权利要求 1 所述的方法，所述根据所述各词对应的各 n 元字符，建立所述各词的特征向量，具体包括：

15 确定所述各词对应的各 n 元字符中不同 n 元字符的总数量；

分别为所述各词建立根据所述总数量确定维度的特征向量，所述特征向量通过各维的取值指示了其对应的词所对应的各 n 元字符。

20 4、如权利要求 1 所述的方法，所述根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练，具体包括：

根据所述各词的特征向量，以及所述各词在所述语料中的上下文词和负样例词的特征向量，对循环神经网络进行训练。

25 5、如权利要求 4 所述的方法，所述根据所述各词的特征向量，以及所述各词在所述语料中的上下文词和负样例词的特征向量，对循环神经网络进行训练，具体包括：

对分词后的所述语料进行遍历，对遍历到的当前词执行：

确定当前词在分词后的所述语料中的一个或多个上下文词以及负样例词；

将当前词的上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行循环计算，得到第一向量；

30 将当前词的特征向量输入所述循环神经网络的全连接层进行计算，得到第二向量，以及将当前词的负样例词的特征向量输入所述循环神经网络的全连接层进行计算，得到

第三向量;

根据所述第一向量、所述第二向量、所述第三向量,以及指定的损失函数,更新所述循环神经网络的参数。

- 5 6、如权利要求 5 所述的方法,所述进行循环计算,具体包括:
按照如下公式,进行循环计算:

$$s_t = \sigma(Ux_t + Ws_{t-1})$$

$$o_t = \text{softmax}(Vs_t)$$

10 其中, x_t 表示序列表示层在 t 时刻的输入单元,即当前词的第 $t+1$ 个上下文词的特征向量, s_t 表示序列表示层在 t 时刻的隐藏单元, o_t 表示对当前词的前 $t+1$ 个上下文词的特征向量进行循环计算得到的向量, U 、 W 、 V 表示序列表示层的权重参数, σ 表示激励函数。

- 15 7、如权利要求 5 所述的方法,所述根据所述第一向量、所述第二向量、所述第三向量,以及指定的损失函数,更新所述循环神经网络的参数,具体包括:

计算所述第二向量与所述第一向量的第一相似度,以及所述第三向量与所述第一向量的第二相似度;

根据所述第一相似度、所述第二相似度,以及指定的损失函数,更新所述循环神经网络的参数。

- 20 8、如权利要求 5 所述的方法,所述损失函数具体包括:

$$l(w, c; U, W, V, \varsigma, \tau) = \log \left(1 + \sum_{m=1}^{\lambda} \exp \left(-\gamma \cdot \left(s(w, c) - s(w'_m, c) \right) \right) \right)$$

25 其中, c 表示所述第一向量, w 表示所述第二向量, w'_m 表示第 m 个负样例词对应的所述第三向量, U 、 W 、 V 表示序列表示层的权重参数, ς 表示全连接层的权重参数, τ 表示全连接层的偏置参数, γ 表示超参数, s 表示相似度计算函数, λ 表示负样例词的数量。

- 9、如权利要求 1 所述的方法,所述根据所述各词的特征向量和训练后的所述循环神经网络,生成所述各词的词向量,具体包括:

30 将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算,

获得计算后输出的向量，作为对应的词向量。

10、一种词向量生成装置，包括：

获取模块，获取对语料分词得到的各词；

5 建立模块，根据所述各词对应的各 n 元字符，建立所述各词的特征向量，所述 n 元字符表征其对应的词的连续 n 个字符；

训练模块，根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练；

10 生成模块，根据所述各词的特征向量和训练后的所述循环神经网络，生成所述各词
的词向量。

11、如权利要求 10 所述的装置，所述词的字符包括构成所述词的各字符，以及添加于所述词的起始位置和/或结束位置的标注字符。

15 12、如权利要求 10 所述的装置，所述建立模块根据所述各词对应的各 n 元字符，建立所述各词的特征向量，具体包括：

所述建立模块确定所述各词对应的各 n 元字符中不同 n 元字符的总数量；

分别为所述各词建立根据所述总数量确定维度的特征向量，所述特征向量通过各维
20 的取值指示了其对应的词所对应的各 n 元字符。

13、如权利要求 10 所述的装置，所述训练模块根据所述各词的特征向量，以及所述各词在所述语料中的上下文词的特征向量，对循环神经网络进行训练，具体包括：

所述训练模块根据所述各词的特征向量，以及所述各词在所述语料中的上下文词和
25 负样例词的特征向量，对循环神经网络进行训练。

14、如权利要求 13 所述的装置，所述训练模块根据所述各词的特征向量，以及所述各词在所述语料中的上下文词和负样例词的特征向量，对循环神经网络进行训练，具体包括：

所述训练模块对分词后的所述语料进行遍历，对遍历到的当前词执行：

30 确定当前词在分词后的所述语料中的一个或多个上下文词以及负样例词；

将当前词的上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行
循环计算，得到第一向量；

将当前词的特征向量输入所述循环神经网络的全连接层进行计算，得到第二向量，
以及将当前词的负样例词的特征向量输入所述循环神经网络的全连接层进行计算，得到
第三向量；

5 根据所述第一向量、所述第二向量、所述第三向量，以及指定的损失函数，更新所
述循环神经网络的参数。

15、如权利要求 14 所述的装置，所述训练模块进行循环计算，具体包括：

所述训练模块按照如下公式，进行循环计算：

$$s_t = \sigma(Ux_t + Ws_{t-1})$$

$$o_t = \text{softmax}(Vs_t)$$

其中， x_t 表示序列表示层在 t 时刻的输入单元，即当前词的第 $t+1$ 个上下文词的特征
向量， s_t 表示序列表示层在 t 时刻的隐藏单元， o_t 表示对当前词的前 $t+1$ 个上下文词的特
征向量进行循环计算得到的向量， U 、 W 、 V 表示序列表示层的权重参数， σ 表示激励
函数。

16、如权利要求 14 所述的装置，所述训练模块根据所述第一向量、所述第二向量、
所述第三向量，以及指定的损失函数，更新所述循环神经网络的参数，具体包括：

所述训练模块计算所述第二向量与所述第一向量的第一相似度，以及所述第三向量
与所述第一向量的第二相似度；

20 根据所述第一相似度、所述第二相似度，以及指定的损失函数，更新所述循环神经
网络的参数。

17、如权利要求 14 所述的装置，所述损失函数具体包括：

$$l(w, c; U, W, V, \varsigma, \tau) = \log \left(1 + \sum_{m=1}^{\lambda} \exp \left(-\gamma \cdot \left(s(w, c) - s(w'_m, c) \right) \right) \right)$$

25 其中， c 表示所述第一向量， w 表示所述第二向量， w'_m 表示第 m 个负样例词对应的
所述第三向量， U 、 W 、 V 表示序列表示层的权重参数， ς 表示全连接层的权重参数，
 τ 表示全连接层的偏置参数， γ 表示超参数， s 表示相似度计算函数， λ 表示负样例词
的数量。

30 18、如权利要求 10 所述的装置，所述生成模块根据所述各词的特征向量和训练后

的所述循环神经网络，生成所述各词的词向量，具体包括：

所述生成模块将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算，获得计算后输出的向量，作为对应的词向量。

5 19、一种词向量生成方法，包括：

步骤 1，建立通过对语料分词得到的各词构成的词汇表，所述各词不包括在所述语料中出现次数少于设定次数的词；跳转步骤 2；

步骤 2，确定各词对应的各 n 元字符的总数量，相同的 n 元字符只计一次，所述 n 元字符表征其对应的词的连续 n 个字符；跳转步骤 3；

10 步骤 3，根据所述各词对应的各 n 元字符，为各词分别建立维度为所述总数量的特征向量，所述特征向量的每维分别对应一个不同的 n 元字符，所述每维的取值表明其对应的 n 元字符是否对应于所述特征向量对应的词；跳转步骤 4；

步骤 4，遍历分词后的所述语料，对遍历到的当前词执行步骤 5，若遍历完成则执行步骤 6，否则继续遍历；

15 步骤 5，以当前词为中心，向两侧分别滑动至多 k 个词建立窗口，将窗口中除当前词以外的词作为上下文词，并将所有上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行循环计算，得到第一向量；将当前词以及在所述语料中选择的负样例词的特征向量输入所述循环神经网络的全连接层进行计算，分别得到第二向量和第三向量；根据所述第一向量、所述第二向量、所述第三向量，以及指定的损失函数，更新所述循环神经网络的参数；

20 所述循环计算按照如下公式进行：

$$s_t = \sigma(Ux_t + Ws_{t-1})$$

$$o_t = \text{softmax}(Vs_t)$$

所述损失函数包括：

25

$$l(w, c; U, W, V, \zeta, \tau) = \log \left(1 + \sum_{m=1}^{\lambda} \exp \left(-\gamma \cdot \left(s(w, c) - s(w'_m, c) \right) \right) \right)$$

其中， x_t 表示序列表示层在 t 时刻的输入单元，即当前词的第 $t+1$ 个上下文词的特征向量， s_t 表示序列表示层在 t 时刻的隐藏单元， o_t 表示对当前词的前 $t+1$ 个上下文词的特征向量进行循环计算得到的向量， U 、 W 、 V 表示序列表示层的权重参数， σ 表示激励函数， c 表示所述第一向量， w 表示所述第二向量， w'_m 表示第 m 个负样例词对应的所述

第三向量, ς 表示全连接层的权重参数, τ 表示全连接层的偏置参数, γ 表示超参数, s 表示相似度计算函数, λ 表示负样例词的数量;

步骤 6, 将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算, 得到对应的词向量。

5

20、一种词向量生成设备, 包括:

至少一个处理器; 以及,

与所述至少一个处理器通信连接的存储器; 其中,

所述存储器存储有可被所述至少一个处理器执行的指令, 所述指令被所述至少一个处理器执行, 以使所述至少一个处理器能够:

10

获取对语料分词得到的各词;

根据所述各词对应的各 n 元字符, 建立所述各词的特征向量, 所述 n 元字符表征其对应的词的连续 n 个字符;

根据所述各词的特征向量, 以及所述各词在所述语料中的上下文词的特征向量, 对循环神经网络进行训练;

15

根据所述各词的特征向量和训练后的所述循环神经网络, 生成所述各词的词向量。

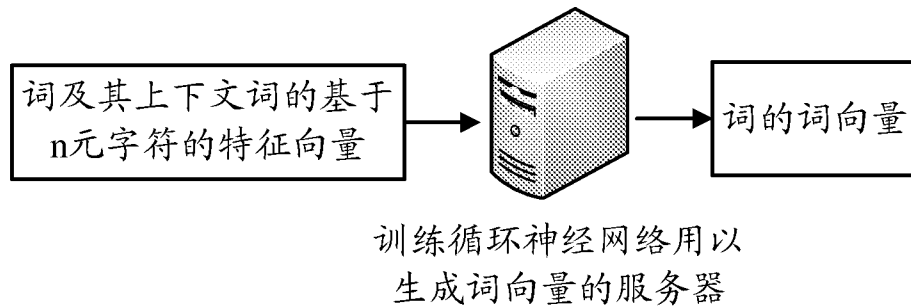


图 1

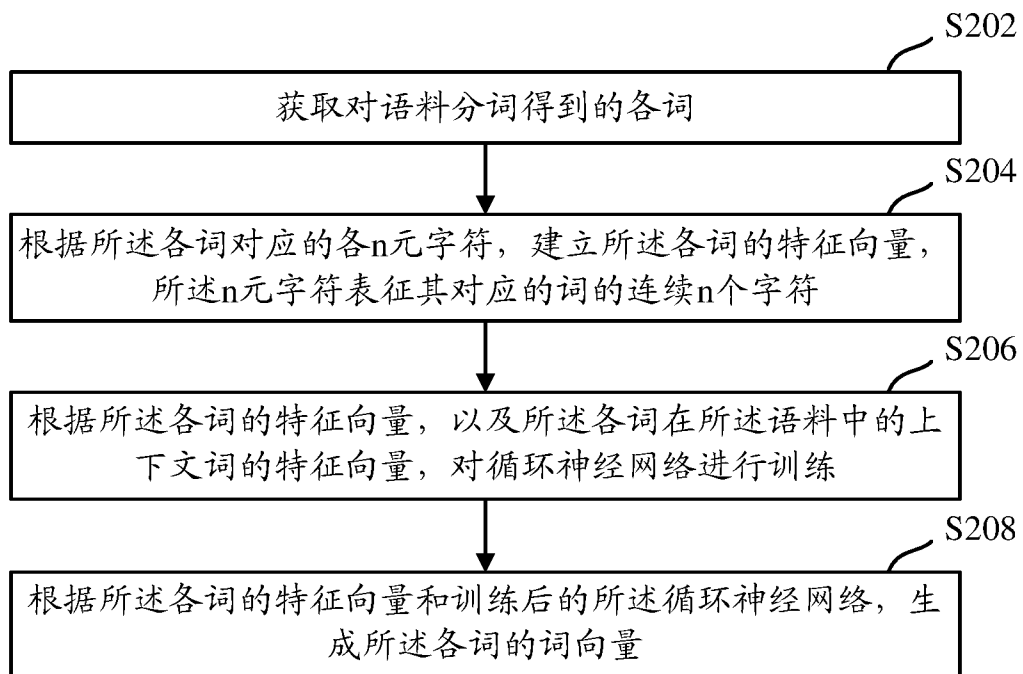


图 2

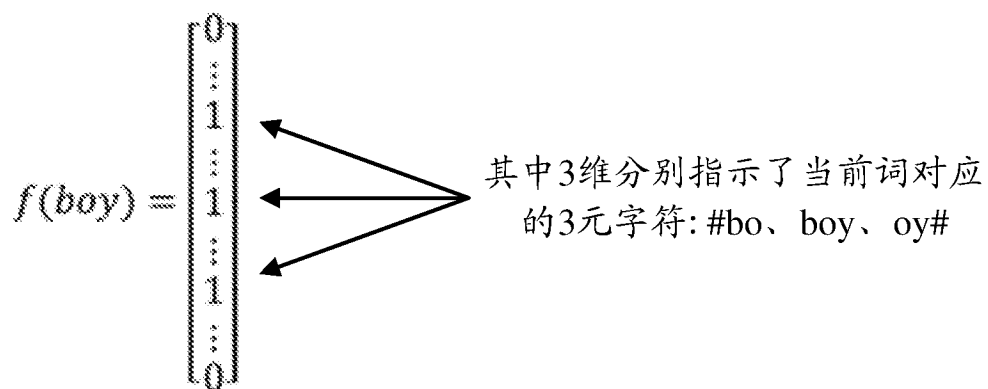


图 3

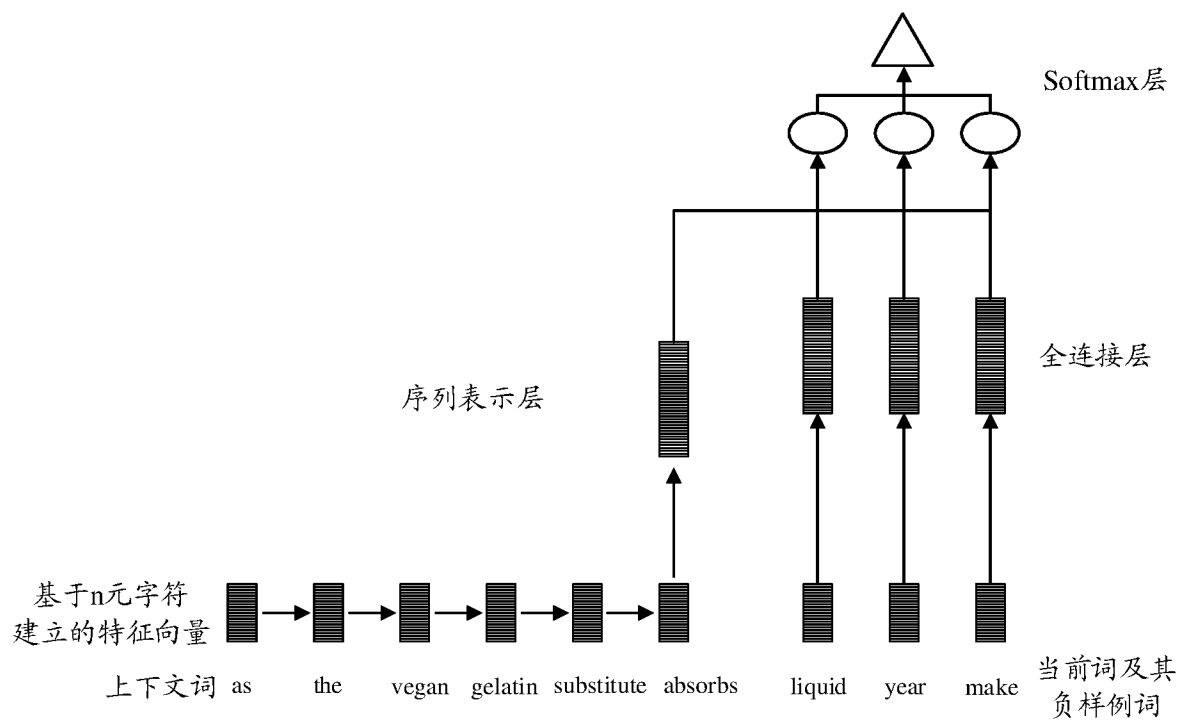


图 4

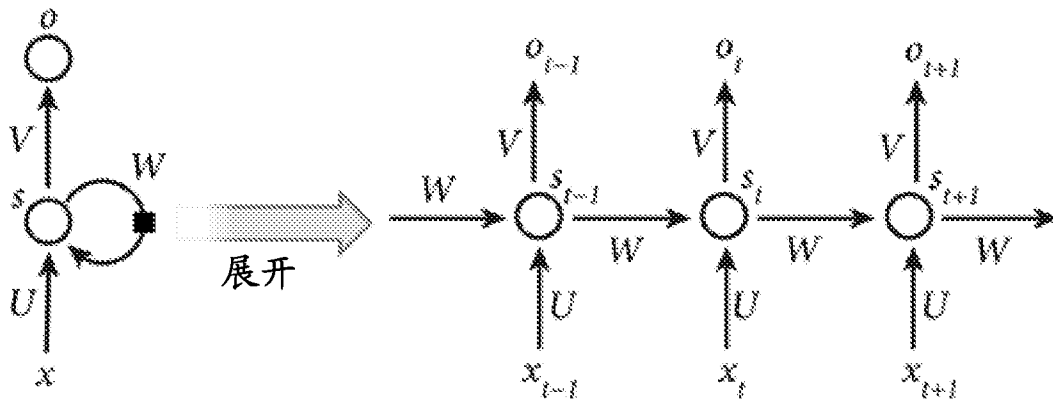


图 5

步骤1，建立通过对语料分词得到的各词构成的词汇表，所述各词不包括在所述语料中出现次数少于设定次数的词；跳转步骤2

步骤2，确定各词对应的各n元字符的总数量，相同的n元字符只计一次，所述n元字符表征其对应的词的连续n个字符；跳转步骤3

步骤3，根据所述各词对应的各n元字符，为各词分别建立维度为所述总数量的特征向量，所述特征向量的每维分别对应一个不同的n元字符，所述每维的取值表明其对应的n元字符是否对应于所述特征向量对应的词；跳转步骤4

步骤4，遍历分词后的所述语料，对遍历到的当前词执行步骤5，若遍历完成则执行步骤6，否则继续遍历

步骤5，以当前词为中心，向两侧分别滑动至多个词建立窗口，将窗口中除当前词以外的词作为上下文词，并将所有上下文词的特征向量构成的序列输入循环神经网络的序列表示层进行循环计算，得到第一向量；将当前词以及在所述语料中选择的负样例词的特征向量输入所述循环神经网络的全连接层进行计算，分别得到第二向量和第三向量；根据所述第一向量、所述第二向量、所述第三向量，以及指定的损失函数，更新所述循环神经网络的参数

步骤6，将所述各词的特征向量分别输入训练后的所述循环神经网络的全连接层进行计算，得到对应的词向量

图 6

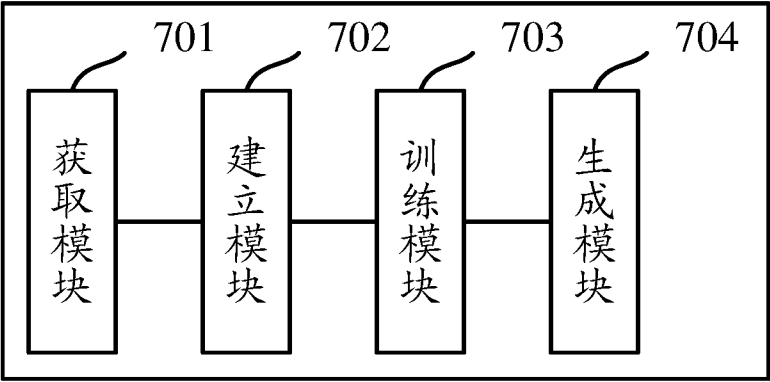


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/072081

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, H04L, G06N

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: 词, 向量, 特征, 循环, 递归, 神经网络, 连续, 字符, 子串, 滑动, 窗口, 上下文, 正样例, 负样例, 生成, 训练, word, vector, feature, RNN, gram, continuous, character, n-gram, slide, window, context, positive, negative, example, generate, train

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 103995805 A (SHENHUA GROUP CORPORATION LIMITED ET AL.) 20 August 2014 (2014-08-20) description, paragraphs [0017]-[0023]	1-4, 9-13, 18, 20
Y	CN 106547735 A (FUDAN UNIVERSITY) 29 March 2017 (2017-03-29) claim 1	1-4, 9-13, 18, 20
A	CN 107273503 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 20 October 2017 (2017-10-20) entire document	1-20
A	CN 105871619 A (INSTITUTE OF INFORMATION ENGINEERING, CHINESE ACADEMY OF SCIENCES) 17 August 2016 (2016-08-17) entire document	1-20
A	CN 107153642 A (NORTH CHINA ELECTRIC POWER UNIVERSITY) 12 September 2017 (2017-09-12) entire document	1-20
A	WO 2017057921 A1 (NAVER CORPORATION) 06 April 2017 (2017-04-06) 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

03 April 2019

Date of mailing of the international search report

17 April 2019

Name and mailing address of the ISA/CN

National Intellectual Property Administration, PRC
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/CN2019/072081

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	103995805	A	20 August 2014	None			
CN	106547735	A	29 March 2017	None			
CN	107273503	A	20 October 2017	US	2018365231	A1	20 December 2018
CN	105871619	A	17 August 2016	None			
CN	107153642	A	12 September 2017	None			
WO	2017057921	A1	06 April 2017	KR	20170039951	A	12 April 2017
				US	2018225553	A1	09 August 2018
				JP	2018533148	A	08 November 2018

国际检索报告

国际申请号

PCT/CN2019/072081

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, H04L, G06N

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: 词, 向量, 特征, 循环, 递归, 神经网络, 连续, 字符, 子串, 滑动, 窗口, 上下文, 正样例, 负样例, 生成, 训练, word, vector, feature, RNN, gram, continuous, character, n-gram, slide, window, context, positive, negative, example, generate, train

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
Y	CN 103995805 A (神华集团有限责任公司 等) 2014年 8月 20日 (2014 - 08 - 20) 说明书第[0017]-[0023]段	1-4, 9-13, 18, 20
Y	CN 106547735 A (复旦大学) 2017年 3月 29日 (2017 - 03 - 29) 权利要求1	1-4, 9-13, 18, 20
A	CN 107273503 A (北京百度网讯科技有限公司) 2017年 10月 20日 (2017 - 10 - 20) 全文	1-20
A	CN 105871619 A (中国科学院信息工程研究所) 2016年 8月 17日 (2016 - 08 - 17) 全文	1-20
A	CN 107153642 A (华北电力大学) 2017年 9月 12日 (2017 - 09 - 12) 全文	1-20
A	WO 2017057921 A1 (NAVER CORPORATION) 2017年 4月 6日 (2017 - 04 - 06) 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

2019年 4月 3日

国际检索报告邮寄日期

2019年 4月 17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

郑宁

电话号码 86-(10)-53961313

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/072081

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	103995805	A	2014年 8月 20日	无			
CN	106547735	A	2017年 3月 29日	无			
CN	107273503	A	2017年 10月 20日	US	2018365231	A1	2018年 12月 20日
CN	105871619	A	2016年 8月 17日	无			
CN	107153642	A	2017年 9月 12日	无			
WO	2017057921	A1	2017年 4月 6日	KR	20170039951	A	2017年 4月 12日
				US	2018225553	A1	2018年 8月 9日
				JP	2018533148	A	2018年 11月 8日

表 PCT/ISA/210 (同族专利附件) (2015年1月)