(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 108509539 B
(45)授权公告日 2021.08.17

(21)申请号 201810218121.1
(22)申请日 2018.03.16
(65)同一申请的已公布的文献号
申请公布号 CN 108509539 A
(43)申请公布日 2018.09.07
(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地西路6号
(72)发明人 杨帆 匡启帆 金宝宝 张成松
(74)专利代理机构 北京派特恩知识产权代理有限公司 11270
代理人 王姗姗 张颖玲
(51)Int.Cl.
G06F 16/36(2019.01)
G06F 40/289(2020.01)

(56)对比文件
CN 107798140 A,2018.03.13
CN 107679225 A,2018.02.09
CN 107368524 A,2017.11.21
CN 107368547 A,2017.11.21
CN 107766559 A,2018.03.06
CN 106844368 A,2017.06.13
US 2017323203 A1,2017.11.09

审查员 刘华桥

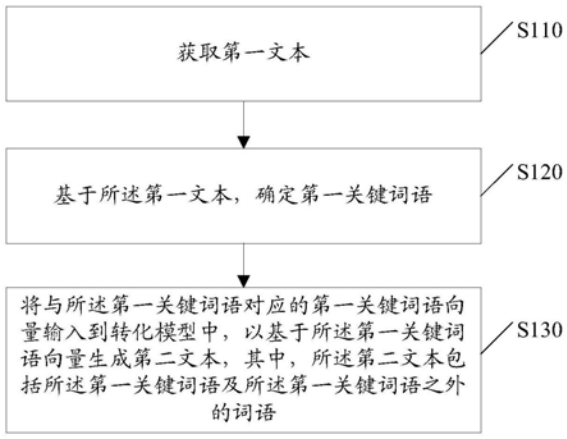
权利要求书2页 说明书11页 附图4页

(54)发明名称

信息处理方法电子设备

(57)摘要

本发明实施例公开了一种信息处理方法及电子设备。所述方法包括:获取第一文本;基于所述第一文本,确定第一关键词语;将与所述第一关键词语对应的第一关键词语向量输入到转化模型中,以基于所述第一关键词语向量生成第二文本,其中,所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语。



CN 108509539 B

1.一种信息处理方法,其特征在于,包括:

获取第一文本;

基于所述第一文本,确定第一关键词语;

将与所述第一关键词语对应的第一关键词语向量输入到转化模型中,以基于所述第一关键词语向量生成第二文本,其中,所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语;其中,所述转化模型为:在训练过程中基于第一损失值对所述转化模型的编码层进行约束并确定是否停止训练的自编码模型,所述第一损失值为:根据包括样本语句的关键词语的第二关键词语和所述样本语句经所述编码层处理得到的第三关键词语的匹配度确定的。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

构建训练语料库,其中,所述训练语料库能够至少用于表征第二关键词语与样本语句的对应关系;

基于所述训练语料库,训练预设模型得到所述转化模型。

3.根据权利要求2所述的方法,其特征在于,

所述构建训练语料库,包括:

对样本语句进行分词,获得与所述样本语句相对应的词序列;

对所述词序列进行编码,获得第一向量;

从所述词序列中提取出关键词语,构成与所述样本语句对应的第二关键词语。

4.根据权利要求3所述的方法,其特征在于,所述基于所述训练语料库,训练预设模型得到所述转化模型,包括:

将所述第一向量输入到预设模型的第一类处理层,得到第三关键词语;其中,所述第一类处理层为所述编码层;

基于所述第二关键词语及所述第三关键词语,确定所述第一损失值;

若所述第一损失值不满足第一预设条件,继续训练所述预设模型。

5.根据权利要求4所述的方法,其特征在于,

所述基于所述训练语料库,训练预设模型得到所述转化模型,还包括:

将所述第三关键词语输出到所述预设模型的第二类处理层,得到第二向量;

基于所述第二向量和与所述第一向量对应的第三向量,确定第二损失值;

所述方法还包括:

若所述第一损失值满足第一预设条件且所述第二损失值满足第二预设条件,确定已成功将所述预设模型训练成所述转化模型。

6.根据权利要求3所述的方法,其特征在于,

所述构建训练语料库,还包括:

对所述词序列中词语进行第一类编码得到与词语含义对应的第四向量;

对所述第四向量进行向量转换得到与所述词序列的上下文内容对应的所述第一向量,获得所述词语和所述第一向量的映射关系。

7.根据权利要求6所述的方法,其特征在于,

所述将与所述第一关键词语对应的第一关键词语向量输入到转化模型中,以基于所述第一关键词语向量生成第二文本,还包括:

基于所述映射关系，将所述第一关键词语转化为所述第一关键词语向量；

将所述第一关键词语向量输入到所述转化模型的第二类处理层，得到第五向量；

基于所述映射关系，获得与所述第五向量对应的第二文本。

8.一种电子设备，其特征在于，包括：

获取模块，用于获取第一文本；

确定模块，用于基于所述第一文本，确定第一关键词语；

生成模块，用于将与所述第一关键词语对应的第一关键词语向量输入到转化模型中，以基于所述第一关键词语向量生成第二文本，其中，所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语；其中，所述转化模型为：在训练过程中基于第一损失值对所述转化模型的编码层进行约束并确定是否停止训练的自编码模型，所述第一损失值为：根据包括样本语句的关键词语的第二关键词语和所述样本语句经所述编码层处理得到的第三关键词语的匹配度确定的。

9.根据权利要求8所述的电子设备，其特征在于，所述电子设备还包括：

构建模块，用于构建训练语料库，其中，所述训练语料库能够至少用于表征第二关键词语与样本语句的对应关系；

训练模块，用于基于所述训练语料库，训练预设模型得到所述转化模型。

10.根据权利要求9所述的电子设备，其特征在于，

所述构建模块，具体用于对样本语句进行分词，获得与所述样本语句相对应的词序列；对所述词序列进行编码，获得第一向量；从所述词序列中提取出关键词语，构成与所述样本语句对应的第二关键词语。

信息处理方法电子设备

技术领域

[0001] 本发明涉及信息技术领域，尤其涉及一种信息处理方法及电子设备。

背景技术

[0002] 随着技术的发展，现有技术中已经有了智能问答系统。智能问答系统可以通过语音或文本接收等方式获得一个问题；基于该问题，智能系统可以通过搜索等操作，提供该问题的答案。现有的智能问答系统一般关注问题回答的准确性，但是有时候智能问答系统提供的答案文本或答案语音并非按照自然语音的方式提供，例如，答案文本是由几个孤立的词或词语组成的，一方面不符合用户的理解，导致用户理解困难；另一方面，即便用户理解也会导致用户使用体验感受差。

发明内容

[0003] 有鉴于此，本发明实施例期望提供一种信息处理方法及电子设备，至少部分解决上述问题。

[0004] 为达到上述目的，本发明的技术方案是这样实现的：第一方面，本发明实施例提供一种信息处理方法，包括：

[0005] 获取第一文本；

[0006] 基于所述第一文本，确定第一关键词语；

[0007] 将与所述第一关键词语对应的第一关键词语向量输入到转化模型中，以基于所述第一关键词语向量生成第二文本，其中，所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语。

[0008] 可选地，所述方法还包括：

[0009] 构建训练语料库，其中，所述训练语料库能够至少用于表征第二关键词语与样本语句的对应关系；

[0010] 基于所述训练语料库，训练预设模型得到所述转化模型。

[0011] 可选地，所述构建训练语料库，包括：

[0012] 对样本语句进行分词，获得与所述样本语句相对应的词序列；

[0013] 对所述词序列进行编码，获得第一向量；

[0014] 从所述词序列中提取出关键词语，构成与所述样本语句对应的第二关键词语。

[0015] 可选地，所述基于所述训练语料库，训练预设模型得到所述转化模型，包括：

[0016] 将所述第一向量输入到预设模型的第一类处理层，得到第三关键词语；

[0017] 基于所述第二关键词语及所述第三关键词语，确定第一损失值；

[0018] 若所述第一损失值不满足第一预设条件，继续训练所述预设模型。

[0019] 可选地，所述基于所述训练语料库，训练预设模型得到所述转化模型，还包括：

[0020] 将所述第三关键词语输出到所述预设模型的第二类处理层，得到第二向量；

[0021] 基于所述第二向量和与所述第一向量对应的第三向量，确定第二损失值；

[0022] 所述方法还包括：

[0023] 若所述第一损失值满足第一预设条件且所述第二损失值满足第二预设条件，确定已成功将所述预设模型训练成所述转化模型。

[0024] 可选地，所述构建训练语料库，还包括：

[0025] 对所述词序列中词语进行第一类编码得到与词语含义对应的第四向量；

[0026] 对所述第四向量进行向量转换得到与所述词序列的上下文内容对应的所述第一向量，获得所述词语和所述第一向量的映射关系。

[0027] 可选地，所述将与所述第一关键词语对应的第一关键词语向量输入到转化模型中，以基于所述第一关键词语向量生成第二文本，还包括：

[0028] 基于所述映射关系，将所述第一关键词语转化为所述第一关键词语向量；

[0029] 将所述第一关键词语向量输入到所述转化模型的第二类处理层，得到第五向量；

[0030] 基于所述映射关系，获得与所述第五向量对应的第二文本。

[0031] 第二方面，本发明实施例提供一种电子设备，包括：

[0032] 获取模块，用于获取第一文本；

[0033] 确定模块，用于基于所述第一文本，确定第一关键词语；

[0034] 生成模块，用于将与所述第一关键词语对应的第一关键词语向量输入到转化模型中，以基于所述第一关键词语向量生成第二文本，其中，所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语。

[0035] 可选地，所述电子设备还包括：

[0036] 构建模块，用于构建训练语料库，其中，所述训练语料库能够至少用于表征第二关键词语与样本语句的对应关系；

[0037] 训练模块，用于基于所述训练语料库，训练预设模型得到所述转化模型。

[0038] 可选地，所述构建模块，具体用于对样本语句进行分词，获得与所述样本语句相对应的词序列；对所述词序列进行编码，获得第一向量；从所述词序列中提取出关键词语，构成与所述样本语句对应的第二关键词语。

[0039] 本发明实施例信息处理方法及电子设备，在基于第一文本获得第一关键词语之后，不会直接输出第一关键词语，而是会将第一关键词语对应的第一关键词向量输入到转化模型中，从而通过转换模型输出包含第一关键词语及第一关键词语以外的第二文本；电子设备在显示第二文本或者语音播报第二文本。如此形成的第二文本是符合语法规则的自然语句，相对于突兀的输出几个关键词，可以降低理解难度，使得用户更好的理解基于第一文本提供的信息，从而提升用户体验及设备只能性；与此同时，减少理解错误的现象。

附图说明

[0040] 图1为本发明实施例提供的第一种信息处理方法的流程示意图；

[0041] 图2为本发明实施例提供的第二种信息处理方法的流程示意图；

[0042] 图3为本发明实施例提供的一种电子设备的结构示意图；

[0043] 图4为本发明实施例提供的第三种信息处理方法的流程示意图；

[0044] 图5为本发明实施例提供的第四种信息处理方法的流程示意图；

[0045] 图6为本发明实施例提供的模型训练获得转化模型的流程示意图；

[0046] 图7为本发明实施例利用转化模型输出为自然语句的文本的流程示意图。

具体实施方式

[0047] 以下结合说明书附图及具体实施例对本发明的技术方案做进一步的详细阐述。如图1所示，本实施例提供一种信息处理方法，包括：

[0048] 步骤S110：获取第一文本；

[0049] 步骤S120：基于所述第一文本，确定第一关键词语；

[0050] 步骤S130：将与所述第一关键词语对应的第一关键词语向量输入到转化模型中，以基于所述第一关键词语向量生成第二文本，其中，所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语。

[0051] 本实施例提供给的信息处理方法可以应用于各种电子设备中，例如，运行有所述转化模型的手机、平板电脑、可穿戴式设备或服务器等电子设备。

[0052] 所述步骤S110可包括以下至少之一：

[0053] 接收语音指令，识别所述语音指令，获得与所述语音指令对应的第一文本；

[0054] 接收用户输入的第一文本；

[0055] 接收其他电子设备提供的第一文本。

[0056] 所述步骤S120可包括：基于第一文本获得与第一文本的文本含义相适配的第一关键词语。例如，第一文本为一个智能问答系统中提出的问题，所述第一关键词语为回答所述第一文本所提出问题的答案。所述第一关键词语的个数可为一个或多个。

[0057] 所述第一关键词语可包括：在自然语句中充当主语的主体(Subject)，可以充当谓语的谓词(Predicate)及充当宾语的宾体(Object)。谓词、宾体(客体)是基于图数据库来定义的。在图数据库中有两种基本的元素，节点和边，主体和宾体对应于图数据库中图的节点，而谓词充当图的边，在图上可以想象成主体和宾体两个节点通过谓词这条边连接起来，并且这条边的方向是由主体节点指向宾体节点。

[0058] 本实施例中的第一关键词语可为前述的主体、谓词及宾体中的一个或多个。

[0059] 在步骤S130中会将第一关键词语对应的第一关键词向量输入到转换模型中，得到所述第二文本。在本实施例中可以利用预设编码方式对所述第一关键词语进行编码，可以得到所述第一关键词向量。例如，电子设备中或网络中存储有编码所述第一关键词语的码本，通过码本的查询，将第一关键词转换为所述第一关键词语。在本实施例中所述第一关键词向量可为一个仅包括单行向量，也可以是包括多行的多行向量，在本实施例中所述多行向量又可以称之为矩阵。

[0060] 在本实施例中利用该转换模型得到的第二文本，包括所述第一关键词语，同时还包括所述第一关键词语以外的其他词语，例如，包括连接两个或两个以上的第一关键词语的连接词语。总之，本实施例提供的第二文本可为满足预设语法规则的自然语句，该自然语句不仅使用到了第一关键词语，同时还包括第一关键词语以外的其他词语；相当于现有技术中直接输出第一关键词语作为第一文本的输出，显然提升了用户的使用感受，同时也避免了突兀的输出第一关键词语导致的用户理解困难或理解错误的问题。

[0061] 例如，若第一文本为：感冒有哪些症状；电子设备通过答案搜索等智能处理得到了若干个关键词词“发烧”、“咳嗽”及“流鼻涕”等。若不经过所述转换模型的处理，直接就可能

输出上述关键词，但是此时若用户同时输入了多个问题时，就会导致用户不知道电子设备突兀输出这几个关键词对一个了哪一个问题。若通过步骤S120及步骤S130的处理，则可能会输出“发烧、咳嗽及流鼻涕是感冒的症状”的自然语句，从而方便用户理解，提升系统的智能化及用户使用体验。

[0062] 可选地，所述方法还包括：

[0063] 构建训练语料库，其中，所述训练语料库能够至少用于表征第二关键词语与样本语句的对应关系；

[0064] 基于所述训练语料库，训练预设模型得到所述转化模型。

[0065] 在本实施例中第二关键词语可为从样本语句中提取出的关键词语，提取第二关键词语的方式可包括：提取出该语句的主体、谓词及宾体等词句；例如，通过绘制有向图来实现，或者，按照语法规则进行拆分等。

[0066] 在提取出所述第二关键词之后，将其与样本语句对应起来，建立所述对应关系；从而形成了一个包括大量对应干系的训练语料库。

[0067] 利用训练语料库中的对应关系，徐连预设模型。该模型可为基于二叉树、多叉树或者回归模型等各种学习模型，例如，向量机学习模型或神经网络学习模型等；通过训练可以得到各个网络参数空白的学习模型的网络参数值，从而形成所述转化模型。

[0068] 进一步地，所述构建训练语料库，包括：

[0069] 步骤S101：对样本语句进行分词，获得与所述样本语句相对应的词序列；

[0070] 步骤S102：对所述词序列进行编码，获得第一向量；

[0071] 步骤S103：从所述词序列中提取出关键词语，构成与所述样本语句对应的第二关键词语。

[0072] 在本实施例中利用分词算法进行分词的处理，得与样本语句对应的词序列。所述分词算法可包括以下任意一种：

[0073] 基于词典的分词算法，该词典可包括：字典或词库，通过基于字典、词库匹配的分词方法；(字符串匹配、机械分词法)；

[0074] 基于统计的分词算法：基于词频度统计的分词算法；

[0075] 基于规则的分词算法：基于知识理解的分词算法，例如，基于语法规则，再例如基于自然语句的上下文内容的分词算法。

[0076] 通过分词，可以得到组成所述自然语句的词序列。该词序列可以包括一个或多个按照其在自然语句中排序进行排列的多个词语。

[0077] 对词序列进行编码，得到第一向量。该第一向量不仅可以反映词序列中各个词语的含义，同时还可以反映词语之间的关联性或上下文内容。

[0078] 通过编码将文本形式或字符串形式的词向量，转换为了可以用于计算等处理的向量，该向量中可包括多个元素，这些元素的取值可为任意数值。

[0079] 与此同时，还会从词序列中进一步提取出关键词。例如，一个自然语句“小丽和小红是中国人”，通过分词之后得到词序列为：“小丽”、“和”、“小红”“是”、“中国人”，再从分词得到的词序列中选取关键词，可以得到第二关键词语“小丽”、“小红”“是”、“中国人”，一般情况下，第二关键词语包括的词语个数，不大于词序列包括的词语个数。

[0080] 在一些实施例中，所述基于所述训练语料库，训练预设模型得到所述转化模型，包

括：

[0081] 将所述第一向量输入到预设模型的第一类处理层，得到第三关键词语；

[0082] 基于所述第二关键词语及所述第三关键词语，确定第一损失值；

[0083] 若所述第一损失值不满足第一预设条件，继续训练所述预设模型。

[0084] 在本实施例中，将第一向量输入到第一类处理层，该第一类处理层可为编码层，使得数值化的第一向量，可以形成文本形式或字符串形式的第三关键词语。在本实施例中还会比对第三关键词语计算损失值。例如，比对所述第二关键词语和第三关键词语，根据两者的匹配度确定所述第一损失值。例如，预先设置有转换模型训练的第一损失函数，将所述第二关键词语和第三关键词语作为第一损失函数的输入量，通过计算自然得到所述第一损失值。将第一损失值与第一阈值进行比较，若大于第一阈值可认为需要继续训练预设模型。

[0085] 可选地，第一类处理层可为编码层，可以将数值化的第一向量转换为文本或字符串形式的第三关键词。所述基于所述训练语料库，训练预设模型得到所述转化模型，还包括：

[0086] 将所述第三关键词语输出到所述预设模型的第二类处理层，得到第二向量；

[0087] 基于所述第二向量和与所述第一向量对应的第三向量，确定第二损失值；

[0088] 所述方法还包括：

[0089] 若所述第一损失值满足第一预设条件且所述第二损失值满足第二预设条件，确定已成功将所述预设模型训练成所述转化模型。

[0090] 在本实施例中，将所述第三关键词语通过第二类处理层的处理，得到第二向量。再将第二向量与第三向量继续比对，可以确定出将第一向量经过第一类处理层处理之后产生的误差。第三向量可为所述第一向量的初始向量。所述第二类处理层可为解码层，例如，将第一向量通过逆向解码得到第二向量。第三向量可为第一向量编码之前的原始向量。

[0091] 若第三向量与第二向量满足预设匹配度，则第二损失值自然就小。在还有一些实施例中还可以设置第二损失函数，将所述第三向量和第二向量作为第二损失函数的输入，通过第二损失函数的计算，会得到第二损失值。

[0092] 在本实施例中为了确保转换模型的可信度，会在第一损失值满足第一预设条件及第二损失值满足第二预设条件时，才停止预设模型的训练，认为当前已经成功完成了模型训练。

[0093] 在一些实施例中，可以在第一损失值满足第一预设条件时，就可以停止模型的训练。

[0094] 此处，第二损失值满足所述第二预设条件可包括：若第二损失值小于第二阈值，可认为第二损失值满足所述第二预设条件。

[0095] 可选地，所述构建训练语料库，还包括：

[0096] 对所述词序列中词语进行第一类编码得到与词语含义对应的第四向量；

[0097] 对所述第四向量进行向量转换得到与所述词序列的上下文内容对应的所述第一向量，获得所述词语和所述第一向量的映射关系。

[0098] 在本实施例中可以对词序列进行第一类表面，得到第四向量。在对第四向量进行转换得到与所述第四向量对应的第一向量。此处，第四向量可与前述的第二向量进行对应。所述第一类编码可为各种了类型的文本或字符串到向量的转换方式。

[0099] 将第四向量进行向量转换时，根据不同词语之间的关联关系(上下文内容个)的一种，以向不同词语对应的向量之间的距离来表征所述上下文内容，从而形成第四向量。

[0100] 直接构建了词语与第一向量的映射关系。在一些实施例中，所述映射关系可包括：第一映射关系表和第二映射关系表，根据第一映射关系表，可以将词语转换为第四向量，在基于第二映射关系表，可以将第四向量转换为所述第一向量。

[0101] 在一些实施例中，所述映射关系可仅包括一张可以直接将第二关键词语转换为所述第一向量的映射关系表。

[0102] 所述映射关系可以是映射关系表，还可以是映射函数。

[0103] 可选地，所述步骤S110可包括：

[0104] 基于所述映射关系，将所述第一关键词语转化为所述第一关键词语向量；

[0105] 将所述第一关键词语向量输入到所述转化模型的第二类处理层，得到第五向量；

[0106] 基于所述映射关系，获得与所述第五向量对应的第二文本。

[0107] 在本实施例中，根据所述映射关系可以将步骤S120中得到的第一关键词语直接转换为可供转化模型处理的第一关键词语向量，再输入到转化模型的第二类处理层之后，就可以得到第五向量，并直接将第五向量进行解码处理，就可以得到同时包括第一关键词语及第一关键词语以外的其他词语的第二文本，而采用这种模式得到的第二文本是自然语句，是满足用户理解习惯的自然语句。

[0108] 在一些实施例中，所述方法还包括：

[0109] 确定所述第一文本对应的应答模式；

[0110] 若第一文本对应的应答模式为第一模式，则执行所述步骤S120至步骤S130，以使电子设备最终输出的是为自然语句的第二文本。

[0111] 在还有一些实施例中，所述方法还包括：

[0112] 若所述第一文本对应的应答模式为第二模式，则直接输出所述第一关键词语。

[0113] 若所述第一文本所提的问题是封闭式问题，例如，封闭式问题中的判断问题或选择问题；则可认为第一文本对应的应答模式为第一模式。针对判断问题，电子设备可以简单回答“是”或“否”，从而无需省略步骤S120至步骤S130的操作。选择问题，在第一文本中已经提供了答案，电子设备也可以简单的输出选择的答案即可，也可以省略执行步骤S120至步骤S130，减少电子设备的处理负荷，降低功耗。

[0114] 若第一文本所提到的问题为：开放式问题，电子设备则可能需要组织语言来进行答复，若还是机械的输出第一关键词语，就会导致误解，故此时执行所述步骤S120至步骤S130。例如，第一文本所提到问题是“请描述猫的习惯和喜好”。

[0115] 在还有一些实施例中，还可以根据电子设备的设置，雀定所述第一文本的应答模式，例如，有的用户理解能力强，不喜欢看大量的文字或听大段的语音，则将应答模式设置为第二模式，否则可以采用电子设备默认模式，该默认模式可为第一模式，也可以是第二模式。

[0116] 在还有一些实施例中，根据当前第一文本的数目确定所述应答模式，例如，若在输出回答之前，电子设备当前待提供答案的第一文本就一个，则可以选择所述第二模式，否则采用第一模式，避免在有多个第一文本在电子设备的显示屏上输出而答案距离问题较远导致的用户混淆和难以理解的问题。

[0117] 如图3所示，本实施例提供一种电子设备，包括：

[0118] 获取模块110，用于获取第一文本；

[0119] 确定模块120，用于基于所述第一文本，确定第一关键词语；

[0120] 生成模块130，用于将与所述第一关键词语对应的第一关键词语向量输入到转化模型中，以基于所述第一关键词语向量生成第二文本，其中，所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语。

[0121] 该电子设备包括或运行有这些模块，这些模块可均为程序模块，能够被处理器等执行后，实现第一文本的获取、第一关键词语的获得，及第二文本的生成等操作。

[0122] 可选地，所述电子设备还包括：

[0123] 构建模块，用于构建训练语料库，其中，所述训练语料库能够至少用于表征第二关键词语与样本语句的对应关系；

[0124] 训练模块，用于基于所述训练语料库，训练预设模型得到所述转化模型。

[0125] 在本实施例中，该电子设备还包括构建模块，该构建模块可以构建出训练语料库，训练模型直接可以用于训练处所述转化模型。

[0126] 可选地，所述构建模块，具体用于对样本语句进行分词，获得与所述样本语句相对应的词序列；对所述词序列进行编码，获得第一向量；从所述词序列中提取出关键词语，构成与所述样本语句对应的第二关键词语。

[0127] 可选地，所述训练模块，可用于将所述第一向量输入到预设模型的第一类处理层，得到第三关键词语；

[0128] 基于所述第二关键词语及所述第三关键词语，确定第一损失值；

[0129] 若所述第一损失值不满足第一预设条件，继续训练所述预设模型。

[0130] 可选地，所述训练模块，还可用于将所述第三关键词语输出到所述预设模型的第二类处理层，得到第二向量；基于所述第二向量和与所述第一向量对应的第三向量，确定第二损失值；

[0131] 所述电子设备还包括：

[0132] 判定模块，用于若所述第一损失值满足第一预设条件且所述第二损失值满足第二预设条件，确定已成功将所述预设模型训练成所述转化模型。

[0133] 可选地，所述构建模块，还用于对所述词序列中词语进行第一类编码得到与词语含义对应的第四向量；对所述第四向量进行向量转换得到与所述词序列的上下文内容对应的所述第一向量，获得所述词语和所述第一向量的映射关系。

[0134] 可选地，所述生成模块130，具体用于基于所述映射关系，将所述第一关键词语转化为所述第一关键词语向量；将所述第一关键词语向量输入到所述转化模型的第二类处理层，得到第五向量；基于所述映射关系，获得与所述第五向量对应的第二文本。

[0135] 以下结合上述任意一个实施例提供几个具体示例：

[0136] 示例1：

[0137] 如图4所示，本示例提出一种基于SPO约束自编码神经网络文本生成的方法和装置，主要包括以下步骤：

[0138] 步骤1：训练语料库准备

[0139] 步骤1.1：收集文本语料，该文本语料可为前述的样本语句，可根据应用场景，分应

用场景的收集对应的文本语料。例如，对于医药行业的场景，可以在一些医疗问答网站上，爬取相关问答文本，作为文本语料。在本示例中收集的文本预料可为满足语法规则的一个或多个自然语句组成。

[0140] 步骤1.2：对文本语料进行预处理，该文本语料的预处理可包括：

[0141] 将文本语料拆分成一各个独立的自然语句；

[0142] 对自然语句进行预处理，主要包括：对自然语句进行分词操作、对词语进行独热(one-hot)编码，基于one-hot编码形成词向量，提取自然语句的主体(Subject)、谓词(Predicate)、宾体(Object)(后面简称为SPO提取)。

[0143] 步骤2：模型训练

[0144] 本示例提出一种带约束的自编码神经网络模型，在模型的自编码层映入SPO的词向量约束信息，然后对于one-hot编码后的自然语句进行自编码学习，从而使得训练之后的自编码神经网络模型，具有能够将一个SPO三元组转化成一个包括该SPO三元组中的主体、谓词及宾体的自然语句。

[0145] 步骤3：文本生成，该自动生成的文本是由按照自然语句的语法规则形成的。

[0146] 对于给定SPO词向量，直接作为自编码训练模型的解码(decode)层的输入，通过decode层映射生成最终包含SPO信息的文本。该SPO信息为所述SPO三元组及所述SPO词向量所表达的含义。

[0147] 本示例提供的信息处理方法，从准备训练语料到模型训练到模型预测的整个流程，基本上无需人工干预，尤其是在训练语料准备阶段，无需传统监督学习的大量人工标注，方便利用大规模数据进行训练。

[0148] 本示例提供的自编码神经网络模型，通过训练自编码模型，便可以得到一个文本生成器；在已知自然语句的主干(SPO)的情况下，生成人友好的自然语言。

[0149] 示例2：

[0150] 基于示例1，本示例提供一种训练语料库准备进行进一步的详细描述，可包括：

[0151] 在步骤1.2中对自然语句进行预处理，如图5所示，可包括：

[0152] 语料收集；

[0153] 句子拆分，将大段的语料拆分成自然语句；

[0154] 自然语句的分词，

[0155] 词语的编码，例如，采用one-hot编码，将one-hot编码形成的向量，通过向量转换，转换成与前述第一向量对应的词向量；

[0156] SPO的提取。

[0157] 所述自然语句的分词，可包括：

[0158] 通过统计学习或者字典匹配等分词算法对每个自然语句进行分词，得到每条自然语句对应的词序列，然后对所有词序列中的词语进行去重汇总，得到词语语料集合C。在本示例中，去重汇总是指：将多条自然语句转化得到的词序列中各个词语进行相同词语的合并，从而使得多条自然语句得到的词序列中各重复词语去除，如此，得到的词语语料集合C就不会包括重复的词语。

[0159] 对于自然语句“发烧是感冒的症状”，通过对自然语句的分词可得到[“发烧”，“是”，“感冒”，“的”，“症状”]的词序列，对于自然语句“咳嗽是感冒的症状”，通过自然语句

的分词可得到[“咳嗽”,“是”,“感冒”,“的”,“症状”]。针对这两个自然语句的去重汇总,则会在词语语料集合C中得到6个词语,分别是“发烧”,“是”,“感冒”,“的”,“症状”及“咳嗽”。

[0160] 在本示例中,词序列是有序排列的多个词语,例如,[“发烧”,“是”,“感冒”,“的”,“症状”]及[“感冒”,“是”,“发烧”,“的”,“症状”]是两个不同的词序列。不同的词序列表面了自然语句中不通的上下文语义。

[0161] 典型的所述统计学习可包括各种类型的机器学习,例如,基于神经网络的机器学习,基于向量机的机器学习。该统计学习可以提供从一个自然语句中提取出词序列的模型。

[0162] 所述字典匹配可包括:可供自然语句分割的字典,该字典中包括各种词语,将自然语句与字典中的词语进行匹配,若一个自然语句有多种拆分方式得到多种词序列时,可以给予前序最大匹配算法,或者,最大概率拆分算法等,选择一种自然语句的拆分方式。例如,以自然语句“北京大学生”,字典中可包括:“北京”、“大学生”、“北京大学生”这几个词条,显然对该自然语句进行拆分,可以得到两种词序列[“北京”,“大学生”]及[“北京大学生”]。若按照前序最大匹配算法,则会选择[北京大学生]作为本自然语句分词得到的词序列。前序最大匹配算法为:在自然语句中靠前的部分按照包含的字数或字符最大进行拆分。若统计信息表明该自然语句分词为“北京大学生”的正确概率或者拆分概率更高一些,则该自然语句可以分词得到的词序列为:[“北京大学生”]。

[0163] 所述词语编码可包括:

[0164] one-hot编码;

[0165] 词向量构建。

[0166] 基于词语语料集合C,对词语进行one-hot编码得到编码向量(对应于前述第四向量),并构建词语到one-hot编码的映射表D(即前述第一映射关系的一种)

[0167] 基于词语one-hot编码和词序列所提供的词语上下文语义信息,通过神经网络语言模型等统计学习算法,生成得到每个词语对应的词向量(对应于前述第一向量),由此构建词语到词向量的映射表P(第二映射关系的一种)。

[0168] 在进行one-hot编码时,词语语料集合C中共得到了N个不同的词语,则为每一个词语构建一个N维的向量,一个N维的向量将包括N个元素。一个词语对应的向量中,仅有一个元素的取值为第一取值,剩余元素的取值为第二取值,不同词语的向量中第一取值的元素的位置不同。例如,所述第一取值为“0”,则所述第二取值可为任意非“0”自然数;在本示例中为了方便编码,第二取值可为“1”;若,第一取值为“1”,则第二取值可为“0”等。例如,若词语语料集合C共有10000条不同的词语,那么通过one-hot编码,每个词语对应10000维的向量(向量中只有一个元素的取值为“1”,其余取值均为0),例如:

[0169] “发烧”:[1,0,0,0,0,..]

[0170] “感冒”:[0,1,0,0,0,..]

[0171] “症状”:[0,0,1,0,0,..]。

[0172] 由上述可知,不同词语中取值为“1”在向量中的位置不同。

[0173] 为了能够在语义上区分不同词语之间的相似性,进一步通过one-hot编码,通过统计学习模型(例如,word2vec)得到了每个词语对应的词向量模型,通常词向量的维度为100～1000维,相比one-hot编码向量有明显的降维,同时不同词语在语意上的相似性可以直接通过欧氏距离来表征。例如,通过one-hot编码生成100维的词向量,其中每个词向量的元素

都是实数。

[0174] 例如，若采用one-hot编码则词语“父亲”、“爸爸”、“母亲”之间的欧式距离必然为1；但是若通过统计学习模型的语义相似性处理，得到上述三个词语的之间的欧式距离会使得“父亲”和“爸爸”对应的两个向量之间的欧式距离，小于“爸爸”、“母亲”对应的两个向量之间的欧式距离。

[0175] 所述SPO提取可包括：

[0176] 基于每条自然语句的词序列，通过基于规则的句法分析或者基于标注的统计学习，从其中抽取出分别表示主体，谓语，宾体的词语，构成SPO三元组。

[0177] 例如，对于自然语句是“发烧是感冒的症状”，SPO分别对应“发烧”，“感冒”，“症状”的三元组；若现有技术中可能直接输出“发烧”，“感冒”，“症状”，或者，仅输出“发烧”。但是若采用本示例的方法，则会根据该SPO三元组，自动补齐这三个词语之间缺失的部分，使之形成有符合语法习惯的自然语句“发烧是感冒的症状”或“发烧为感冒的症状”等用户容易理解的文本。

[0178] 模型训练主要包含以下几个步骤：

[0179] 从用于模型训练的自然语句，提取出词序列和SPO(具体提取方式见训练语料准备)；

[0180] 将词序列和SPO三元组中的词语分别通过映射表D(可对应于前述的第一映射表)和映射表P(可对应于前述第二映射表)分别进行编码转换，这样词序列转换成了输入one-hot列表(一个二维矩阵，每行依次对应一个词向量，矩阵的行数等于词序列中词语的个数，矩阵的列数等于one-hot向量的维度)，SPO三元组转换成了SPO词向量(一个一维向量，将SPO对应的3个词向量依次拼接，向量维度等于3*词向量的维度)。

[0181] 将训练自然语句得到的one-hot列表作为自编码的输入，通过如图6中所示的编码(encode)层1(对应于前述第一类处理层)的编码操作，得到一个一维的编码层向量，编码层向量的维度设置为3*词向量的维度(和SPO词向量维度保持一致)，然后通过计算编码层向量和SPO词向量的距离损失(例如欧氏距离)，作为自编码模型的第一损失值loss1。若将编码层1的输出，通过编码层2之后得到词向量，该词向量可以提取SPO词向量(对应于前述第一向量)。SPO词向量通过解码层之后，得到一个向量，通过查询one-hot列表，可以得到一个生成第二文本的向量，从而可以计算出loss2。

[0182] 因为引入了loss1，对编码层进行了约束，使得编码层的编码形式变得可控，从而可以通过控制编码层的输入，对后续的文本生成进行有效干预。这也是本示例对于通常自编码模型主要改进(图6)，称这种改进方案为约束自编码。注意，在图6中，尽管只是展示了两个编码层，但encode层内部可以包含多个隐藏层，具体而言，可以通过递归神经网络或者卷积神经网络等方式来实现。

[0183] 将编码层向量通过decode层的解码操作，生成一个one-hot列表(一个二维自然语句，每行依次对应一个词语的one-hot编码向量)，然后通过对输出的one-hot列表和输入的one-hot列表进行比较(例如计算欧氏距离)，作为自编码模型的第二损失值loss2。注意，在图6中，尽管只是展示了一个decode层(对应于前述第二类处理层)，但decode层内部可以包含多个隐藏层，具体而言，可以通过递归神经网络等方式来实现。

[0184] 四，将loss1和loss2作为自编码模型的优化目标函数，通过训练数据对模型参数

进行优化(例如,采用随机梯度下降方法)。

[0185] 如图7所示,文本生成主要包含以下几个步骤:

[0186] 对于指定的SPO三元组,通过映射表P转化为SPO词向量(具体形式和模型训练中一致),将SPO词向量作为编码层的输入。关于SPO三元组的获取,可以根据具体的智能问答系统来生成,例如,智能医疗问答系统提取出(“感冒”,“症状”,“发烧”)。

[0187] 通过解码层的解码操作,输出one-hot列表(具体形式和模型训练中一致)。

[0188] 最后将输出one-hot列表中,每行的one-hot向量,替换为对应的词语,生成输出词序列,也就是生成的最终文本。例如,将SPO三元组(“感冒”,“症状”,“发烧”)通过转换生成词序列(“发烧”,“是”,“感冒”,“的”,“症状”),最后生成文本:“发烧是感冒的症状”。

[0189] 本发明实施例还提供一种计算机存储介质,该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被处理器执行后,能够实现前述一个或多个技术方案提供的信息处理方法。

[0190] 所述计算机存储介质可为:移动存储设备、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质等存储介质,可选为非瞬间存储介质。

[0191] 在本申请所提供的几个实施例中,应该理解到,所揭露的设备和方法,可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,如:多个单元或组件可以结合,或可以集成到另一个系统,或一些特征可以忽略,或不执行。另外,所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口,设备或单元的间接耦合或通信连接,可以是电性的、机械的或其它形式的。

[0192] 上述作为分离部件说明的单元可以是、或也可以不是物理上分开的,作为单元显示的部件可以是、或也可以不是物理单元,即可以位于一个地方,也可以分布到多个网络单元上;可以根据实际的需要选择其中的部分或全部单元来实现本实施例方案的目的。

[0193] 另外,在本发明各实施例中的各功能单元可以全部集成在一个处理模块中,也可以是各单元分别单独作为一个单元,也可以两个或两个以上单元集成在一个单元中;上述集成的单元既可以采用硬件的形式实现,也可以采用硬件加软件功能单元的形式实现。

[0194] 本领域普通技术人员可以理解:实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成,前述的程序可以存储于一计算机可读取存储介质中,该程序在执行时,执行包括上述方法实施例的步骤。

[0195] 以上所述,仅为本发明的具体实施方式,但本发明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发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应以所述权利要求的保护范围为准。

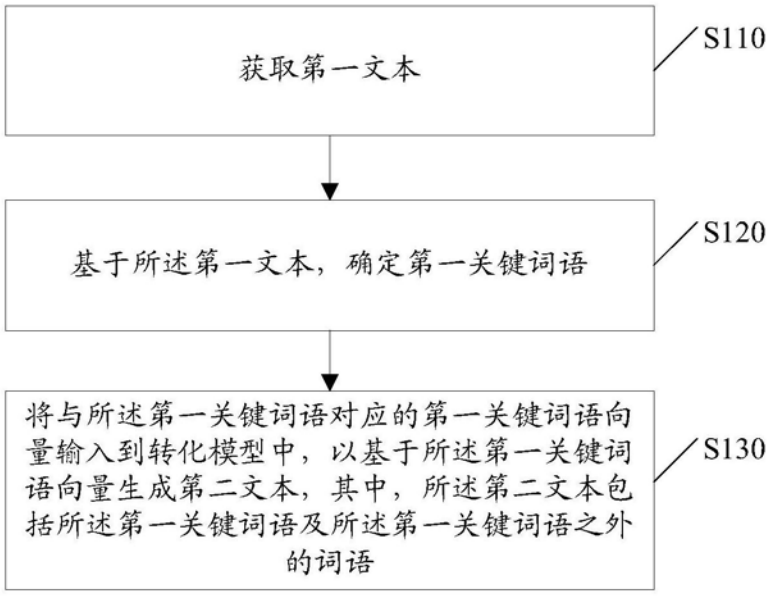

获取第一文本
S110
基于所述第一文本，确定第一关键词语
S120
将与所述第一关键词语对应的第一关键词语向量输入到转化模型中，以基于所述第一关键词语向量生成第二文本，其中，所述第二文本包括所述第一关键词语及所述第一关键词语之外的词语
S130

图1

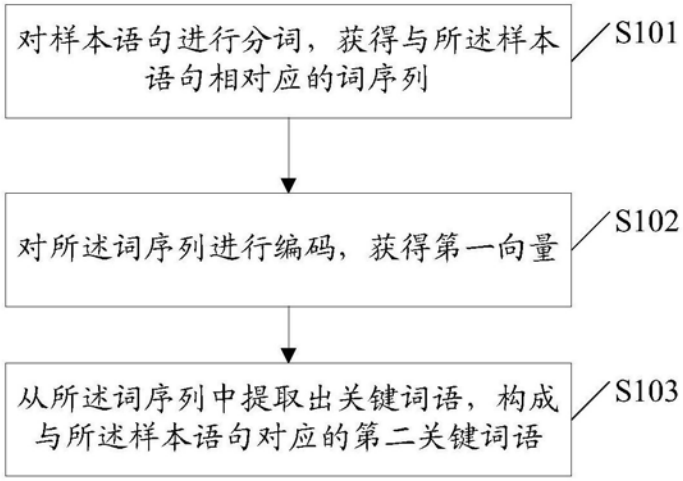

对样本语句进行分词，获得与所述样本语句相对应的词序列
S101
对所述词序列进行编码，获得第一向量
S102
从所述词序列中提取出关键词语，构成与所述样本语句对应的第二关键词语
S103

图2

图3

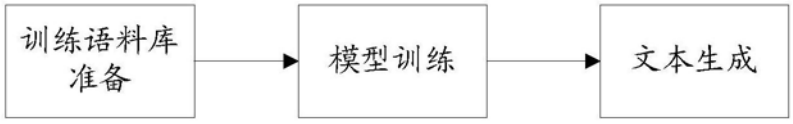


图4

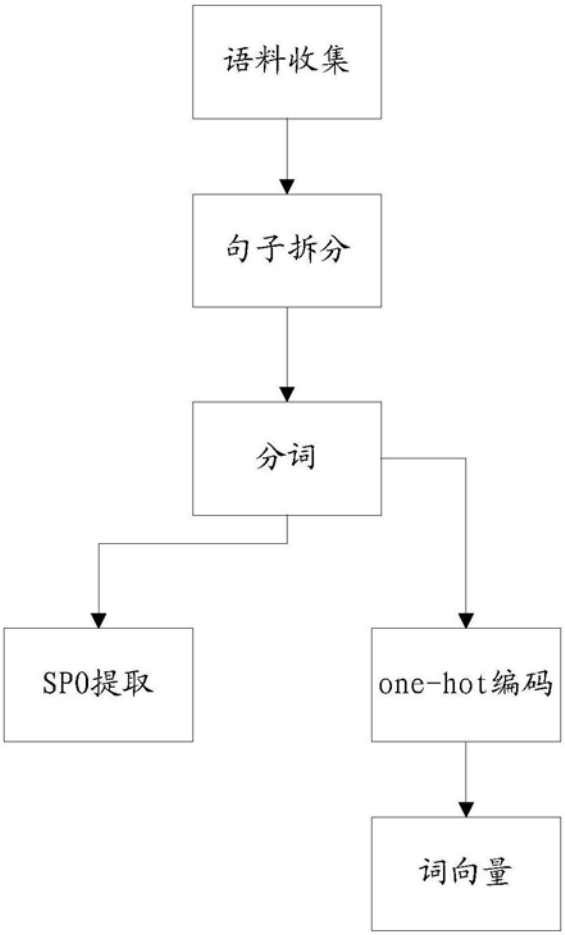

语料收集
句子拆分
分词
SPO提取
one-hot编码
词向量

图5

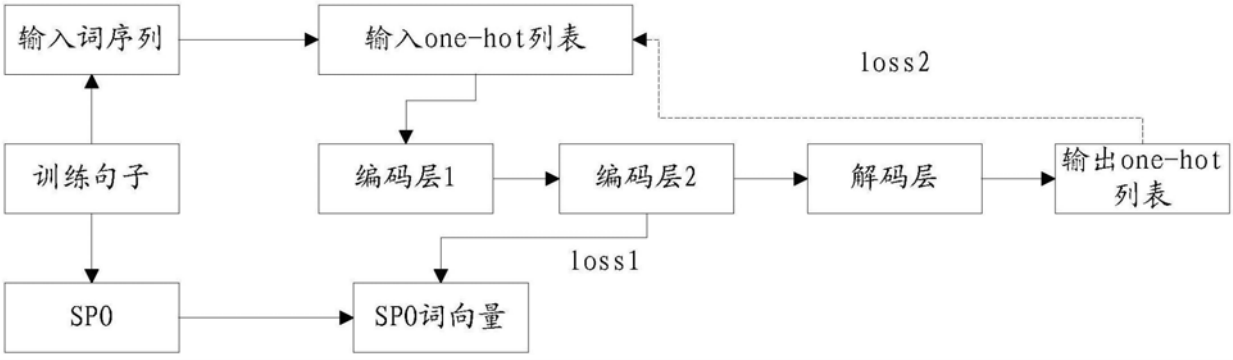

输入词序列
输入one-hot列表
loss2
训练句子
编码层1
编码层2
解码层
输出one-hot列表
loss1
SPO
SPO词向量

图6

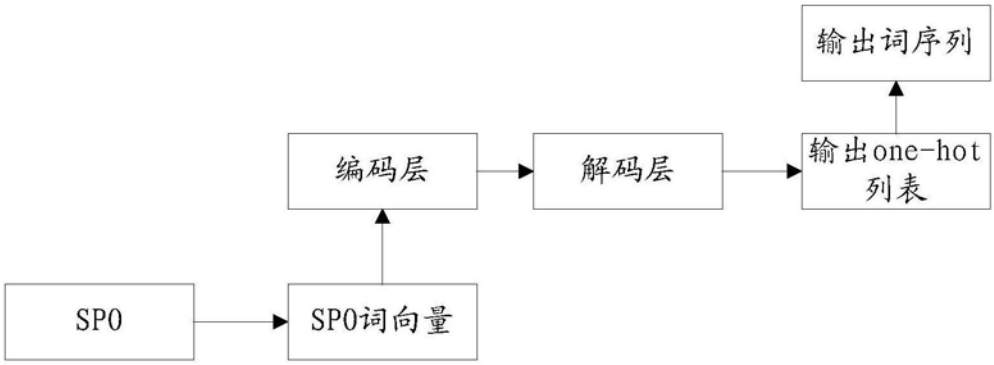


图7