

(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
2017 年 9 月 28 日 (28.09.2017)



(10) 国际公布号
WO 2017/162134 A1

- (51) 国际专利分类号:
G06F 17/27 (2006.01) *G06F 17/21* (2006.01)
- (21) 国际申请号: PCT/CN2017/077473
- (22) 国际申请日: 2017 年 3 月 21 日 (21.03.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610166105.3 2016 年 3 月 22 日 (22.03.2016) CN
- (71) 申请人 (对除美国外的所有指定国): **索尼公司 (SONY CORPORATION)** [JP/JP]; 日本东京都港区港南 1-7-1, Tokyo 〒108-0075 (JP)。
- (72) 发明人: 及
- (71) 申请人 (仅对美国): **吴友政 (WU, Youzheng)** [CN/CN]; 中国北京市朝阳区太阳宫中路 12 号冠城大厦 701, Beijing 100028 (CN)。 **祁均 (QI, Jun)** [CN/CN]; 中国北京市朝阳区太阳宫中路 12 号冠城大厦 701, Beijing 100028 (CN)。
- (74) 代理人: **北京集佳知识产权代理有限公司 (UNITALEN ATTORNEYS AT LAW)**; 中国北京市朝阳区建国门外大街 22 号赛特广场 7 层, Beijing 100004 (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, 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))。

(54) Title: ELECTRONIC DEVICE AND METHOD FOR TEXT PROCESSING

(54) 发明名称: 用于文本处理的电子设备和方法

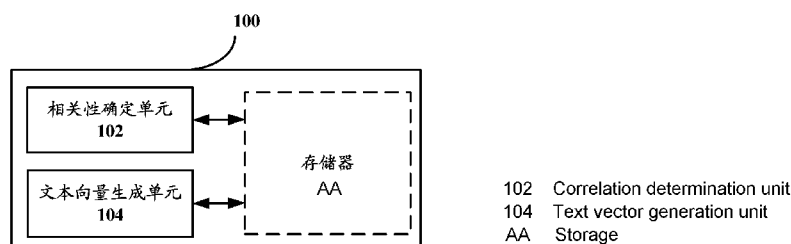


图 1

(57) Abstract: An electronic device and method for text processing, the electronic device comprises a processor (100), and the processor is configured to: determine a correlation between a first text vector and a second text vector, wherein the first text vector and the second text vector are multi-dimensional, real number vectors generated on the basis of a same text, respectively; obtain, according to the correlation, a third text vector representing the text, wherein a vector space in which the third text vector is located is correlated to vector spaces in which first text and second text vectors are located. The electronic device and method of the present invention can be used to create a text-feature representation model which represents text features by combining a plurality of view angles, thereby improving the performance of natural language processing.

(57) 摘要: 一种用于文本处理的电子设备和方法, 该电子设备包括处理器 (100), 该处理器被配置为: 确定第一文本向量与第二文本向量之间的相关性, 第一文本向量和第二文本向量是分别基于同一文本生成的多维实数向量; 以及根据相关性获得第三文本向量以用于表示该文本, 其中, 第三文本向量所在的向量空间与第一文本向量和第二文本向量所在的向量空间相关。该电子设备和方法可以建立结合多个视角进行文本特征表示的文本特征表示模型, 从而能够提高自然语言处理的性能。

WO 2017/162134 A1

用于文本处理的电子设备和方法

本申请要求于 2016 年 3 月 22 日提交中国专利局、申请号为 201610166105.3、发明名称为“用于文本处理的电子设备和方法”的中国专利
5 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

本公开涉及自然语言处理领域，更具体地，涉及一种用于文本处理的
10 的电子设备和方法，其基于两个或更多个词特征表示模型之间的相关性来构建多视角词特征表示模型，以实现对文本对象的特征的深度共享视角表示，从而更有利于后续的自然语言处理。

背景技术

在传统的自然语言理解（Natural Language Understanding, NLU）算法中，
15 文本（例如，词语）被当作一个离散的符号，词语的表示是独立的、离散的，使得词之间并没有很大的关联。例如：

“中国”的表示为 [0001000000000000...]

“北京”的表示为 [00000000000001000...]

该离散向量的维度是词典的大小，因此维度通常较高。这种简洁的离散
20 表示方法通过配合最大熵、支持向量机（Support Vector Machine, SVM）、条件随机场（Condition Random Field, CRF）等统计模型已经很好地完成了自然语言处理（Natural Language Processing, NLP）领域的各种主流任务，例如，词性标注（Part-of-Speech Tagging）、要素抽取（Slot Filling）、命名实体识别（Named Entity Recognition）等。

25 然而，这种离散表示方法通常也意味着我们需要更多的训练数据去成功地训练统计模型，因此运算量较大，并且这种词的独立表示往往不能反映词语之间的语义关联，从而对于自然语言理解可能是不利的。

近年来发展起来的词嵌入（word embedding）技术克服了这些缺点。词嵌入简单来说就是把离散的文本（例如，词语、短语或句子）表示为低维空间的

向量。以词语为例，利用词嵌入技术的词向量表示通常为例如：

“中国” 的表示为 [0.0172, -0.77, -0.507, 0.1, -0.42, ...]

“北京” 的表示为 [0.01, -0.8, -0.5, 0.123, -0.142, ...]

在词嵌入技术中，词向量的纬度以 50 维、100 维、300 维比较常见。由于词嵌入技术考虑了各个文本之间的语义关系，因此各个词语的向量表示并不是完全独立的而是存在一定的语义关联，这样，不仅使得词向量表示的维度大大降低从而降低了计算复杂度，而且还使得这样的词向量表示更加有利于自然语言处理和口语理解中的任务。

C&W, Word2vec 和 GloVe 是近年来被广泛使用的几种词嵌入技术。随着深度学习的发展，词嵌入技术已经成为自然语言处理和口语理解中不可缺少的重要分支，且该技术已经取得了一定的成功。

然而，现有的词嵌入技术仅是从一个视角出发（例如，采用同一种训练机制或者基于同一个训练语料库）来进行词特征表示，这样的词特征表示通常具有局限性，即，在某一方面具有较突出的优点而在其它方面有所欠缺。例如，Word2Vec 依赖于跳元（skip-grams）或连续词袋（Continuous Bag of Words, CBOX）模型来创建词向量从而可以获得长的词上下文，而 GloVe 是基于全局词共现矩阵的非零项训练的，这需要对整个语料库进行遍历以收集统计信息。又例如，针对新闻报导的训练语料库和针对日常口语的训练语料库训练得到的词特征表示在各个词语间的语义关联上各有偏重而具有局限性。

发明内容

在下文中给出了关于本公开的简要概述，以便提供关于本公开的某些方面的基本理解。但是，应当理解，这个概述并不是关于本公开的穷举性概述。它并不是意图用来确定本公开的关键性部分或重要部分，也不是意图用来限定本公开的范围。其目的仅仅是以简化的形式给出关于本公开的某些概念，以此作为稍后给出的更详细描述的前序。

鉴于以上问题，本公开的目的是提供一种用于文本处理的电子设备和方法，其从多个视角出发，根据以不同视角表示文本对象的不同文本特征表示之间的相关性来提供该文本对象的深度共享视角特征表示，以优化执行自然语言

处理和口语理解等任务时的系统性能。

根据本公开的一方面，提供了一种用于文本处理的电子设备，该电子设备包括处理器，该处理器被配置为：确定第一文本向量与第二文本向量之间的相关性，第一文本向量和第二文本向量是分别基于同一文本生成的多维实数向量；以及根据相关性获得第三文本向量以用于表示该文本，其中，第三文本向量所在的向量空间与第一文本向量和第二文本向量所在的向量空间相关。

根据本公开的优选实施例，文本对应于词语。

根据本公开的另一优选实施例，文本对应于以下之一：多个词语组成的短语；以及多个短语组成的句子。

10 根据本公开的另一优选实施例，第一文本向量和第二文本向量分别基于第一词特征表示模型和第二词特征表示模型。

根据本公开的另一优选实施例，第一词特征表示模型和第二词特征表示模型是分别基于不同的词特征表示训练机制得到的。

15 根据本公开的另一优选实施例，词特征表示训练机制包括以下至少之一：Word2Vec 机制、GloVe 机制和 C&W 机制。

根据本公开的另一优选实施例，处理器进一步被配置为：基于典型相关分析来确定第一文本向量与第二文本向量之间的相关性，并且以使得相关性满足预定条件为目标来调整典型相关分析的参数。

20 根据本公开的另一优选实施例，处理器进一步被配置为：利用神经网络对第一文本向量和第二文本向量进行处理以得到第一文本向量的变量和第二文本向量的变量，基于第一文本向量的变量和第二文本向量的变量确定相关性，并且以使得相关性满足预定条件为目标来调整神经网络的参数。

25 根据本公开的另一优选实施例，处理器进一步被配置为：利用自动编码器对第一文本向量的变量和第二文本向量的变量进行处理以重构第一文本向量和第二文本向量，并且以还使得重构后的第一文本向量和第二文本向量与第一文本向量和第二文本向量之间的误差满足预定条件为目标来调整自动编码器和神经网络的参数，以确定相关性。

根据本公开的另一优选实施例，处理器进一步被配置为针对多个文本分别确定相应的第一文本向量与第二文本向量之间的相关性并获得相应的第三

文本向量，并且该电子设备还包括存储器，该存储器被配置为存储多个文本的第三文本向量以用于建立多视角文本特征表示模型。

根据本公开的另一优选实施例，处理器进一步被配置成针对多个文本中的每个文本，还基于关于其它文本的相关性来确定该文本的相应的第一文本向量与第二文本向量之间的相关性。

根据本公开的另一方面，还提供了一种用于文本处理的方法，该方法包括：确定第一文本向量与第二文本向量之间的相关性，第一文本向量和第二文本向量是分别基于同一文本生成的多维实数向量；以及根据相关性获得第三文本向量以用于表示该文本，其中，第三文本向量所在的向量空间与第一文本向量和第二文本向量所在的向量空间相关。

根据本公开的另一方面，还提供了一种用于文本处理的电子设备，该电子设备包括：存储器，被配置为存储多视角文本特征表示模型，其中，该多视角文本特征表示模型是利用上述方法建立的；以及处理器，被配置为从存储器读取多视角文本特征表示模型，并且基于该多视角文本特征表示模型将待处理的文本对象映射为相应的多维实数向量。

根据本公开的另一方面，还提供了一种用于文本处理的方法，该方法包括：从存储器读取多视角文本特征表示模型，其中，该多视角文本特征表示模型是利用上述方法建立的；以及基于多视角文本特征表示模型将待处理的文本对象映射为相应的多维实数向量。

根据本公开的其它方面，还提供了用于实现上述根据本公开的方法的计算机程序代码和计算机程序产品以及其上记录有该用于实现上述根据本公开的方法的计算机程序代码的计算机可读存储介质。另外，还提供了用于承载本公开的多视角文本特征表示模型的计算机可读存储介质。

根据本公开的实施例，通过结合多个视角来表示文本特征，以此来建立多视角文本特征表示模型，可以克服现有技术中单一视角的文本特征表示模型的不足，从而能够提高应用于自然语言处理时的性能。

在下面的说明书部分中给出本公开实施例的其它方面，其中，详细说明用于充分地公开本公开实施例的优选实施例，而不对其施加限定。

附图说明

本公开可以通过参考下文中结合附图所给出的详细描述而得到更好的理解,其中在所有附图中使用了相同或相似的附图标记来表示相同或者相似的部件。所述附图连同下面的详细说明一起包含在本说明书中并形成说明书的一部分,用来进一步举例说明本公开的优选实施例和解释本公开的原理和优点。其中:

图 1 是示出根据本公开的实施例的用于文本处理的电子设备的功能配置示例的框图;

图 2 是示出根据本公开的实施例的基于典型相关分析 (Canonical Correlation Analysis, CCA) 来确定文本向量间的相关性的实现方案的示意图;

图 3 是示出对图 2 所示的方案进一步应用神经网络来确定文本向量间的相关性的实现方案的示意图;

图 4 是示出对图 3 所示的方案进一步应用自动编码器来确定文本向量间的相关性的实现方案的示意图;

图 5 是示出根据本公开的实施例的用于文本处理的电子设备的功能配置示例的框图;

图 6 是示出根据本公开的实施例的用于文本处理的方法的过程示例的流程图;

图 7 是示出根据本公开的实施例的用于文本处理的方法的过程示例的流程图; 以及

图 8 是示出作为本公开的实施例中可采用的信息处理设备的个人计算机的示例结构的框图。

具体实施方式

在下文中将结合附图对本公开的示范性实施例进行描述。为了清楚和简明起见,在说明书中并未描述实际实施方式的所有特征。然而,应该了解,在开发任何这种实际实施方式的过程中必须做出很多特定于实施方式的决定,以便实现开发人员的具体目标,例如,符合与系统及业务相关的那些限制条件,并且这些限制条件可能会随着实施方式的不同而有所改变。此外,还应该了解,

虽然开发工作有可能是非常复杂和费时的,但对得益于本公开内容的本领域技术人员来说,这种开发工作仅仅是例行的任务。

在此,还需要说明的一点是,为了避免因不必要的细节而模糊了本公开,在附图中仅仅示出了与根据本公开的方案密切相关的设备结构和/或处理步骤,而省略了与本公开关系不大的其它细节。

接下来,将参照图 1 至图 8 详细描述本公开的实施例。

首先,将参照图 1 描述根据本公开的实施例的用于文本处理的电子设备的功能配置示例。图 1 是示出根据本公开的实施例的用于文本处理的电子设备的功能配置示例的框图。

如图 1 所示,根据该实施例的电子设备 100 可包括相关性确定单元 102 和文本向量生成单元 104。应指出,这里的相关性确定单元 102 和文本向量生成单元 104 可以是分立的物理实体或逻辑实体,或者也可由同一个物理实体(例如,中央处理单元(CPU)、大规模集成电路(ASIC)等)来实现。

相关性确定单元 102 可被配置成确定第一文本向量与第二文本向量之间的相关性,其中,这里的第一文本向量和第二文本向量是分别基于同一文本生成的多维实数向量。该文本例如可以是词语、由多个词语组成的短语或者由多个短语组成的句子。

以文本是词语为例,第一文本向量和第二文本向量分别基于第一词特征表示模型和第二词特征表示模型,这两个词特征表示模型是分别从不同的视角所建立的词特征表示模型。例如,这两个词特征表示模型是分别基于不同的词特征表示训练机制得到,优选地,这里的词特征表示训练机制可包括 Word2Vec 机制、GloVe 机制和 C&W 机制中的至少一个,即,可从这三种训练机制中选择两种分别作为用于第一词特征表示模型和第二次特征模型的训练机制。这些机制均是现有技术中常用的词嵌入技术,在此不再对其进行详细描述。作为示例,第一词特征表示模型是基于 Word2Vec 机制获得的,而第二词特征表示模型是基于 GloVe 机制获得的。可以理解,随着技术的发展、改进,可能出现其他的主流词特征表示训练机制,本领域技术人员显然也可根据本公开的构思融合基于其他的两种主流词特征表示训练机制得到的词特征表示模型。

另一方面,备选地,这两个词特征表示模型是分别基于不同的训练语料

(corpus) 得到的。例如, 第一词特征表示模型是基于一般语料(例如, 大规模的新闻语料或网页文本)得到的, 而第二词特征表示模型是基于用户固有语料(例如, 邮件语料、口语语料等)训练得到的, 其中用于训练第一词特征表示模型和第二词特征表示模型的训练机制可以相同也可以不同。

5 需要注意, 上述的两种词特征表示模型在一个示例中是执行本公开技术方案的人员自行根据相应的训练机制和语料训练得到的(在线或离线形式), 例如根据其具体的语言处理任务进行针对性的训练得到的, 在另一个示例中是直接

10 从外部获取的, 例如从学术研究共享平台上获取的他人已经训练好的词特征表示模型来作为待融合的词特征表示模型。另外, 本公开主要以两个词特征表示模型的融合作为示例, 然而, 本领域技术人员可以理解, 根据实际需要, 还可以基于本公开执行多于两个词特征表示模型的融合, 例如, 先依据本公开的方案针对第一及第二词特征表示模型进行融合, 将融合得到的第三词特征表示模型再根据本公开的方案与第四词特征表示模型进行融合; 亦可以先依据本公开的方案针对第一及第二词特征表示模型进行融合以得到第三词特征表示模型, 同时针对第四及第五词特征表示模型进行融合以得到第六词特征表示模型, 再将第三词特征表示模型与第六词特征表示模型进行融合, 在此不再赘述。

15

优选地, 相关性确定单元 102 可进一步被配置成基于典型相关分析(CCA)来确定第一文本向量与第二文本向量之间的相关性, 并且以使得该相关性满足预定条件为目标来调整典型相关分析的参数。典型相关分析(CCA)是用于分析

20 两组变量之间的相关关系的一种常用统计分析方法, 在此将其应用于确定词嵌入技术中的两组词特征表示(即, 词向量)之间的相关性。然而, 应理解, 本领域技术人员显然也可想到利用其它相关性分析方法(包括现有的或者未来可能出现的分析方法)来确定第一文本向量与第二文本向量之间的相关性。

在这里, 将简要介绍 CCA。CCA 是用于找到两个随机向量的相关性最大的线性投影的无监督数据分析的标准技术。在数学上, 我们定义两个随机向量

25 (X_1, X_2) , 其协方差矩阵定义为 $(\Sigma_{11}, \Sigma_{22})$ 并且互协方差矩阵定义为 Σ_{12} 。
 $(r_1, r_2) > 0$ 是协方差矩阵 Σ_{11} 和 Σ_{22} 的两个正则项以保证样本协方差的非特异

性。CCA 试图找到两个视角 A_1 、 A_2 的相关性最大的一对线性投影 (A_1^*, A_2^*) ，如以下表达式 (1) 所示：

$$\max_{A_1, A_2} \text{tr}(A_1^T \Sigma_{12} A_2) \quad (1)$$

$$\begin{aligned} \text{s.t.}, & A_1^T (\Sigma_{11} + r_1 I) A_1 = A_2^T (\Sigma_{22} + r_2 I) A_2 = I \\ & u_i^T f(X) g(Y)^T v_j = 0, \quad \forall i \neq j. \end{aligned}$$

表达式 (1) 是经典的半定规划 (semi-definite programming)。假设中间

5 项为 $T = \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$ ，并且令 U_k 和 V_k 为 T 的前 k 个左奇异向量和前 k 个右奇异向量，则最优解为以下表达式 (2) 所示：

$$(A_1^*, A_2^*) = (\Sigma_{11}^{-\frac{1}{2}} U_k, \Sigma_{22}^{-\frac{1}{2}} V_k). \quad (2)$$

在实施例的以下描述中，将以典型相关分析为例来描述本公开的技术，
10 但是应理解，本公开并不限于此。

图 2 是示出根据本公开的实施例的基于典型相关分析来确定文本向量间的相关性的实现方案的示意图。

如图 2 所示，假设 X 和 Y 分别为第一文本向量和第二文本向量，并且 U 和 V 分别为典型相关分析的线性变换参数。根据典型相关分析，这里例如以
15 使得经线性变换后的第一文本向量 ($U^T X$) 与第二文本向量 ($V^T Y$) 之间的相关性最高为优化目标来调整参数 U 和 V ，即，在数学上可以表示为例如以使得 $U^T X$ 与 $V^T Y$ 之间的协方差最小为优化目标来确定参数 U 和 V 的值，其中 $(\cdot)^T$ 表示矩阵的转置。这里应理解，尽管这里以使得 $U^T X$ 与 $V^T Y$ 之间的相关性最高为例来描述如何调整典型相关分析的参数 U 和 V ，但是本公开不限于此，
20 而是也可根据实际情况（例如，计算能力等）而以满足其它预定条件（例如，预定相关性阈值、预定迭代次数等）的相关性为目标来确定典型相关分析的参

数，这同样适用于随后的实施例中的描述。根据优化目标函数来确定线性变换参数 U 和 V 的具体过程是本领域技术人员根据相关数学知识可以实现的，在此不再详细描述。

返回参照图 1，文本向量生成单元 104 可被配置成根据所确定的相关性来
5 获得第三文本向量以表示同一文本。

在图 2 所示的示例中，根据在相关性满足预定条件时所确定的 U 和 V ，可以获得同一文本的两个文本特征表示，即 $U^T X$ 、 $V^T Y$ ，其中任一者皆可作为第三文本向量，换言之，第三文本向量例如可以表示为 $U^T X$ 、 $V^T Y$ 或者基于 $U^T X$ 和 $V^T Y$ 中至少之一确定的向量（例如两者的加权平均等变换形式）。

10 可以理解，如上所述，由于参数 U 和 V 是基于第一文本向量与第二文本向量之间的相关性所确定的，因此所生成的第三文本向量所在的向量空间与第一文本向量和第二文本向量所在的向量空间具有相关性。这样，所生成的第三文本向量考虑了基于不同视角得到的第一文本向量和第二文本向量之间的相关性，因此其是对同一文本的多视角、深度特征表示，能够提高后续的自然语言处理的性能。另外，本公开的技术方案可以是针对既得的至少两个文本特征表示模型进行融合从而易于实现和推广，不必再例如重新统合两种语料进行训练。
15

以上描述了一个文本进行处理以得到新的第三文本向量的示例，类似地，可对多个文本为了进行类似处理以得到相应的第三文本向量的集合，以
20 用于建立多视角文本特征表示模型。

优选地，上述相关性确定单元 102 可进一步被配置成针对多个文本，通过上述方式分别确定相应的第一文本向量与第二文本向量之间的相关性，并且文本向量生成单元 104 可进一步被配置成根据关于各个文本所确定的相关性来获得相应的第三文本向量。

25 优选地，该电子设备 100 还可包括存储器 106，存储器 106 可被配置为存储这多个文本的第三文本向量以用于建立多视角文本特征表示模型，该多视角文本特征表示模型表示从文本对象到基于多个视角确定的文本向量的映射，可用于执行后续的自然语言处理中的各种任务。

此外，优选地，相关性确定单元 102 可进一步被配置成针对多个文本中

的每个文本，还基于关于其它文本的相关性来确定该文本的第一文本向量与第二文本向量之间的相关性。根据上述方式，针对每个文本，可仅基于该文本自身的第一个文本向量与第二个文本向量之间的相关性来确定其对应的第三个文本向量，然后根据这些分别确定的第三个文本向量的集合来建立新的多视角文本特征表示模型。然而，通常地，当对特定文本集合进行处理以建立新的文本特征表示模型时，取代逐文本地确定第三个文本向量，还可基于针对该文本集合的第一个文本向量集合与第二个文本向量集合之间的整体相关性来确定针对该文本集合的典型相关分析的参数，即，在确定关于特定文本的相关性时还需要将其它文本的相关性纳入考虑，由此来确定针对该文本集合的第三个文本向量集合，从而用于建立多视角文本特征表示模型。

利用典型相关分析技术、以文本集合作为整体来确定第一个文本向量集合与第二个文本向量集合之间的相关性的具体实现过程可参见典型相关分析技术的原理，在此不再详细描述。此外，应指出，在以下参照图 3 和图 4 描述的确定的相关性的示例实现方案中，均是以多个文本的集合作为整体来确定相关性，但是替代地也可逐文本地来确定相关性，从而根据相关性来确定相应的第三个文本向量集合以用于建立多视角文本特征表示模型，本领域技术人员可根据实际情况而选择具体的实现方式，本公开对此不作限制。

优选地，还可进一步利用神经网络来确定上述相关性。图 3 是示出对图 2 所示的方案进一步应用神经网络来确定文本向量间的相关性的实现方案的示意图。

如图 3 所示，在图 2 所示的方案的基础上，进一步添加了两个独立的深度神经网络（Deep Neural Network, DNN）以对所输入的两个文本向量 X 和 Y（这里的 X 和 Y 也可表示文本向量集合）进行非线性变换，然后再利用典型相关分析（CAA）来确定非线性变换后的向量之间的相关性，该方案在下文中也可以称为深度典型相关分析（Deep Canonical Correlation Analysis, DCCA）。然而，应理解，尽管这里以深度神经网络与典型相关分析的组合为例来确定文本向量间的相关性，但是如上所述，也可利用深度神经网络与其它相关性分析技术的组合来执行该确定。此外，这里利用两个独立的深度神经网络来进行非线性变换是为了降低计算复杂度，在不考虑计算复杂度的情况下，当然也可利

用一个深度神经网络来对第一和第二文本向量进行非线性变换。

在图 3 所示的示例中, 符号 X、Y、U 和 V 的含义与以上参照图 2 描述的相同, 在此不再重复, $f(\cdot)$ 和 $g(\cdot)$ 分别表示两个深度神经网络的非线性变换, 其参数分别为 W_f 和 W_g 。根据图 3 所示的方案, 第一文本向量 X 和第二文本向量 Y 首先经过深度神经网络以接受非线性变换, 并且变换后的第一文本向量的变量和第二文本向量的变量分别记为 $f(X)$ 和 $g(Y)$ 。然后, 利用 CCA 对 $f(X)$ 和 $g(Y)$ 分别进行线性变换, 并且以使得线性变换后的 $f(X)$ 和 $g(Y)$ (即, $U^T f(X)$ 和 $V^T g(Y)$) 之间的相关性最大化为目标来调整典型相关分析的参数 (即, U 和 V) 和深度神经网络的参数, 深度神经网络的参数可包括上述 W_f 和 W_g , 另外还可以包括其结构参数 (包括深度神经网络的层数和每层上的维度), 从而可以确定最终的第三文本向量为 $U^T f(X)$ 、 $V^T g(Y)$ 或者基于 $U^T f(X)$ 和 $V^T g(Y)$ 中至少之一确定的向量 (例如两者的加权平均等变换形式)。其中, 深度神经网络的结构参数也可以是根据运算系统环境等因素预定义的, 根据本发明的一个示例预定结构为 4 层, 每层的维度分别为 100、1024、1024 和 100。

上述计算过程在数学上可以表示为寻找使得 $U^T f(X)$ 与 $V^T g(Y)$ 之间的协方差最小的 U、V、 W_f 和 W_g , 例如可以表示为如下表达式 (3):

$$\begin{aligned} \min_{W_f, W_g, U, V} & -\frac{1}{N} \text{tr}(U^T f(X) g(Y)^T V) \\ \text{s.t.}, & U^T \left(\frac{1}{N} f(X) f(X)^T + r_x I \right) U = I \\ & V^T \left(\frac{1}{N} g(Y) g(Y)^T + r_y I \right) V = I \\ & u_i^T f(X) g(Y)^T v_j = 0, \forall i \neq j, \end{aligned} \quad (3)$$

其中, N 表示文本向量集合的总数, I 表示单位矩阵, 并且 $(r_x, r_y) > 0$ 是用于协方差估计的正则化参数。

如何根据上述目标优化函数对模型进行训练以确定深度神经网络的参数 W_f 和 W_g 以及 CCA 的线性变换参数 U 和 V 是本领域技术人员根据掌握的数学知识可以实现的, 这并不是本公开的技术的重点, 因此在此不再详细描述。例如, 可以使用例如受限玻尔兹曼 (Restricted Boltzmann Machine, RBM) 技术来进行模型的预训练, 然后使用反向传播 (Back-propagation) 例如随机梯度下降 (Stochastic Gradient Descent, SGD) 技术, 基于使得文本向量间的相关

性最大的目标函数对深度神经网络的参数 W_f 和 W_g 以及 CCA 的线性变换参数 U 和 V 进行联合优化学习。根据一个示例，利用上述的随机梯度下降方案对 DNN 的参数进行精细调整，例如先确定 DNN 顶层的梯度 (δ)，再根据梯度调整 DNN 顶层参数如 $W'_f = W_f + \text{调整系数} \times \text{梯度}$ ，进而再推算 DNN 其他层

5 次的参数。其中，DNN 顶层的梯度可以通过基于目标函数（即公式（3））分别针对 H_x 和 H_y 对 $\text{corr}(H_x, H_y)$ 进行求导来获得，其中， $\text{corr}(H_x, H_y)$ 表示 H_x 和 H_y 的相关度， $H_x = U^T f(X)$ 以及 $H_y = V^T g(Y)$ 。

在利用局部或优选的全局训练数据基于 DCCA 方案对深度神经网络的参数以及 CCA 的线性变换参数完成训练之后，可以得到确定的深度神经网络的

10 参数 W_f 和 W_g 及 CCA 参数 U 和 V 。届时，逐个地针对每个文本，将其第一文本向量 X 和第二文本向量 Y 分别输入相应的神经网络 $f(\cdot)$ 和 $g(\cdot)$ 中，之后再由 CCA 进行变换即可获得该文本的目标第三文本向量例如 $U^T f(X)$ ，直至完成所有待处理文本的文本向量变换。其中，局部或全局是相对于待处理的全部文本集合而言的，本领域技术人员可以根据其面对的具体语言处理任务，从全部

15 文本集合中抽取有关的文本作为局部训练数据进行训练以提高效率，也可以根据其对模型精度的要求和运算资源来选择全局或局部训练数据。

此外，应指出，以上给出的目标优化函数仅为示例而非限制，并且本领域技术人员也可根据具体的优化目标，基于本公开的原理而设计适合实际需要的目标函数。

20 应指出，上述典型相关分析 (CCA) 和深度典型相关分析 (DCCA) 均属于无监督学习过程，因此，在确定相关性的过程中，尽管可能获得了第一文本向量与第二文本向量之间的较高相关性，但是在此过程中，可能会使得此时获得的第三文本向量与第一文本向量和/或第二文本向量的差别较大，即，导致较大的失真，这有可能在一定程度上影响后续的自然语言处理的性能。鉴于此，

25 为了进一步优化系统性能，还可通过利用自动编码器重构第一文本向量和第二文本向量，以在最大化相关性的同时最小化自动编码错误来调整相关参数，从而确定相应的第三文本向量。以下将参照图 4 描述该情况下的实现方案。

图 4 是示出对图 3 所示的方案进一步应用自动编码器来确定文本向量间的相关性的实现方案的示意图。

如图 4 所示, 在图 3 所示的方案的基础上, 进一步加入了两个自动编码器 (auto-encoder) 以对经过深度神经网络非线性变换后的第一文本向量和第二文本向量进行重构, 该方案在下文可称为深度典型相关自动编码 (Deep Canonically Correlated Auto-Encoders, DCCAE)。类似地, 如上所述, 在该实现方案中, 也可应用除 CCA 之外的技术来确定相关性。

在图 4 所示的示例中, 符号 X 、 Y 、 U 、 V 、 $f(\cdot)$ 和 $g(\cdot)$ 的含义与以上参照图 3 描述的相同, 在此不再重复, 符号 $p(\cdot)$ 和 $q(\cdot)$ 分别表示用于重构的自动编码器 (即, 深度神经网络) 的非线性变换, 其参数分别为 W_p 和 W_q 。

根据图 4 所示的方案, 经过深度神经网络的非线性变换后的第一文本向量的变量 $f(X)$ 和第二文本向量的变量 $g(Y)$ 同时被输入到 CCA 模块和自动编码器模块以分别接受相关性分析和重构, 并且以在使得线性变换后的 $f(X)$ 和 $g(Y)$ (即, $U^T f(X)$ 和 $V^T g(Y)$) 之间的相关性最大化的同时使得自动编码误差 (即, 重构后的第一文本向量 $p(f(X))$ 和第二文本向量 $q(g(Y))$ 分别与原始的第一文本向量 X 和第二文本向量 Y 之间的差的绝对值 $|p(f(X)) - X|$ 和 $|q(g(Y)) - Y|$) 最小化为优化目标来调整典型相关分析的参数 (即, U 和 V)、深度神经网络的参数 (即, W_f 和 W_g) 以及自动编码器的参数 (即, W_p 和 W_q), 从而可以确定最终的第三文本向量为 $U^T f(X)$ 、 $V^T g(Y)$ 或者基于 $U^T f(X)$ 和 $V^T g(Y)$ 中至少之一确定的向量。

上述计算过程在数学上可以例如表示为寻找使得 $U^T f(X)$ 与 $V^T g(Y)$ 之间的协方差与 $p(f(X))$ 与 X 之间的差的绝对值以及 $q(g(Y))$ 与 Y 之间的差的绝对值的最小的 U 、 V 、 W_f 、 W_g 、 W_p 和 W_q , 这例如可以表示为如下表达式 (4):

$$\begin{aligned}
 & \min_{W_f, W_g, W_p, W_q, U, V} -\frac{1}{N} \text{tr}(U^T f(X) g(Y)^T V) \\
 & + \frac{\lambda}{N} (\|X - p(f(X))\|_F^2 + \|Y - q(g(Y))\|_F^2) \\
 & s.t., U^T \left(\frac{1}{N} f(X) f(X)^T + r_x I \right) U = I \\
 & \quad V^T \left(\frac{1}{N} g(Y) g(Y)^T + r_y I \right) V = I \\
 & \quad u_i^T f(X) g(Y)^T v_j = 0, \forall i \neq j,
 \end{aligned} \tag{4}$$

在表达式 (4) 中, 与以上表达式 (3) 中的符号相同的符号表示相同的

含义, 在此不再重复。 λ 是与用于控制自动编码器的水平的归一化常数 (实际上是控制自动编码错误在目标函数中所占的比例), 其为经验值或者通过有限次实验确定的值。

如何根据该目标表达式来对相关参数进行联合优化学习可参见以上针对 DCCA 方案的描述, 在此不再重复。此外, 应理解, 该目标函数仅是示例而非限制, 并且本领域技术人员可以根据实际的设计目标而对该目标函数进行修改。在利用局部或全局训练数据基于 DCCAE 方案对深度神经网络的参数以及 CCA 的线性变换参数完成训练之后, 可以得到确定的深度神经网络的参数 W_f 和 W_g 及 CCA 参数 U 和 V 。届时, 逐个地针对每个文本, 将其第一文本向量 X 和第二文本向量 Y 分别输入相应的神经网络 $f(\cdot)$ 和 $g(\cdot)$ 中, 之后再由 CCA 进行变换即可获得该文本的目标第三文本向量例如 $U^T f(X)$, 直至完成所有待处理文本的文本向量变换。

应指出, 尽管以上参照图 2 至图 4 描述了确定文本向量间的相关性的示例实现方案, 但是应理解, 这仅是示例而非限制, 并且本领域技术人员可根据本公开的原理而对上述实现方案进行修改。例如, 优化目标函数可以不是使得相关性最大化, 而是预设的最大迭代次数或者满足预定阈值的相关性等, 或者也可采用除 CCA 之外的相关性分析技术等等, 并且这样的变型均认为落入本公开的范围。

通过利用上述 CCA、DCCA、DCCAE 方案来获得文本向量间的相关性, 由于结合了多个视角来表示文本特征, 因此能够获得深度多视角文本特征表示模型, 从而能够提高自然语言理解等中的任务的性能。

接下来, 将参照图 5 描述利用上述获得的多视角文本特征表示模型来进行文本处理的实施例。图 5 是示出根据本公开的实施例的用于文本处理的电子设备的功能配置示例的框图。

如图 5 所示, 根据该实施例的电子设备 500 可包括存储器 502 和处理器 504。

存储器 502 可被配置为存储上述所建立的多视角文本特征表示模型。

处理器 504 可被配置为从存储器 502 读取多视角文本特征表示模型, 并且基于该多视角文本特征表示模型而将待处理的文本对象映射为相应的多维

实数向量。该待处理的文本对象可存储在存储器 502 或者外部存储器中，或者也可以是由用户输入的，例如用户输入语音，由语音识别模块将语音转化为文本，进而由本公开的方案进行处理。

该文本对象例如可以是词语，并且该多视角文本特征表示模型例如是词特征表示模型。在该情况下，当对短语、句子或段落进行处理时，处理器 504 可通过利用现有的词划分技术将该短语、句子或段落适当地划分为多个词语单元，并基于该词特征表示模型而将这多个词语单元分别映射为相应的词向量，以用于执行要素抽取、语句分类、自动翻译等自然语言理解处理。

5 备选地，在所建立的多视角文本特征表示模型例如是短语或句子等文本对象的特征表示模型时，取代将短语、句子、段落等划分为相应的词单元，可通过直接映射、将句子或段落划分为短语或者将段落划分为句子等方式，基于多视角文本特征表示模型将这些文本对象映射为相应的文本向量，并且基于这些文本向量对这些短语、句子或段落进行理解。即，在实际处理过程中，可能还需要进行词语划分的处理，该处理可采用现有技术中公知的技术，并且与本发明的发明点不相关，因此在此不详细描述。

15 利用所建立的文本特征表示模型来进行自然语言理解等处理的具体过程与现有技术中相同，在此不再详细描述。

在这里，应指出，尽管以上参照图 1 和图 5 描述了用于文本处理的电子设备的功能配置示例，但是这仅是示例而非限制，并且本领域技术人员可根据本公开的原理而对上述功能配置进行修改。例如，所示出的各个功能单元可以进行组合、进一步划分或者添加另外的功能单元，并且这样的变型应认为落入本公开的范围。

与上述装置实施例相对应的，本公开还提供了以下方法实施例。接下来，将参照图 6 和图 7 描述根据本公开的实施例的用于文本处理的方法的过程示例。

25 图 6 是示出根据本公开的实施例的用于文本处理的方法的过程示例的流程图。该方法对应于以上参照图 1 描述的用于文本处理的电子设备的实施例。

如图 6 所示，首先，在步骤 S610 中，确定第一文本向量与第二文本向量之间的相关性，第一文本向量和第二文本向量是分别基于同一文本生成的多维

实数向量。

接下来，在步骤 S620 中，根据所确定的相关性获得第三文本向量以用于表示该文本，第三文本向量所在的向量空间与第一文本向量和第二文本向量所在的向量空间相关。

5 优选地，该文本对应于词语、由多个词语构成的短语或者由多个短语构成的句子。

10 优选地，第一文本向量和第二文本向量分别基于第一词特征表示模型和第二词特征表示模型，第一词特征表示模型和第二词特征表示模型是分别基于不同的词特征表示训练机制以及/或者不同的训练语料得到的。词特征表示训练机制可包括以下至少之一：Word2Vec 机制、GloVe 机制和 C&W 机制，即，可从这三种训练机制中选择两种分别作为用于第一次特征表示模型和第二词特征表示模型的训练机制。

15 优选地，该方法还包括：基于典型相关分析确定第一文本向量与第二文本向量之间的相关性，并且以使得相关性满足预定条件为目标来调整典型相关分析的参数。

优选地，该方法还包括：针对多个文本分别确定相应的第一文本向量与第二文本向量之间的相关性并获得相应的第三文本向量；以及基于多个文本的第三文本向量建立多视角文本特征表示模型。

20 此外，优选地，该方法还可包括基于上述 DCCA 和 DCCAE 等方案来确定文本向量间的相关性。

利用 CCA、DCCA 和 DCCAE 等方案来确定文本向量间的相关性以生成相应的第三文本向量从而建立多视角文本特征表示模型的具体处理过程可参见以上装置实施例中相应位置的描述，在此不再重复。

25 图 7 是示出根据本公开的实施例的用于文本处理的方法的过程示例的流程图。该方法对应于以上参照图 5 描述的用于文本处理的电子设备的实施例。

如图 7 所示，首先，在步骤 S710 中，从存储器读取上述建立的多视角文本特征表示模型。接下来，在步骤 S720 中，基于该多视角文本特征表示模型而将待处理的文本对象映射为相应的多维实数向量。该待处理的文本对象可存储在内部存储器或者外部存储器中，或者也可以是用户输入的。

优选地，该文本对象可对应于词语，并且该方法还可包括基于该文本对象的多维实数向量而对包含该文本对象的短语、句子和段落中至少之一进行文本理解。

应理解，图 6 和图 7 所示的方法实施例的流程图仅是示例而非限制，并且本领域技术人员可根据本公开的原理而对上述处理步骤进行修改，例如，对上述处理步骤进行添加、删除、组合和/或变更等，并且这样的变型都应认为落入本公开的范围。

此外，还应指出，这里参照图 6 和图 7 描述的方法实施例分别与以上参照图 1 和图 5 描述的装置实施例相对应，因此在此未详细描述的内容可参见以上装置实施例中相应位置的描述，而在此不再重复。

当将根据本公开的实施例所建立的多视角文本特征表示模型应用于执行自然语言理解中的任务时，其能够有效地优化处理性能。下面将作为示例给出当分别将根据现有技术构建的文本特征表示模型和根据本发明的 CCA、DCCA 和 DCCAE 方案分别建立的多视角文本特征表示模型应用于口语理解中的要素抽取任务时，各个模型之间的处理性能对比。

应理解，尽管这里给出了要素抽取任务作为示例来检验本发明的实际效果，但是本发明还可以应用于自然语言理解中的任何其它任务，诸如上述词性标注、命名实体识别等任务。也就是说，本公开的例如电子设备 500 实际上还可以包括要素提取模块、词性标注模块或命名实体识别模块等高层的自然语言处理模块，响应于基于多视角文本特征表示模型对待处理的文本映射得到的多维实数向量，上述高层语言处理模块进一步执行相应的自然语言理解。其中，要素抽取的任务具体来说就是抽取输入句子中的要素并且进行标记。例如，在该对比实验中，作为示例，采用的数据集合是航空交通信息系统（Air Travel Information System, ATIS），并且要素抽取的具体任务是：今天从波士顿到西雅图的航班，执行要素抽取后的结果为以下表 1 所示：

表 1 要素抽取的结果

输入的句子	今天	从	波士顿	到	西雅图	的	航班
输出的要素标注结果	B-日期	0	B-出发地	0	B-到达地	0	0

其中，今天是日期的起始词（B-日期），波士顿是出发地的起始词（B-出发地），西雅图是到达地的起始词（B-到达地），并且“0”表示非要素词。应指出，根据该示例，本公开的方案可应用于例如航旅订票系统、日程安排系统等产品中。当然，由于本公开的方案涉及基础的词嵌入技术，还可以广泛地被应用于多种其他语言理解场景。

近年来的研究表明，循环神经网络（Recurrent Neural Network，RNN）在要素抽取任务中能够获得更好的性能，因此，在本实验中，分别采用两种类型的RNN（即，埃尔曼型RNN和乔丹型RNN）来验证本发明的效果，并且参与实验对比的词嵌入技术包括：随机法、Word2Vec、GloVe、基于Word2Vec和GloVe的CCA方案、DCCA方案以及DCCAE方案。

这里用于衡量要素抽取任务中的性能的指标定义为F1测度，其表示准确率和召回率的调和平均值。以下表2示出了实验对比结果：

表2 性能比较结果

	随机法	Word2Vec	GloVe	CCA	DCCA	DCCAE
埃尔曼型 RNN	94.77	94.89	94.65	95.02	95.44	95.48
乔丹型 RNN	93.40	93.90	93.67	94.15	94.37	94.45

从上表可以看出，无论是哪种类型的循环神经网络，根据本公开的技术所建立的多视角文本特征表示模型均能够实现更优的性能。

此外，尽管这里未具体描述，但是根据本公开的技术所建立的多视角文本特征表示模型在其它自然语言理解任务中同样可以实现更优的性能。

应理解，根据本公开的实施例的存储介质和程序产品中的机器可执行的指令还可以被执行以上描述的用于文本处理的方法，因此在此未详细描述的部分可参考先前相应位置的描述，在此不再重复进行描述。

相应地，用于承载上述存储有机器可读取的指令代码的程序产品的存储介质以及用于承载本公开的多视角文本特征表示模型的存储介质也包括在本发明的公开中。所述存储介质包括但不限于软盘、光盘、磁光盘、存储卡、存储棒等等。

另外，还应该指出的是，上述系列处理和装置也可以通过软件和/或固件

实现。在通过软件和/或固件实现的情况下，从存储介质或网络向具有专用硬件结构的计算机，例如图 8 所示的通用个人计算机 800 安装构成该软件的程序，该计算机在安装有各种程序时，能够执行各种功能等等。

5 在图 8 中，中央处理单元（CPU）801 根据只读存储器（ROM）802 中存储的程序或从存储部分 808 加载到随机存取存储器（RAM）803 的程序执行各种处理。在 RAM 803 中，也根据需要存储当 CPU 801 执行各种处理等等时所需的数据。

CPU 801、ROM 802 和 RAM 803 经由总线 804 彼此连接。输入/输出接口 805 也连接到总线 804。

10 下述部件连接到输入/输出接口 805：输入部分 806，包括键盘、鼠标等等；输出部分 807，包括显示器，比如阴极射线管（CRT）、液晶显示器（LCD）等等，和扬声器等等；存储部分 808，包括硬盘等等；和通信部分 809，包括网络接口卡比如 LAN 卡、调制解调器等等。通信部分 809 经由网络比如因特网执行通信处理。

15 根据需要，驱动器 810 也连接到输入/输出接口 805。可拆卸介质 811 比如磁盘、光盘、磁光盘、半导体存储器等等根据需要被安装在驱动器 810 上，使得从中读出的计算机程序根据需要被安装到存储部分 808 中。

在通过软件实现上述系列处理的情况下，从网络比如因特网或存储介质比如可拆卸介质 811 安装构成软件的程序。

20 本领域的技术人员应当理解，这种存储介质不局限于图 8 所示的其中存储有程序、与设备相分离地分发以向用户提供程序的可拆卸介质 811。可拆卸介质 811 的例子包含磁盘（包含软盘（注册商标）、光盘（包含光盘只读存储器（CD-ROM）和数字通用盘（DVD）、磁光盘（包含迷你盘（MD）（注册商标）和半导体存储器。或者，存储介质可以是 ROM 802、存储部分 808 中
25 包含的硬盘等等，其中存有程序，并且与包含它们的设备一起被分发给用户。

还需要指出的是，执行上述系列处理的步骤可以自然地根据说明的顺序按时间顺序执行，但是并不需要一定根据时间顺序执行。某些步骤可以并行或彼此独立地执行。

例如，在以上实施例中包含在一个单元中的多个功能可以由分开的装置

来实现。备选地，在以上实施例中由多个单元实现的多个功能可分别由分开的装置来实现。另外，以上功能之一可由多个单元来实现。无需说，这样的配置包括在本公开的技术范围内。

5 在该说明书中，流程图中所描述的步骤不仅包括以所述顺序按时间序列执行的处理，而且包括并行地或单独地而不是必须按时间序列执行的处理。此外，甚至在按时间序列处理的步骤中，无需说，也可以适当地改变该顺序。

10 虽然已经详细说明了本公开及其优点，但是应当理解在不脱离由所附的权利要求所限定的本公开的精神和范围的情况下可以进行各种改变、替代和变换。而且，本公开实施例的术语“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、物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、物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、物品或者设备中还存在另外的相同要素。

15

权 利 要 求

1. 一种用于文本处理的电子设备，所述电子设备包括处理器，所述处理器被配置为：

5 确定第一文本向量与第二文本向量之间的相关性，所述第一文本向量和所述第二文本向量是分别基于同一文本生成的多维实数向量；以及

根据所述相关性获得第三文本向量以用于表示所述文本，其中，所述第三文本向量所在的向量空间与所述第一文本向量和所述第二文本向量所在的向量空间相关。

2. 根据权利要求 1 所述的电子设备，其中，所述文本对应于词语。

10 3. 根据权利要求 1 所述的电子设备，其中，所述文本对应于以下之一：多个词语组成的短语；以及多个短语组成的句子。

4. 根据权利要求 2 所述的电子设备，其中，所述第一文本向量和所述第二文本向量分别基于第一词特征表示模型和第二词特征表示模型。

15 5. 根据权利要求 4 所述的电子设备，其中，所述第一词特征表示模型和所述第二词特征表示模型是分别基于不同的词特征表示训练机制得到的。

6. 根据权利要求 5 所述的电子设备，其中，所述词特征表示训练机制包括以下至少之一：Word2Vec 机制、GloVe 机制和 C&W 机制。

7. 根据权利要求 4 或 5 所述的电子设备，其中，所述第一词特征表示模型和所述第二词特征表示模型是分别基于不同的训练语料得到的。

20 8. 根据权利要求 1 所述的电子设备，其中，所述处理器进一步被配置为：基于典型相关分析来确定所述第一文本向量与所述第二文本向量之间的相关性，并且以使得所述相关性满足预定条件为目标来调整所述典型相关分析的参数。

25 9. 根据权利要求 1 或 8 所述的电子设备，其中，所述处理器进一步被配置为：利用神经网络对所述第一文本向量和所述第二文本向量进行处理以得到所述第一文本向量的变量和所述第二文本向量的变量，基于所述第一文本向量的变量和所述第二文本向量的变量确定所述相关性，并且以使得所述相关性满足预定条件为目标来调整所述神经网络的参数。

10. 根据权利要求 9 所述的电子设备，其中，所述处理器进一步被配置为：

利用自动编码器对所述第一文本向量的变量和所述第二文本向量的变量进行处理以重构所述第一文本向量和所述第二文本向量,并且以还使得重构后的第一文本向量和第二文本向量与所述第一文本向量和所述第二文本向量之间的误差满足预定条件为目标来调整所述自动编码器和所述神经网络的参数,以确
5 定所述相关性。

11. 根据权利要求 1 至 10 中任一项所述的电子设备,其中,所述处理器进一步被配置为针对多个文本分别确定相应的第一文本向量与第二文本向量之间的相关性并获得相应的第三文本向量,并且所述电子设备还包括存储器,所述存储器被配置为存储所述多个文本的第三文本向量以用于建立多视角文
10 本特征表示模型。

12. 根据权利要求 11 所述的电子设备,其中,所述处理器进一步被配置成针对所述多个文本中的每个文本,还基于关于其它文本的所述相关性来确定该文本的相应的第一文本向量与第二文本向量之间的相关性。

13. 一种用于文本处理的方法,包括:

15 确定第一文本向量与第二文本向量之间的相关性,所述第一文本向量和所述第二文本向量是分别基于同一文本生成的多维实数向量;以及

根据所述相关性获得第三文本向量以用于表示所述文本,其中,所述第三文本向量所在的向量空间与所述第一文本向量和所述第二文本向量所在的向量空间相关。

20 14. 根据权利要求 13 所述的方法,其中,所述文本对应于词语。

15. 根据权利要求 14 所述的方法,其中,所述第一文本向量和所述第二文本向量分别基于第一词特征表示模型和第二词特征表示模型,所述第一词特征表示模型和所述第二词特征表示模型是分别基于不同的词特征表示训练机制以及/或者不同的训练语料得到的。

25 16. 根据权利要求 13 所述的方法,其中,还包括:基于典型相关分析确定所述第一文本向量与所述第二文本向量之间的相关性,并且以使得所述相关性满足预定条件为目标来调整所述典型相关分析的参数。

17. 根据权利要求 13 至 16 中任一项所述的方法,还包括:针对多个文本分别确定相应的第一文本向量与第二文本向量之间的相关性并获得相应的第

三文本向量；以及基于所述多个文本的第三文本向量建立多视角文本特征表示模型。

18. 一种用于文本处理的电子设备，包括

5 存储器，被配置为存储多视角文本特征表示模型，其中，所述多视角文本特征表示模型是利用根据权利要求 17 所述的方法建立的；以及

处理器，被配置为从所述存储器读取所述多视角文本特征表示模型，并且基于所述多视角文本特征表示模型将待处理的文本对象映射为相应的多维实数向量。

19. 根据权利要求 18 所述的电子设备，其中，所述文本对象对应于词语。

10 20. 根据权利要求 19 所述的电子设备，其中，所述处理器进一步被配置为基于所述文本对象的多维实数向量对包含所述文本对象的短语、句子和段落中至少之一进行文本理解。

21. 一种用于文本处理的方法，包括：

15 从存储器读取多视角文本特征表示模型，其中，所述多视角文本特征表示模型是利用根据权利要求 17 所述的方法建立的；以及

基于所述多视角文本特征表示模型将待处理的文本对象映射为相应的多维实数向量。

22. 根据权利要求 21 所述的方法，其中，所述文本对象对应于词语。

20 23. 根据权利要求 22 所述的方法，还包括：基于所述文本对象的多维实数向量对包含所述文本对象的短语、句子和段落中至少之一进行文本理解。

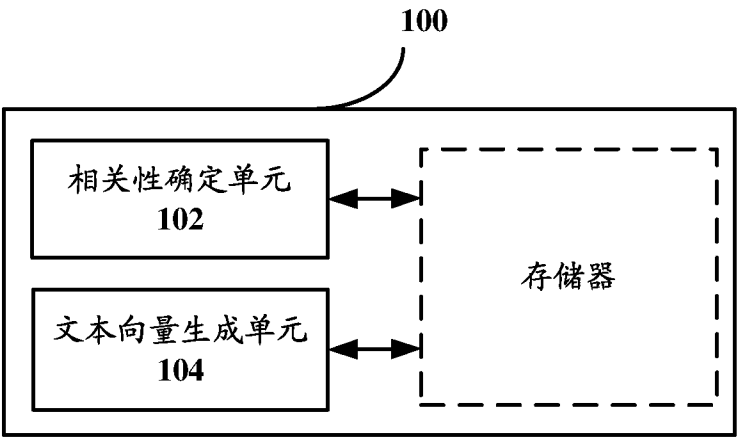


图 1

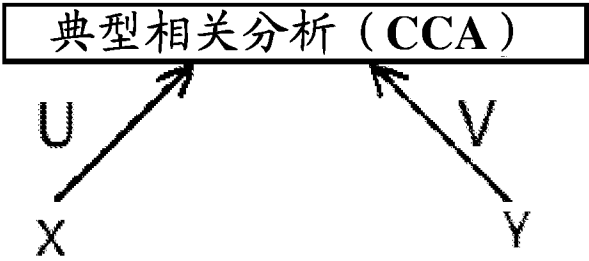


图 2

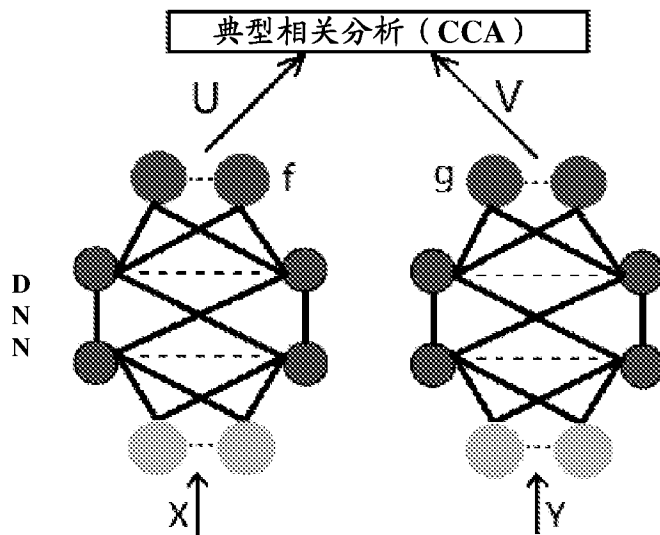


图 3

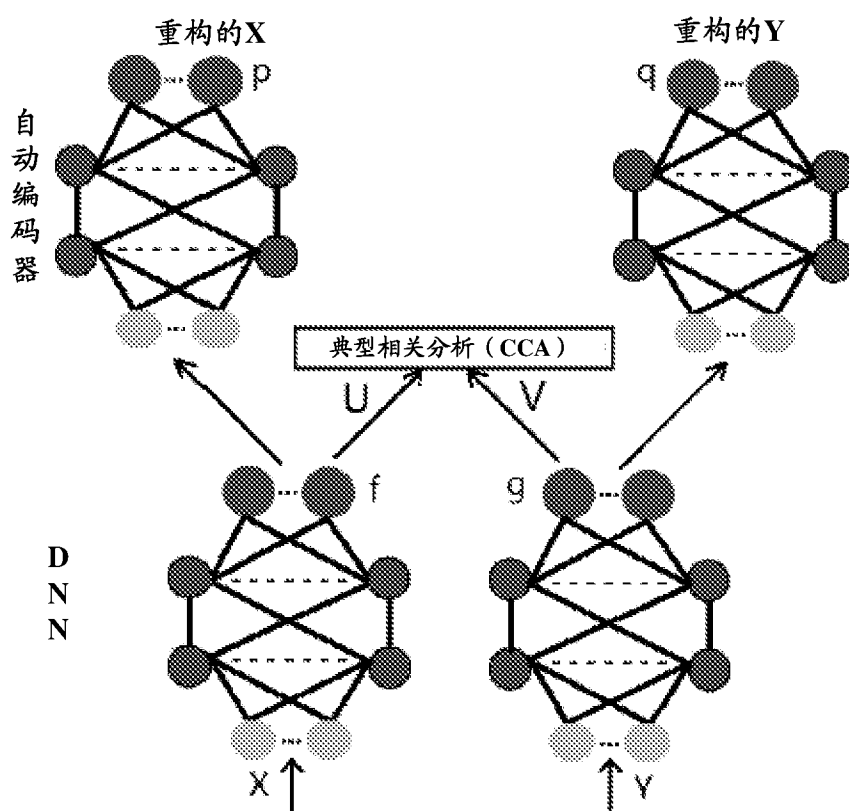


图 4

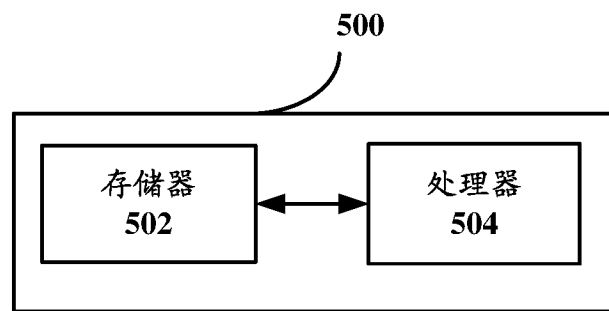


图 5

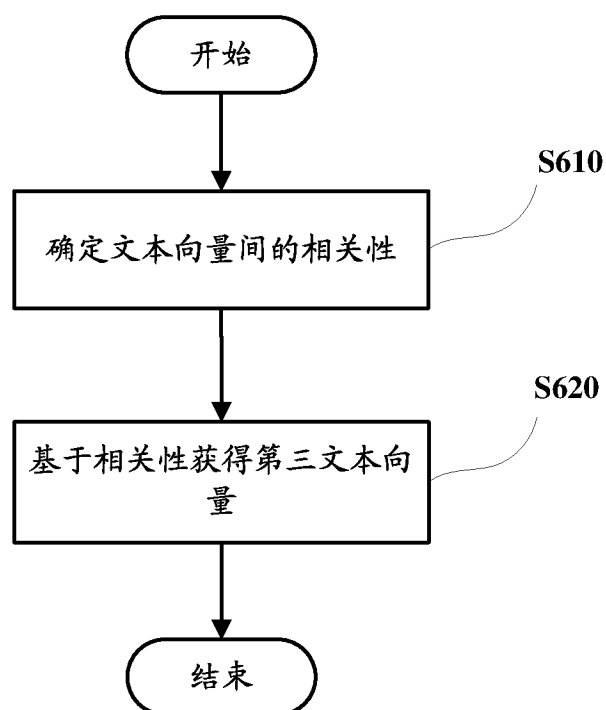


图 6

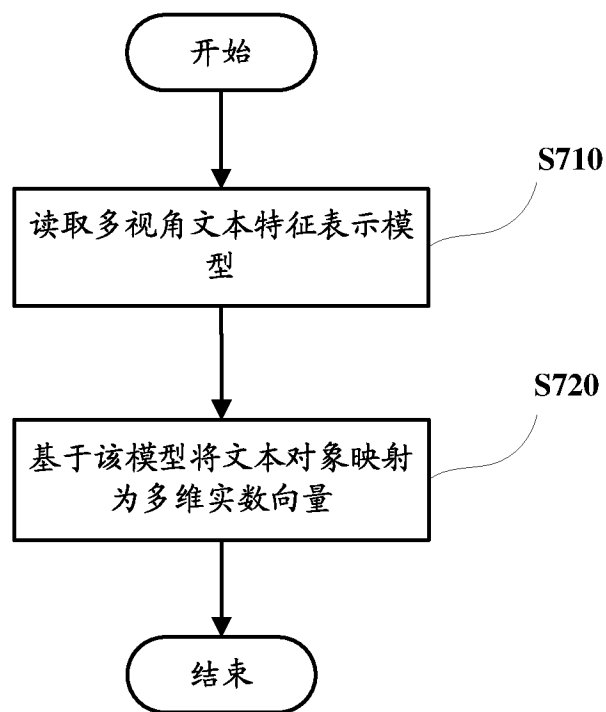
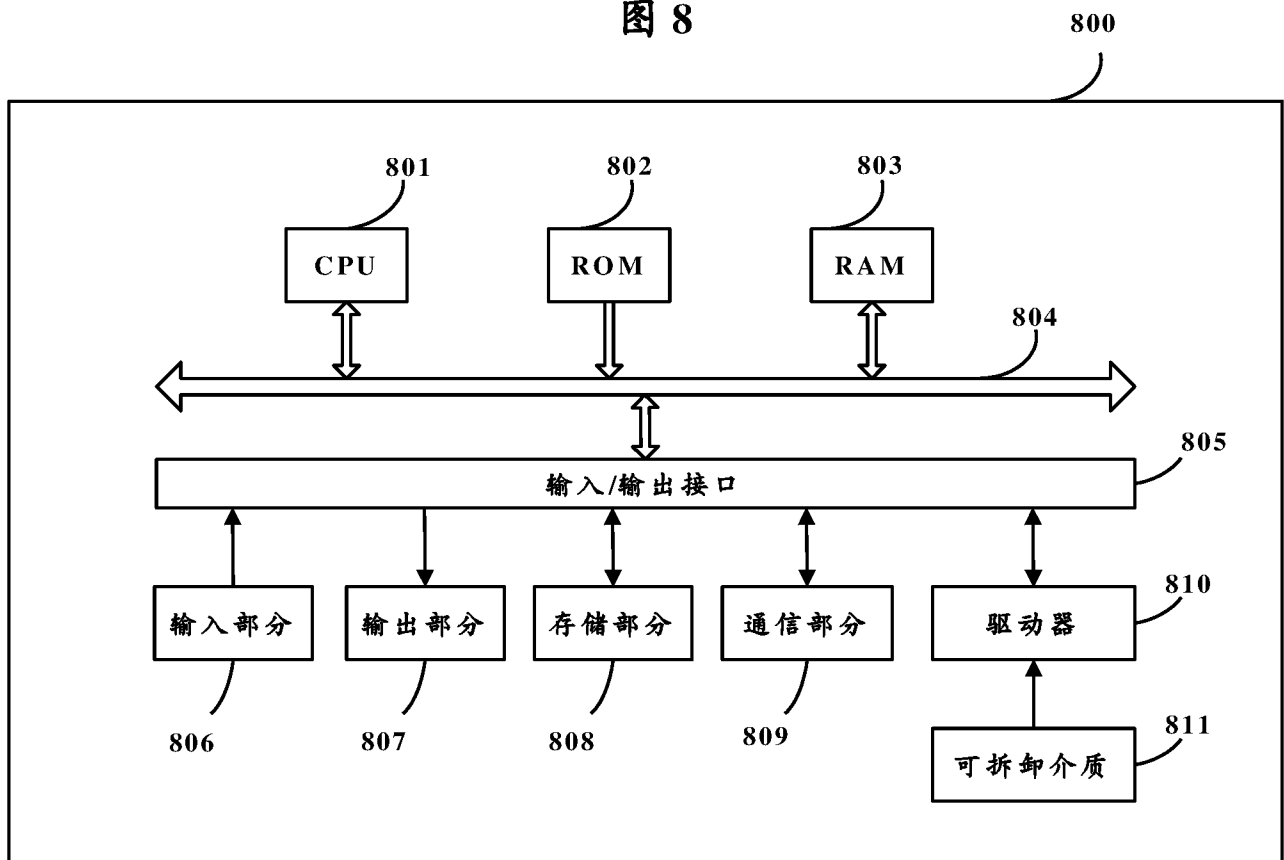


图 7

图 8



INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/077473

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27 (2006.01) i; G06F 17/21 (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, CNKI, EPODOC, WPI, IEEE: natural language, train, depth, word embedding, word feature, neural network, natural, language, text+, vector, model, merge, correlation, analysis, CCA, DCCA, deep, word, embedding, feature, Word2vec, GloVe, "C&W", phrase, sentence, corpus, auto, encoder, neural, network

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104881401 A (DALIAN UNIVERSITY OF TECHNOLOGY), 02 September 2015 (02.09.2015), description, paragraphs [0052]-[0072], and figure 1	1-23
A	CN 104657350 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES), 27 May 2015 (27.05.2015), the whole document	1-23
A	CN 104199809 A (JIANGSU UNIVERSITY), 10 December 2014 (10.12.2014), the whole document	1-23
A	CN 104915448 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES), 16 September 2015 (16.09.2015), the whole document	1-23
A	US 2004148154 A1 (ACERO, A. et al.), 29 July 2004 (29.07.2004), the whole document	1-23

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 23 May 2017 (23.05.2017)	Date of mailing of the international search report 20 June 2017 (20.06.2017)
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 XU, Shuxian Telephone No.: (86-10) 53318920

International application No.
PCT/CN2017/077473

Form PCT/ISA/210 (patent family annex) (July 2009)

国际检索报告

国际申请号

PCT/CN2017/077473

A. 主题的分类

G06F 17/27 (2006.01)i; G06F 17/21 (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, CNKI, EPDOC, WPI, IEEE: 自然语言, 文本, 向量, 模型, 融合, 相关, 分析, 训练, 深度, 词嵌入, 词特征, 词语, 短语, 句子, 语料, 自动, 编码器, 神经网络, natural, language, text+, vector, model, merge, correlation, analysis, CCA, DCCA, deep, word, embedding, feature, Word2vec, GloVe, "C&W", phrase, sentence, corpus, auto, encoder, neural, network

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	CN 104881401 A (大连理工大学) 2015年 9月 2日 (2015 - 09 - 02) 说明书第[0052]-[0072]段, 附图1	1-23
A	CN 104657350 A (中国科学院自动化研究所) 2015年 5月 27日 (2015 - 05 - 27) 全文	1-23
A	CN 104199809 A (江苏大学) 2014年 12月 10日 (2014 - 12 - 10) 全文	1-23
A	CN 104915448 A (中国科学院自动化研究所) 2015年 9月 16日 (2015 - 09 - 16) 全文	1-23
A	US 2004148154 A1 (ACERO, ALEJANDRO 等) 2004年 7月 29日 (2004 - 07 - 29) 全文	1-23

☐ 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

2017年 5月 23日

国际检索报告邮寄日期

2017年 6月 20日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局 (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

徐淑娴

电话号码 (86-10) 53318920

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/077473

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	104881401	A	2015年 9月 2日	无	
CN	104657350	A	2015年 5月 27日	无	
CN	104199809	A	2014年 12月 10日	无	
CN	104915448	A	2015年 9月 16日	无	
US	2004148154	A1	2004年 7月 29日	US 2004148170	A1 2004年 7月 29日

表 PCT/ISA/210 (同族专利附件) (2009年7月)