(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117056490 A
(43)申请公布日 2023.11.14

(21)申请号 202311093501.4

(22)申请日 2023.08.28

(71)申请人 平安银行股份有限公司
地址 518000 广东省深圳市罗湖区深南东路5047号

(72)发明人 詹乐 龚静

(74)专利代理机构 深圳中细软知识产权代理有限公司 44528
专利代理师 曹远浩

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/34(2019.01)
G06Q 40/02(2023.01)
G06N 5/022(2023.01)
G06N 5/04(2023.01)

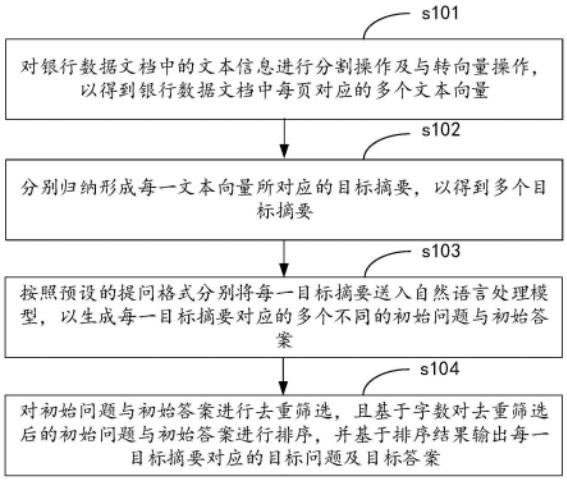
权利要求书2页 说明书9页 附图2页

(54)发明名称

问题提取及答案生成方法、装置、介质和设备

(57)摘要

本发明公开了一种问题提取及答案生成方法、装置、介质和设备，对银行数据文档中的文本信息进行分割和向量化，得到每页的多个文本向量；对每个文本向量进行摘要，得到多个目标摘要；对每个目标摘要使用预设的提问格式，送入自然语言处理模型，得到多个初始问题和答案；对初始问题和答案进行去重和排序，输出每个目标摘要对应的目标问题和答案。本发明能对银行数据文档进行解析，自动生成相关问题和答案，帮助用户快速理解文档内容，及时掌握行业信息动态。



CN 117056490 A

1.一种问题提取及答案生成方法,其特征在于,所述方法包括:

对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量;

分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要;

按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案;

对所述初始问题与所述初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

2.根据权利要求1所述的方法,其特征在于,所述对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量,包括:

读取所述银行数据文档中每页的文本信息,并将目标页的文本信息与相邻页的浮动变量作为所述目标页的文本块,以得到多页的文本块;其中,所述目标页为银行数据文档中的任意一页,所述相邻页为所述目标页的上一页和/或下一页,所述浮动变量为所述相邻页中与所述目标页临近的预设长度的文本信息;

以字数和/或段落为划分单位对每页的文本块进行划分,以得到每页的子文本块;

使用预设的句子相似度模型将每页的子文本块嵌入为对应的文本向量。

3.根据权利要求1所述的方法,其特征在于,所述分别归纳形成每一文本向量所对应的目标摘要,包括:

分别将每一向量块输入到预设的摘要生成模型中,并获取输出的初始摘要;其中,每一向量块由一子文本块和对应的文本向量构成;

分别基于句子边界和/或关键词对每一初始摘要进行提取,以得到每一初始摘要对应的少于预设字数阈值的目标摘要。

4.根据权利要求1所述的方法,其特征在于,所述对所述初始问题与所述初始答案进行去重筛选,包括:

在所有的初始问题中,分别计算两两初始问题之间的余弦相似度并进行排序,以得到多个问题对组成的第一排序结果;

在所有的初始答案中,分别计算两两初始答案之间的余弦相似度并进行排序,以得到多个答案对组成的第二排序结果;

在所述第一排序结果中,对余弦相似度最大的前N个问题及对应的初始答案进行第一去重操作;其中,N为预设值,所述第一去重操作指示删除每一问题对中的任意一个初始问题及删除对应的初始答案;

在所述第二排序结果中,对余弦相似度最大的前M个答案及对应的初始问题进行第二去重操作;其中,M为预设值,所述第二去重操作指示删除每一答案对中的任意一个初始答案及删除对应的初始问题。

5.根据权利要求1所述的方法,其特征在于,所述基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案,包括:

返回字数最多的前L个答案作为目标答案,并将所述目标答案对应的问题作为目标问题与所述目标答案一同输出;其中,L为预设值。

6.一种问题提取及答案生成装置,其特征在于,所述装置包括:

文本向量生成模块,用于对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量;

目标摘要生成模块,用于分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要;

结果输出模块,用于按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案;及对所述初始问题与所述初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

7.根据权利要求6所述的问题提取及答案生成装置,其特征在于,所述文本向量生成模块,具体用于:

读取所述银行数据文档中每页的文本信息,并将目标页的文本信息与相邻页的浮动变量作为所述目标页的文本块,以得到多页的文本块;其中,所述目标页为银行数据文档中的任意一页,所述相邻页为所述目标页的上一页和/或下一页,所述浮动变量为所述相邻页中与所述目标页临近的预设长度的文本信息;

以字数和/或段落为划分单位对每页的文本块进行划分,以得到每页的子文本块;

使用预设的句子相似度模型将每页的子文本块嵌入为对应的文本向量。

8.根据权利要求6所述的问题提取及答案生成装置,其特征在于,所述目标摘要生成模块,具体用于:

分别将每一向量块输入到预设的摘要生成模型中,并获取输出的初始摘要;其中,每一向量块由一子文本块和对应的文本向量构成;

分别基于句子边界和/或关键词对每一初始摘要进行提取,以得到每一初始摘要对应的少于预设字数阈值的目标摘要。

9.一种计算机可读存储介质,其特征在于,存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行如权利要求1至5中任一项所述方法的步骤。

10.一种问题提取及答案生成设备,其特征在于,包括存储器和处理器,所述存储器存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行如权利要求1至5中任一项所述方法的步骤。

问题提取及答案生成方法、装置、介质和设备

技术领域

[0001] 本发明涉及银行技术领域,尤其是涉及一种问题提取及答案生成方法、装置、介质和设备。

背景技术

[0002] 当前,智能文档解析是银行的重要业务场景,然而长银行数据文档解析是一项较为困难的处理场景。如果文档过长,用户并不清楚文档中有哪些内容,无法提出相应问题来让系统解答。这时,就可以为用户自动生成相关问题和答案,帮助用户快速理解文档内容,及时掌握行业信息动态。

发明内容

[0003] 基于此,有必要提供问题提取及答案生成方法、装置、介质和设备,以解决文档过长的情况下,难以自动提取文档中的相关问题和答案的问题。

[0004] 一种问题提取及答案生成方法,所述方法包括:

[0005] 对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量;

[0006] 分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要;

[0007] 按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案;

[0008] 对所述初始问题与所述初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

[0009] 在其中一个实施例中,所述对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量,包括:

[0010] 读取所述银行数据文档中每页的文本信息,并将目标页的文本信息与相邻页的浮动变量作为所述目标页的文本块,以得到多页的文本块;其中,所述目标页为银行数据文档中的任意一页,所述相邻页为所述目标页的上一页和/或下一页,所述浮动变量为所述相邻页中与所述目标页临近的预设长度的文本信息;

[0011] 以字数和/或段落为划分单位对每页的文本块进行划分,以得到每页的子文本块;

[0012] 使用预设的句子相似度模型将每页的子文本块嵌入为对应的文本向量。

[0013] 在其中一个实施例中,所述分别归纳形成每一文本向量所对应的目标摘要,包括:

[0014] 分别将每一向量块输入到预设的摘要生成模型中,并获取输出的初始摘要;其中,每一向量块由一子文本块和对应的文本向量构成;

[0015] 分别基于句子边界和/或关键词对每一初始摘要进行提取,以得到每一初始摘要对应的少于预设字数阈值的目标摘要。

[0016] 在其中一个实施例中,所述对所述初始问题与所述初始答案进行去重筛选,包括:

[0017] 在所有的初始问题中,分别计算两两初始问题之间的余弦相似度并进行排序,以得到多个问题对组成的第一排序结果;

[0018] 在所有的初始答案中,分别计算两两初始答案之间的余弦相似度并进行排序,以得到多个答案对组成的第二排序结果;

[0019] 在所述第一排序结果中,对余弦相似度最大的前N个问题及对应的初始答案进行第一去重操作;其中,N为预设值,所述第一去重操作指示删除每一问题对中的任意一个初始问题及删除对应的初始答案;

[0020] 在所述第二排序结果中,对余弦相似度最大的前M个答案及对应的初始问题进行第二去重操作;其中,M为预设值,所述第二去重操作指示删除每一答案对中的任意一个初始答案及删除对应的初始问题。

[0021] 在其中一个实施例中,所述基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案,包括:

[0022] 返回字数最多的前L个答案作为目标答案,并将所述目标答案对应的问题作为目标问题与所述目标答案一同输出;其中,L为预设值。

[0023] 一种问题提取及答案生成装置,所述装置包括:

[0024] 文本向量生成模块,用于对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量;

[0025] 目标摘要生成模块,用于分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要;

[0026] 结果输出模块,用于按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案;及对所述初始问题与所述初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

[0027] 在其中一个实施例中,所述文本向量生成模块,具体用于:

[0028] 读取所述银行数据文档中每页的文本信息,并将目标页的文本信息与相邻页的浮动变量作为所述目标页的文本块,以得到多页的文本块;其中,所述目标页为银行数据文档中的任意一页,所述相邻页为所述目标页的上一页和/或下一页,所述浮动变量为所述相邻页中与所述目标页临近的预设长度的文本信息;

[0029] 以字数和/或段落为划分单位对每页的文本块进行划分,以得到每页的子文本块;

[0030] 使用预设的句子相似度模型将每页的子文本块嵌入为对应的文本向量。

[0031] 在其中一个实施例中,所述目标摘要生成模块,具体用于:

[0032] 分别将每一向量块输入到预设的摘要生成模型中,并获取输出的初始摘要;其中,每一向量块由一子文本块和对应的文本向量构成;

[0033] 分别基于句子边界和/或关键词对每一初始摘要进行提取,以得到每一初始摘要对应的少于预设字数阈值的目标摘要。

[0034] 一种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,使得所述处理器执行上述问题提取及答案生成方法的步骤。

[0035] 一种问题提取及答案生成设备,包括存储器和处理器,所述存储器存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行上述问题提取及答案生

成方法的步骤。

[0036] 本发明提供了问题提取及答案生成方法、装置、介质和设备，对银行数据文档中的文本信息进行分割和向量化，得到每页的多个文本向量；对每个文本向量进行摘要，得到多个目标摘要；对每个目标摘要使用预设的提问格式，送入自然语言处理模型，得到多个初始问题和答案；对初始问题和答案进行去重和排序，输出每个目标摘要对应的目标问题和答案。本发明能对银行数据文档进行解析，自动生成相关问题和答案，帮助用户快速理解文档内容，及时掌握行业信息动态。

附图说明

[0037] 为了更清楚地说明本发明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发明的一些实施例，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。

[0038] 其中：

[0039] 图1为问题提取及答案生成方法的流程示意图；

[0040] 图2为问题提取及答案生成装置的结构示意图；

[0041] 图3为问题提取及答案生成设备的结构框图。

具体实施方式

[0042] 下面将结合本发明实施例中的附图，对本发明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发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

[0043] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用于描述特定顺序。此外，术语“包括”和“具有”以及它们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可选地还包括没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其他步骤或单元。

[0044] 在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

[0045] 如图1所示，图1为一个实施例中问题提取及答案生成方法的流程示意图，本实施例中问题提取及答案生成方法提供的步骤包括：

[0046] S101，对银行数据文档中的文本信息进行分割操作及与转向量操作，以得到银行数据文档中每页对应的多个文本向量。

[0047] 其中，银行数据文档是指包含银行相关数据和信息的文档，例如银行概况、业务介绍、财务报告等。分割操作是指将文本信息按照一定的规则或标准划分为若干个部分，每个部分包含一定的语义信息。例如，我们可以根据段落、句子、标点符号等来分割文本信息。转

向量操作是指将文本信息转换为数值向量,以便于计算机处理和分析。例如,我们可以使用词嵌入技术,将每个词或短语映射到一个高维空间中的一个点,从而得到词向量或短语向量。

[0048] 在一个具体实施中,通过如下的具体步骤来获得每页对应的多个文本向量,包括:

[0049] (1)、读取银行数据文档中每页的文本信息,并将目标页的文本信息与相邻页的浮动变量作为目标页的文本块,以得到多页的文本块。

[0050] 其中,目标页为银行数据文档中的任意一页,相邻页为目标页的上一页和/或下一页,浮动变量为相邻页中与目标页临近的预设长度的文本信息。

[0051] 这一步骤是为了将每一页的文本信息扩展一些上下文,以便于捕捉更多的语义信息。例如,假设银行数据文档有三页,每页有一段文字,如下:

[0052] 第一部分:银行概况

[0053] 银行是一种金融机构,主要从事存款、贷款、支付、结算、信用卡等业务。银行是金融体系的核心组成部分,对经济发展和社会稳定起着重要作用。

[0054] 第二部分:银行业务

[0055] 银行业务主要分为两大类:存款业务和贷款业务。存款业务是指银行接受客户存入的资金,并按照约定支付利息的业务。贷款业务是指银行向客户提供资金,并收取利息和手续费的业务。

[0056] 第三部分:银行风险

[0057] 银行风险是指银行在经营过程中可能遭受的损失或损害。银行风险主要包括信用风险、市场风险、流动性风险、操作风险等。银行需要采取有效的风险管理措施,以保障资产安全和盈利能力。

[0058] 假设我们选择第二页作为目标页,那么我们可以将第二页的文本信息与第一页和第三页的最后一句话作为浮动变量,拼接起来作为第二页的文本块,如下:

[0059] 银行是金融体系的核心组成部分,对经济发展和社会稳定起着重要作用。银行业务主要分为两大类:存款业务和贷款业务。存款业务是指银行接受客户存入的资金,并按照约定支付利息的业务。贷款业务是指银行向客户提供资金,并收取利息和手续费的业务。银行风险是指银行在经营过程中可能遭受的损失或损害。

[0060] 同理,我们可以对其他两页也进行类似的操作,得到三个文本块。当然可以理解的是,这里的预设长度是可以根据需求自行设定的。

[0061] (2)、以字数和/或段落为划分单位对每页的文本块进行划分,以得到每页的子文本块。

[0062] 这一步骤是为了将每个文本块进一步细化为若干个子文本块,以便于提取更精细的语义信息。例如,我们可以根据段落或者固定的字数来划分第二页的文本块,如下:

[0063] 子文本块1:银行是金融体系的核心组成部分,对经济发展和社会稳定起着重要作用。

[0064] 子文本块2:银行业务主要分为两大类:存款业务和贷款业务。

[0065] 子文本块3:银行是金融体系的核心组成部分,对经济发展和社会稳定起着重要作用。

[0066] 子文本块4:银行业务主要分为两大类:存款业务和贷款业务。存款业务是指银行

接受客户存入的资金,并按照约定支付利息的业务。

[0067] 子文本块5:贷款业务是指银行向客户提供资金,并收取利息和手续费的业务。

[0068] 子文本块6:银行风险是指银行在经营过程中可能遭受的损失或损害。

[0069] 子文本块7:存款业务是指银行接受客户存入的资金,并按照约定支付利息的业务。

[0070] 子文本块8:银行风险是指银行在经营过程中可能遭受的损失或损害。

[0071] (3)、使用预设的句子相似度模型将每页的子文本块嵌入为对应的文本向量。

[0072] 这一步骤是为了将每个子文本块转换为数值向量,以便于计算机处理和分析。例如选用simcse模型,simcse是一种基于对比学习的句子相似度模型,它可以通过自身预测自身的方式,学习到句子的语义表示。例如,我们可以使用simcse模型将第二页的所有子文本块嵌入为一个768维的向量,如下:

[0073] 子文本块1->文本向量1:[0.12,-0.34,...,-0.45]

[0074] 子文本块2->文本向量2:[-0.23,0.56,...,-0.67]

[0075] 子文本块3->文本向量3:[0.34,-0.78,...,-0.89]

[0076] 子文本块4->文本向量4:[-0.23,0.56,...,-0.67]

[0077] 子文本块5->文本向量5:[0.45,-0.89,...,-0.12]

[0078] 子文本块6->文本向量6:[0.67,-0.34,...,-0.45]

[0079] 子文本块7->文本向量7:[0.78,-0.12,...,-0.56]

[0080] 子文本块8->文本向量8:[0.45,-0.89...,-0.50]

[0081] 这样,我们就得到了每页对应的多个文本向量,它们可以反映每个子文本块的语义信息,也可以用于后续的计算和分析。

[0082] S102,分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要。

[0083] 其中,该步骤S102是一个自然语言处理的任务,它的目的是根据每个文本向量的语义内容,归纳出一个简短的目标摘要,以便于概括文本的主要信息。目标摘要可以是一句话,也可以是一个短语或一个词。

[0084] 在一个具体实施中,形成对应的目标摘要的具体步骤,包括:

[0085] (1)、分别将每一向量块输入到预设的摘要生成模型中,并获取输出的初始摘要。

[0086] 其中,每一向量块由一子文本块和对应的文本向量构成。这一子步骤是为了利用预设的摘要生成模型,例如gpt-2模型,根据每个向量块的文本信息和数值表示,生成一个初始摘要。

[0087] 示例性的,在另一个银行卡的示例中:

[0088] 向量块1->初始摘要1:银行卡是一种银行发行的电子支付工具,它可以储存用户的资金和个人信息,方便用户在自动取款机、POS机或网络平台上进行消费或转账等操作。银行卡通常有磁条和芯片两种形式,芯片卡相比磁条卡更安全和稳定。

[0089] 向量块2->初始摘要2:银行卡的种类有多种,主要分为借记卡、信用卡和准贷记卡三大类。借记卡是指用户存入资金后才能使用的卡片,信用卡是指银行给用户提供一定额度的贷款,用户可以先消费后还款的卡片,准贷记卡是指银行给用户提供一定额度的预授权,用户可以在额度内透支消费的卡片。不同的卡种有不同的功能和费用,用户应根据自己的需求和能力选择合适的卡种。

[0090] 向量块3->初始摘要3:银行卡的使用需要注意安全和合理,避免泄露密码或个人信息,及时还款或查询余额。使用银行卡时,应保管好卡片和密码,不要将密码写在卡片上或告诉他人,不要在不安全的设备或网站上输入银行卡信息。使用信用卡时,应注意还款期限和利息,按时还清欠款,避免产生逾期费或滞纳金。使用借记卡时,应注意查询余额和交易记录,及时发现并处理异常情况。

[0091] (2)、分别基于句子边界和/或关键词对每一初始摘要进行提取,以得到每一初始摘要对应的少于预设字数阈值的目标摘要。

[0092] 其中,句子边界是指句子开始和结束的位置,通常用标点符号来划分;我们可以根据句子的长度和位置,选择最能代表原文主旨的句子,或者将多个句子进行合并或删减,以形成一个简洁的目标摘要。可选的,此处的预设字数阈值可以是50字。例如:

[0093] 初始摘要1->目标摘要1:银行卡是一种电子支付工具,可以储存资金和信息,支持多种操作。芯片卡比磁条卡更安全。

[0094] 而关键词是指能够反映文本主题或核心内容的词语。我们可以根据原文中出现的关键词或短语,选择最能反映原文主题和内容的词语,或者将多个词语进行组合或替换,以形成一个精炼的目标摘要。可选的,此处的预设字数阈值可以是50字。

[0095] 初始摘要2->目标摘要2:银行卡分为借记卡、信用卡和准贷记卡。借记卡需存款,信用卡需还款,准贷记卡可透支。不同卡种有不同特点和费用。

[0096] S103,按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案。

[0097] 具体的,假设我们有以下的目标摘要作为输入:银行卡是一种银行发行的电子支付工具,它可以储存用户的资金和个人信息,方便用户在自动取款机、POS机或网络平台上进行消费或转账等操作。银行卡通常有磁条和芯片两种形式,芯片卡相比磁条卡更安全和稳定。

[0098] 那么我们可以使用不同的方法来生成与文本相关的初始问题与初始答案:例如基于预训练语言模型(pre-trained language model)的模型:这种模型可以利用大规模的语料库进行预训练,学习语言的通用知识和规律,然后在特定的任务上进行微调,提高生成质量和效率。例如,我们可以使用BERT、GPT-2等预训练语言模型,在文本前加上一个特殊的标记(如[Q]),然后让模型根据标记生成相应的问题。其中,可能的初始问题与初始答案:

[0099] 初始问题:银行卡有哪两种形式?初始答案:银行卡通常有磁条和芯片两种形式。

[0100] 初始问题:什么是银行卡?初始答案:银行卡是一种银行发行的电子支付工具,它可以储存用户的资金和个人信息,方便用户在自动取款机、POS机或网络平台上进行消费或转账等操作。

[0101] 初始问题:芯片卡相比磁条卡有什么优势?初始答案:芯片卡相比磁条卡更安全和稳定。

[0102] S104,对初始问题与初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

[0103] 可以理解的是,在步骤S103生成的初始问题及初始答案中可能存在重复或相似的内容,而这些重复的内容对于用户来说并非是必须的,因此还需进行一系列的筛选操作。

[0104] 在一个具体实施中,对初始问题与初始答案进行去重筛选,包括:

[0105] (1)、在所有的初始问题中,分别计算两两初始问题之间的余弦相似度并进行排序,以得到多个问题对组成的第一排序结果;

[0106] 这一步是为了找出相似度最高的初始问题对,也就是那些内容重复或者相近的初始问题。余弦相似度是一种通过计算两个向量的夹角余弦值来评估他们的相似度的方法。在这里,每个初始问题可以看作是一个由词语组成的向量,词语可以用词频、TF-IDF等方法进行权重赋值。计算两个初始问题之间的余弦相似度,就可以得到它们之间的相似程度,越接近1表示越相似,越接近0表示越不相关。将所有初始问题两两进行余弦相似度计算,并按照从大到小的顺序进行排序,就可以得到第一排序结果,它是由多个问题对组成的列表,每个问题对都有一个对应的余弦相似度值。

[0107] (2)、在所有的初始答案中,分别计算两两初始答案之间的余弦相似度并进行排序,以得到多个答案对组成的第二排序结果。

[0108] 这一步是为了找出相似度最高的初始答案对,也就是那些内容重复或者相近的初始答案。余弦相似度的计算方法和上一步一样,只是将初始问题换成了初始答案。

[0109] (3)、在第一排序结果中,对余弦相似度最大的前N个问题及对应的初始答案进行第一去重操作;其中,N为预设值,第一去重操作指示删除每一问题对中的任意一个初始问题及删除对应的初始答案。

[0110] 这一步是为了删除那些重复或者相近的初始问题及其对应的初始答案,以减少冗余信息。N是一个预设值,表示要删除多少个问题对。

[0111] (4)、在第二排序结果中,对余弦相似度最大的前M个答案及对应的初始问题进行第二去重操作;其中,M为预设值,第二去重操作指示删除每一答案对中的任意一个初始答案及删除对应的初始问题。

[0112] 这一步是为了进一步删除那些重复或者相近的初始答案及其对应的初始问题,以进一步减少冗余信息。M是一个预设值,表示要删除多少个答案对。

[0113] 在一个具体实施中,基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案,包括:返回字数最多的前L个答案作为目标答案,并将目标答案对应的问题作为目标问题与目标答案一同输出;其中,L为预设值。

[0114] 这一步是为了从去重筛选后的初始问题与初始答案中选择最能覆盖目标摘要内容的目标问题及目标答案,并输出。L是一个预设值,表示要选择多少个答案。根据初始答案的字数,按照从多到少的顺序进行排序,然后选取字数最多的前L个答案作为目标答案,并将它们对应的初始问题作为目标问题与目标答案一同输出。

[0115] 上述问题提取及答案生成方法,对银行数据文档中的文本信息进行分割和向量化,得到每页的多个文本向量;对每个文本向量进行摘要,得到多个目标摘要;对每个目标摘要使用预设的提问格式,送入自然语言处理模型,得到多个初始问题和答案;对初始问题和答案进行去重和排序,输出每个目标摘要对应的目标问题和答案。可见,本发明能对银行数据文档进行解析,自动生成相关问题和答案,帮助用户快速理解文档内容,及时掌握行业信息动态。

[0116] 在一个实施例中,如图2所示,提出了一种问题提取及答案生成装置,该装置包括:

[0117] 文本向量生成模块201,用于对银行数据文档中的文本信息进行分割操作及与转

向量操作,以得到银行数据文档中每页对应的多个文本向量;

[0118] 目标摘要生成模块202,用于分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要;

[0119] 结果输出模块203,用于按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案;及对所述初始问题与所述初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

[0120] 在其中一个实施例中,所述文本向量生成模块201,具体用于:

[0121] 读取所述银行数据文档中每页的文本信息,并将目标页的文本信息与相邻页的浮动变量作为所述目标页的文本块,以得到多页的文本块;其中,所述目标页为银行数据文档中的任意一页,所述相邻页为所述目标页的上一页和/或下一页,所述浮动变量为所述相邻页中与所述目标页临近的预设长度的文本信息;

[0122] 以字数和/或段落为划分单位对每页的文本块进行划分,以得到每页的子文本块;

[0123] 使用预设的句子相似度模型将每页的子文本块嵌入为对应的文本向量。

[0124] 在其中一个实施例中,所述目标摘要生成模块202,具体用于:

[0125] 分别将每一向量块输入到预设的摘要生成模型中,并获取输出的初始摘要;其中,每一向量块由一子文本块和对应的文本向量构成;

[0126] 分别基于句子边界和/或关键词对每一初始摘要进行提取,以得到每一初始摘要对应的少于预设字数阈值的目标摘要。

[0127] 结果输出模块203,具体用于:在所有的初始问题中,分别计算两两初始问题之间的余弦相似度并进行排序,以得到多个问题对组成的第一排序结果;

[0128] 在所有的初始答案中,分别计算两两初始答案之间的余弦相似度并进行排序,以得到多个答案对组成的第二排序结果;

[0129] 在所述第一排序结果中,对余弦相似度最大的前N个问题及对应的初始答案进行第一去重操作;其中,N为预设值,所述第一去重操作指示删除每一问题对中的任意一个初始问题及删除对应的初始答案;

[0130] 在所述第二排序结果中,对余弦相似度最大的前M个答案及对应的初始问题进行第二去重操作;其中,M为预设值,所述第二去重操作指示删除每一答案对中的任意一个初始答案及删除对应的初始问题。

[0131] 结果输出模块203,具体用于:返回字数最多的前L个答案作为目标答案,并将所述目标答案对应的问题作为目标问题与所述目标答案一同输出;其中,L为预设值。

[0132] 图3示出了一个实施例中问题提取及答案生成设备的内部结构图。如图3所示,该问题提取及答案生成设备包括通过系统总线连接的处理器、存储器和网络接口。其中,存储器包括非易失性存储介质和内存储器。该问题提取及答案生成设备的非易失性存储介质存储有操作系统,还可存储有计算机程序,该计算机程序被处理器执行时,可使得处理器实现问题提取及答案生成方法。该内存储器中也可储存有计算机程序,该计算机程序被处理器执行时,可使得处理器执行问题提取及答案生成方法。本领域技术人员可以理解,图3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问题提取及答案生成设备的限定,具体的问题提取及答案生成设备可以包括比图中

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。

[0133] 一种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实现如下步骤:对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量;分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要;按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案;对所述初始问题与所述初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

[0134] 一种问题提取及答案生成设备,包括存储器、处理器以及存储在该存储器中并可在该处理器上执行的计算机程序,该处理器执行该计算机程序时实现如下步骤:对银行数据文档中的文本信息进行分割操作及与转向量操作,以得到银行数据文档中每页对应的多个文本向量;分别归纳形成每一文本向量所对应的目标摘要,以得到多个目标摘要;按照预设的提问格式分别将每一目标摘要送入自然语言处理模型,以生成每一目标摘要对应的多个不同的初始问题与初始答案;对所述初始问题与所述初始答案进行去重筛选,且基于字数对去重筛选后的初始问题与初始答案进行排序,并基于排序结果输出每一目标摘要对应的目标问题及目标答案。

[0135] 需要说明的是,上述问题提取及答案生成方法、装置、设备及计算机可读存储介质属于一个总的发明构思,问题提取及答案生成方法、装置、设备及计算机可读存储介质实施例中的内容可相互适用。

[0136] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,该程序可存储于一非易失性计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0137] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0138] 以上实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对本申请专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范围。因此,本申请专利的保护范围应以所附权利要求为准。

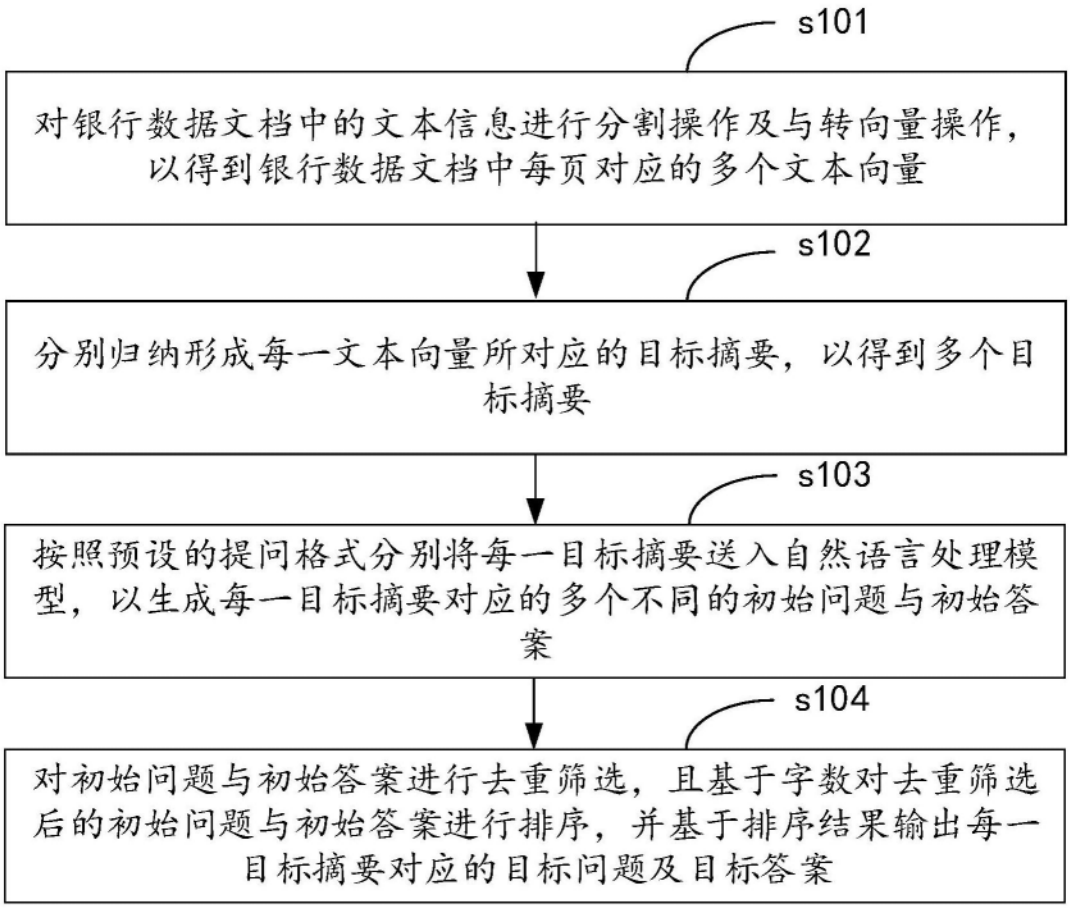

s101
对银行数据文档中的文本信息进行分割操作及与转向量操作，以得到银行数据文档中每页对应的多个文本向量
s102
分别归纳形成每一文本向量所对应的目标摘要，以得到多个目标摘要
s103
按照预设的提问格式分别将每一目标摘要送入自然语言处理模型，以生成每一目标摘要对应的多个不同的初始问题与初始答案
s104
对初始问题与初始答案进行去重筛选，且基于字数对去重筛选后的初始问题与初始答案进行排序，并基于排序结果输出每一目标摘要对应的目标问题及目标答案

图1

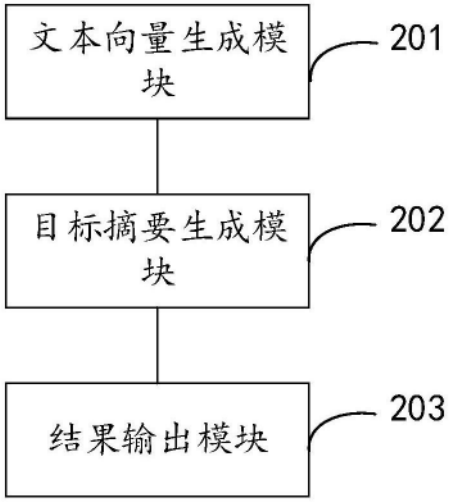

文本向量生成模块
201
目标摘要生成模块
202
结果输出模块
203

图2

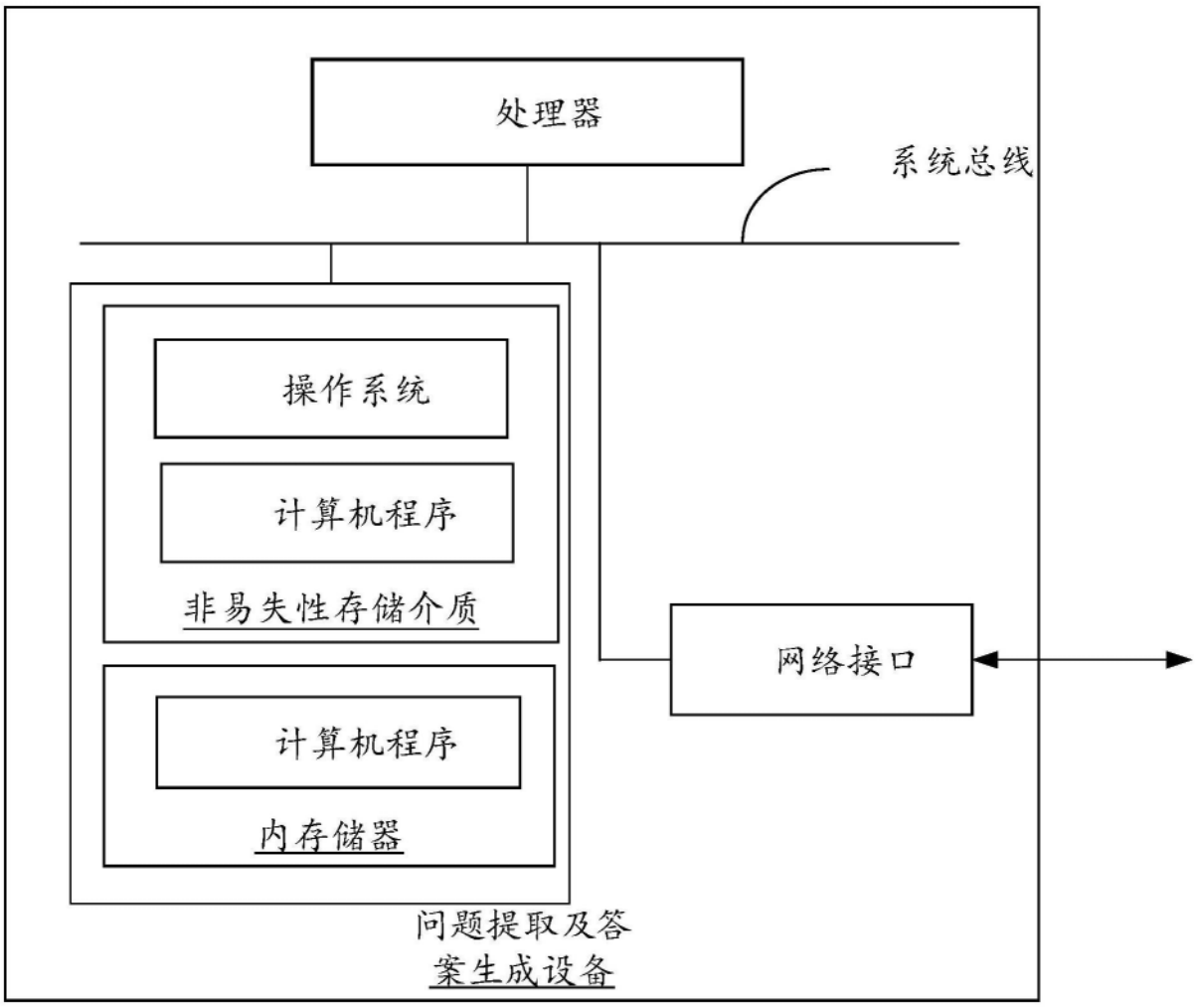


图3