(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113590768 B
(45)授权公告日 2023.10.27

(21)申请号 202010363564.7

(22)申请日 2020.04.30

(65)同一申请的已公布的文献号
申请公布号 CN 113590768 A

(43)申请公布日 2021.11.02

(73)专利权人 北京金山数字娱乐科技有限公司
地址 100085 北京市海淀区小营西路33号金山软件大厦2层西区

(72)发明人 李长亮 冯晓阳 唐剑波

(74)专利代理机构 北京智信禾专利代理有限公司 11637
专利代理师 王治东

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/36(2019.01)
G06F 40/289(2020.01)
G06N 3/0442(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 110781663 A,2020.02.11
CN 110032632 A,2019.07.19
WO 2019214145 A1,2019.11.14
JP 2017228272 A,2017.12.28
余正涛;樊孝忠;宋丽哲;高盛祥.汉语问答系统答案提取方法研究.计算机工程.2006,(03),全文.

审查员 陈飞

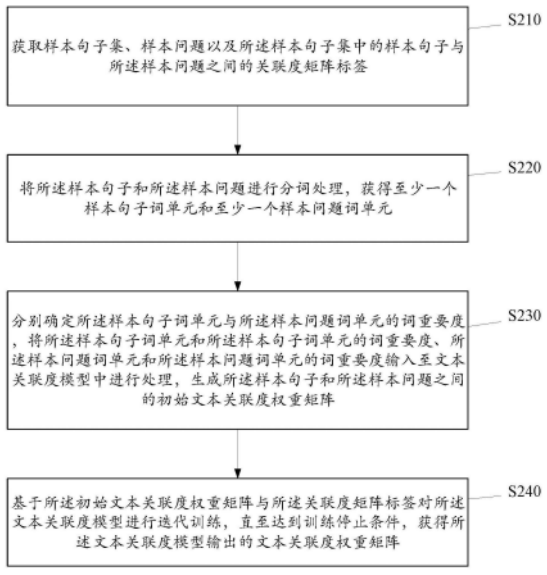
权利要求书3页 说明书19页 附图4页

(54)发明名称

一种文本关联度模型的训练方法及装置、问答方法及装置

(57)摘要

本申请提供一种文本关联度模型的训练方法及装置、问答方法及装置。其中，所述问答方法，包括：获取待回答问题，对所述待回答问题进行分词处理，获得多个问题词单元；确定所述问题词单元中的关键词单元以及所述关键词单元的词重要度，并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵；基于所述关键词矩阵与所述文本关联度权重矩阵确定目标句子，并基于所述目标句子生成所述待回答问题的答案。本申请提供的文本关联度模型的训练方法及装置，不仅可以为问答系统智能度的提高提供助力，还可以加快训练过程中参数的收敛，提高训练速度；本申请所述的问答方法及装置可以有效提高问题回答的智能度以及生成答案的准确度和流畅度。



CN 113590768 B

1.一种文本关联度模型的训练方法,其特征在于,包括:

获取样本句子集、样本问题以及所述样本句子集中的样本句子与所述样本问题之间的关联度矩阵标签;

将所述样本句子和所述样本问题进行分词处理,获得至少一个样本句子词单元和至少一个样本问题词单元;

分别确定所述样本句子词单元与所述样本问题词单元的词重要度,将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理,生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵;

基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练,直至达到训练停止条件,获得所述文本关联度模型输出的文本关联度权重矩阵。

2.根据权利要求1所述的文本关联度模型的训练方法,其特征在于,在所述获取样本句子集之前,还包括:

获取语料文本,通过主题分类算法对所述语料文本进行分类,获得具有类别标签的多个样本句子集。

3.根据权利要求1所述的文本关联度模型的训练方法,其特征在于,所述分别确定所述样本句子词单元与所述样本问题词单元的词重要度,包括:

分别确定所述样本句子词单元与所述样本问题词单元的词频、词性和情感极性;

基于所述样本句子词单元的词频、词性和情感极性确定所述样本句子词单元的词重要度;

基于所述样本问题词单元的词频、词性和情感极性确定所述样本问题词单元的词重要度。

4.根据权利要求1所述的文本关联度模型的训练方法,其特征在于,所述将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理,生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵,包括:

将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中;

基于所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度确定所述样本句子词单元与所述样本问题词单元之间的关联度;

基于所述样本句子词单元与所述样本问题词单元之间的关联度,生成所述样本句子与所述样本问题之间的初始文本关联度权重矩阵。

5.根据权利要求1所述的文本关联度模型的训练方法,其特征在于,所述基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练,包括:

基于所述初始文本关联度权重矩阵与所述关联度矩阵标签确定损失值,并判断所述损失值是否大于预设阈值;

若是,则基于所述损失值对所述文本关联度模型进行调整;

若否,则结束训练并输出文本关联度权重矩阵。

6.一种问答方法,其特征在于,包括:

获取待回答问题,对所述待回答问题进行分词处理,获得多个问题词单元;

确定所述问题词单元中的关键词单元以及所述关键词单元的词重要度,并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵;

基于所述关键词矩阵与文本关联度权重矩阵确定目标句子,并基于所述目标句子生成所述待回答问题的答案,其中,所述文本关联度权重矩阵通过上述权利要求1-5任意一项所述方法确定。

7.根据权利要求6所述的问答方法,其特征在于,所述确定所述问题词单元中的关键词单元,包括:

确定每一个所述问题词单元的词频、词性和/或情感极性,并基于所述问题词单元的词频、词性和/或情感极性确定关键词单元。

8.根据权利要求6所述的问答方法,其特征在于,所述确定所述关键词单元的词重要度,包括:

确定每一个所述关键词单元的词频、词性和情感极性;

基于所述关键词单元的词频、词性和情感极性确定所述关键词单元的词重要度。

9.根据权利要求6所述的问答方法,其特征在于,所述基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵,包括:

基于所述关键词单元获得词向量矩阵;

基于所述关键词单元的词重要度获得词重要度矩阵;

将所述词向量矩阵和所述词重要度矩阵进行融合处理,生成关键词矩阵。

10.根据权利要求6所述的问答方法,其特征在于,所述基于所述关键词矩阵与文本关联度权重矩阵确定目标句子,包括:

确定所述关键词矩阵与预设句子集的类别标签之间的类别关联度,并基于所述类别关联度确定目标句子集;

基于关键词矩阵与目标句子集中每一个句子的文本关联度权重矩阵确定所述待回答问题与所述目标句子集中每一个句子之间的内容关联度,并基于所述内容关联度确定至少一个目标句子。

11.根据权利要求6所述的问答方法,其特征在于,所述基于所述目标句子生成所述待回答问题的答案,包括:

通过语义分析单元提取所述目标句子中的目标词单元;

基于所述目标词单元生成所述待回答问题的答案。

12.根据权利要求11所述的问答方法,其特征在于,还包括:

通过实体识别单元识别所述目标句子中的时间标签,并基于所述时间标签对所述待回答问题的答案进行更新。

13.根据权利要求11或12所述的问答方法,其特征在于,还包括:

通过净化单元过滤所述答案中的负面词单元,并对所述答案进行更新。

14.一种文本关联度模型的训练装置,其特征在于,包括:

样本获取模块,被配置为获取样本句子集、样本问题以及所述样本句子集中的样本句

子与所述样本问题之间的关联度矩阵标签；

分词处理模块，被配置为将所述样本句子和所述样本问题进行分词处理，获得至少一个样本句子词单元和至少一个样本问题词单元；

矩阵生成模块，被配置为分别确定所述样本句子词单元与所述样本问题词单元的词重要度，将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵；

迭代训练模块，被配置为基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，直至达到训练停止条件，获得所述文本关联度模型输出的文本关联度权重矩阵。

15.一种问答装置，其特征在于，包括：

问题分词模块，被配置为获取待回答问题，对所述待回答问题进行分词处理，获得多个问题词单元；

关键词矩阵生成模块，被配置为确定所述问题词单元中的关键词单元以及所述关键词单元的词重要度，并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵；

答案生成模块，被配置为基于所述关键词矩阵与文本关联度权重矩阵确定目标句子，并基于所述目标句子生成所述待回答问题的答案，其中，所述文本关联度权重矩阵通过上述权利要求1-5任意一项所述方法确定。

16.一种计算设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机指令，其特征在于，所述处理器执行所述指令时实现权利要求1-5或者6-13任意一项所述方法的步骤。

17.一种计算机可读存储介质，其存储有计算机指令，其特征在于，该指令被处理器执行时实现权利要求1-5或者6-13任意一项所述方法的步骤。

一种文本关联度模型的训练方法及装置、问答方法及装置

技术领域

[0001] 本申请涉及计算机技术领域，特别涉及一种文本关联度模型的训练方法及装置、问答方法及装置、计算设备及计算机可读存储介质。

背景技术

[0002] 智能问答系统是信息检索系统的一种高级形式，是基于人们对快速、准确地获取信息的需求而兴起的，可以用准确、简洁的自然语言回答用户用自然语言提出的问题。

[0003] 目前，现有的智能问答系统是将积累的无序语料信息进行有序和科学的整理，并建立分类模型，用以指导新增加的语料咨询和服务信息，节约人力资源，提高信息处理的自动性，降低网站运行成本，基于对网站多年积累的常见问题及其解答，整理为规范的问答库形式，以支撑各种形式问题的智能问答。

[0004] 但是，现有的智能问答系统训练时间长、生成的答案较为宽泛，导致整个智能问答系统的智能度不够高，这成为亟待解决的问题。

发明内容

[0005] 有鉴于此，本申请实施例提供了一种文本关联度模型的训练方法及装置、问答方法及装置、计算设备及计算机可读存储介质，以解决现有技术中存在的技术缺陷。

[0006] 本申请实施例公开了一种文本关联度模型的训练方法，包括：

[0007] 获取样本句子集、样本问题以及所述样本句子集中的样本句子与所述样本问题之间的关联度矩阵标签；

[0008] 将所述样本句子和所述样本问题进行分词处理，获得至少一个样本句子词单元和至少一个样本问题词单元；

[0009] 分别确定所述样本句子词单元与所述样本问题词单元的词重要度，将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵；

[0010] 基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，直至达到训练停止条件，获得所述文本关联度模型输出的文本关联度权重矩阵。

[0011] 进一步地，在所述获取样本句子集之前，还包括：

[0012] 获取语料文本，通过主题分类算法对所述语料文本进行分类，获得具有类别标签的多个样本句子集。

[0013] 进一步地，所述分别确定所述样本句子词单元与所述样本问题词单元的词重要度，包括：

[0014] 分别确定所述样本句子词单元与所述样本问题词单元的词频、词性和情感极性；

[0015] 基于所述样本句子词单元的词频、词性和情感极性确定所述样本句子词单元的词

重要度；

[0016] 基于所述样本问题词单元的词频、词性和情感极性确定所述样本问题词单元的词重要度。

[0017] 进一步地，所述将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵，包括：

[0018] 将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中；

[0019] 基于所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度确定所述样本句子词单元与所述样本问题词单元之间的关联度；

[0020] 基于所述样本句子词单元与所述样本问题词单元之间的关联度，生成所述样本句子与所述样本问题之间的初始文本关联度权重矩阵。

[0021] 进一步地，所述基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，包括：

[0022] 基于所述初始文本关联度权重矩阵与所述关联度矩阵标签确定损失值，并判断所述损失值是否大于预设阈值；

[0023] 若是，则基于所述损失值对所述文本关联度模型进行调整；

[0024] 若否，则结束训练并输出文本关联度权重矩阵。

[0025] 本申请实施例还提供了一种问答方法，包括：

[0026] 获取待回答问题，对所述待回答问题进行分词处理，获得多个问题词单元；

[0027] 确定所述问题词单元中的关键词单元以及所述关键词单元的词重要度，并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵；

[0028] 基于所述关键词矩阵与所述文本关联度权重矩阵确定目标句子，并基于所述目标句子生成所述待回答问题的答案。

[0029] 进一步地，所述确定所述问题词单元中的关键词单元，包括：

[0030] 确定每一个所述问题词单元的词频、词性和/或情感极性，并基于所述问题词单元的词频、词性和/或情感极性确定关键词单元。

[0031] 进一步地，所述确定所述关键词单元的词重要度，包括：

[0032] 确定每一个所述关键词单元的词频、词性和情感极性；

[0033] 基于所述关键词单元的词频、词性和情感极性确定所述关键词单元的词重要度。

[0034] 进一步地，所述基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵，包括：

[0035] 基于所述关键词单元获得词向量矩阵；

[0036] 基于所述关键词单元的词重要度获得词重要度矩阵；

[0037] 将所述词向量矩阵和所述词重要度矩阵进行融合处理，生成关键词矩阵。

[0038] 进一步地，所述基于所述关键词矩阵与所述文本关联度权重矩阵确定目标句子，包括：

[0039] 确定所述关键词矩阵与预设句子集的类别标签之间的类别关联度，并基于所述类

别关联度确定目标句子集；

[0040] 基于关键词矩阵与目标句子集中每一个句子的文本关联度权重矩阵确定所述待回答问题与所述目标句子集中每一个句子之间的内容关联度，并基于所述内容关联度确定至少一个目标句子。

[0041] 进一步地，所述基于所述目标句子生成所述待回答问题的答案，包括：

[0042] 通过语义分析单元提取所述目标句子中的目标词单元；

[0043] 基于所述目标词单元生成所述待回答问题的答案。

[0044] 进一步地，所述问答方法，还包括：

[0045] 通过实体识别单元识别所述目标句子中的时间标签，并基于所述时间标签对所述待回答问题的答案进行更新。

[0046] 进一步地，所述问答方法，还包括：

[0047] 通过净化单元过滤所述答案中的负面词单元，并对所述答案进行更新。

[0048] 本申请还提供一种文本关联度模型的训练装置，包括：

[0049] 样本获取模块，被配置为获取样本句子集、样本问题以及所述样本句子集中的样本句子与所述样本问题之间的关联度矩阵标签；

[0050] 分词处理模块，被配置为将所述样本句子和所述样本问题进行分词处理，获得至少一个样本句子词单元和至少一个样本问题词单元；

[0051] 矩阵生成模块，被配置为分别确定所述样本句子词单元与所述样本问题词单元的词重要度，将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵；

[0052] 迭代训练模块，被配置为基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，直至达到训练停止条件，获得所述文本关联度模型输出的文本关联度权重矩阵。

[0053] 本申请还提供一种问答装置，包括：

[0054] 问题分词模块，被配置为获取待回答问题，对所述待回答问题进行分词处理，获得多个问题词单元；

[0055] 关键词矩阵生成模块，被配置为确定所述问题词单元中的关键词单元以及所述关键词单元的词重要度，并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵；

[0056] 答案生成模块，被配置为基于所述关键词矩阵与所述文本关联度权重矩阵确定目标句子，并基于所述目标句子生成所述待回答问题的答案。

[0057] 本申请还提供一种计算设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机指令，所述处理器执行所述指令时实现所述文本关联度模型的训练方法或者所述问答方法的步骤。

[0058] 本申请还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现所述文本关联度模型的训练方法或者所述问答方法的步骤。

[0059] 本申请提供的文本关联度模型的训练方法及装置，通过文本关联度模型对样本句子词单元及其词重要度、样本问题词单元及其词重要度进行处理，生成样本句子和样本问

题之间的初始文本关联度权重矩阵，进而获得样本句子与样本问题之间的语义关联度；再基于初始文本关联度权重矩阵和关联度矩阵标签对文本关联度模型进行迭代训练，初始文本关联度矩阵的权重系数随着训练过程的不断推进而逐步更新，不断在细粒度的层面上学习样本问题与样本句子的语义关联，从而实现初始文本关联度权重矩阵的最优化，即获得用于识别提问意图、提高智能问答准确性的文本关联度权重矩阵，不仅可以为问答系统智能度的提高提供助力，还可以加快训练过程中参数的收敛，提高训练速度。

[0060] 本申请提供的问答方法及装置，通过确定待回答问题中的关键词单元及其重要度，获得关键词矩阵，将其与文本关联度权重矩阵一同处理确定目标句子，可以更好的捕捉回答问题与句子之间的语义关联，选取语义关联度高的句子作为目标句子后，再基于目标句子生成待回答问题的答案，可以有效提高问题回答的智能度以及生成答案的准确度和流畅度。

[0061] 此外，本申请提供的问答方法及装置，可以通过实体识别单元提取目标句子中的时间标签，对答案进行优化，以提高答案与现实时间线之间的匹配度，进而提高答案的准确度；还可以通过净化单元过滤答案中的负面词单元，以实现答案中冗余信息的去除，敏感词、争议词等负面词的过滤。

[0062] 本申请提供的问答方法及装置，还可以灵活的应用于政务问答、历史问答、常识问答等各种领域。以政务问答为例，本实施例所述的问答方法，能够全面地捕捉用户提问问题与政务文档之间的语义关联，精准地实现政务文本中的时间线匹配，以及敏感词、争议短语等的过滤，保证答案句子生成的准确度和流畅度，用准确、简洁的自然语言回答用户提出的政务领域的问题，满足人们对快速、准确地获取政务信息的需求。

附图说明

[0063] 图1是本申请一实施例的计算设备的结构示意图；

[0064] 图2是本申请一实施例的文本关联度模型的训练方法的步骤流程示意图；

[0065] 图3是本申请一实施例的双向LSTM模型的结构示意图；

[0066] 图4是本申请一实施例的问答方法的步骤流程示意图；

[0067] 图5是本申请一实施例的文本关联度模型的训练装置的结构示意图；

[0068] 图6是本申请一实施例的问答装置的结构示意图。

具体实施方式

[0069] 在下面的描述中阐述了很多具体细节以便于充分理解本申请。但是本申请能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况下做类似推广，因此本申请不受下面公开的具体实施的限制。

[0070] 在本说明书一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本说明书一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0071] 应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描

述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一也可以被称为第二，类似地，第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0072] 首先，对本发明一个或多个实施例涉及的名词术语进行解释。

[0073] 长短期记忆网络(Long Short-Term Memory，LSTM)模型：是为了解决一般的循环神经网络(Recurrent Neural Network，RNN)存在的长期依赖问题而专门设计出来的一种时间循环神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。LSTM模型可以用来连接先前的信息到当前的任务上，例如使用过去的语句来推测对当前语句的理解。

[0074] 双向长短期记忆网络(Bi-directional Long Short-Term Memory，BiLSTM)模型：由前向LSTM与后向LSTM组合而成的模型，在自然语言处理任务中常被用来建模上下文信息，并生成对应的隐藏层向量表征。

[0075] 文本关联度模型：是用于生成文本关联度权重矩阵的模型。本申请中的文本关联度模型可以为BiLSTM模型。

[0076] 样本句子集：由多个属于同一个主题类别的样本句子组成的集合。

[0077] 样本句子：在文本关联度模型的训练阶段输入至文本关联度模型中的句子，样本句子包括以句号、感叹号、问号等语句结束标志为分隔符分隔而成的词单元集合。

[0078] 样本问题：在文本关联度模型的训练阶段，输入文本关联度模型的问题句子。

[0079] 关联度矩阵标签：基于样本句子与样本问题之间的真实关联度而生成的该样本句子与该样本问题之间的文本关联度权重矩阵。

[0080] 词单元(token)：对输入文本做任何实际处理前，都需要将其分割成诸如词、标点符号、数字或纯字母数字等语言单元，这些单元被称为词单元。对于英文文本，词单元可以是一个单词、一个标点符号、一个数字等，对于中文文本，最小的词单元可以是一个字、一个标点符号、一个数字、一个词等。

[0081] 词重要度：基于词单元在句子中的词频、词性以及情感极性而计算得到的一种权重参数。其中，句子包括样本句子、样本问题、待回答问题等。

[0082] 词频：是指词单元在句子中出现的次数。其中，句子包括样本句子、样本问题、待回答问题等。

[0083] 词性：是以语法特征(包括句法功能和形态变化)为主要依据、兼顾词汇意义对词单元进行划分的结果，包括名词、动词、副词等。

[0084] 情感极性：是指词单元具有的感情色彩，包括正向、负向等。

[0085] 文本关联度权重矩阵：基于样本句子与样本问题之间的关联度产生的权重矩阵。

[0086] 语料文本：由多个样本句子组合而成的自然语言文本。

[0087] 待回答问题：用户输入至交互界面中的亟待解决的问题句子。

[0088] 词向量矩阵：对输入的句子的词单元进行向量化处理形成的矩阵。

[0089] 词重要度矩阵：对词单元的词重要度进行向量化处理形成的矩阵。

[0090] 关键词矩阵：词单元的词向量矩阵及其词重要度矩阵融合而成的矩阵。

[0091] 预设句子集：预先根据句子主题类别的不同划分而成的句子集。

[0092] 目标句子集:与待回答问题主题相符的预设句子集。

[0093] 目标句子:目标句子集中与待回答问题之间的关联度大于预设阈值的句子。

[0094] 语义分析单元:即语义依存分析工具,用于构造答案句子主成分。

[0095] 实体识别单元:即NER命名实体识别模块,用于识别并提取出时间标签。

[0096] 时间标签:可以表示时间的字、词或短语。

[0097] 净化单元:通过净化词典过滤答案句子中负面词语的执行单元。

[0098] 净化词典:是一种反面语料库,包括各种反动暴力、色情淫秽、人身攻击、低俗偏激等方面的反面词语语料。

[0099] 在本申请中,提供了一种文本关联度模型的训练方法及装置、问答方法及装置、计算设备及计算机可读存储介质,在下面的实施例中逐一进行详细说明。

[0100] 图1是示出了根据本说明书一实施例的计算设备100的结构框图。该计算设备100的部件包括但不限于存储器110和处理器120。处理器120与存储器110通过总线130相连接,数据库150用于保存数据。

[0101] 计算设备100还包括接入设备140,接入设备140使得计算设备100能够经由一个或多个网络160通信。这些网络的示例包括公用交换电话网(PSTN)、局域网(LAN)、广域网(WAN)、个域网(PAN)或诸如因特网的通信网络的组合。接入设备140可以包括有线或无线的任何类型的网络接口(例如,网络接口卡(NIC))中的一个或多个,诸如IEEE802.11无线局域网(WLAN)无线接口、全球微波互联接入(Wi-MAX)接口、以太网接口、通用串行总线(USB)接口、蜂窝网络接口、蓝牙接口、近场通信(NFC)接口,等等。

[0102] 在本说明书的一个实施例中,计算设备100的上述部件以及图1中未示出的其他部件也可以彼此相连接,例如通过总线。应当理解,图1所示的计算设备结构框图仅仅是出于示例的目的,而不是对本说明书范围的限制。本领域技术人员可以根据需要,增添或替换其他部件。

[0103] 计算设备100可以是任何类型的静止或移动计算设备,包括移动计算机或移动计算设备(例如,平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等)、移动电话(例如,智能手机)、可佩戴的计算设备(例如,智能手表、智能眼镜等)或其他类型的移动设备,或者诸如台式计算机或PC的静止计算设备。计算设备100还可以是移动式或静止式的服务器。

[0104] 其中,处理器120可以执行图2所示方法中的步骤。图2是示出了根据本申请一实施例的文本关联度模型的训练方法的示意性流程图,包括步骤S210至步骤S240。

[0105] S210、获取样本句子集、样本问题以及所述样本句子集中的样本句子与所述样本问题之间的关联度矩阵标签。

[0106] 具体地,在获取样本句子集之前,先获取语料文本,通过主题分类算法对所述语料文本进行分类,获得具有类别标签的多个样本句子集。

[0107] 其中,语料文本是由多个句子组合而成的自然语言文本,可以是一篇文章、多篇文章等各种篇幅长度的文本,也可以是中文文本、英文文本等各种语言类型的文本,本申请对此不做限制。主题分类算法是用于确定语料文本中的句子主题类别,并将属于相同主题类别的句子归为一个集合即样本句子集的算法,每一个样本句子集具有的类别标签可以表示该样本句子集中全部样本句子的主题类别。

[0108] 假设获取100篇语料文本，此100篇语料文本中共包括10000个句子，通过任务启发式主题分类算法对上述100篇语料文本的10000个句子进行分类，获得具有类别标签的多个样本句子集，分类过程如下：

[0109] 将上述100篇语料文本的10000个句子作为数据集D，$D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_{10000},y_m)\}$，其中，$x_i$是每一个句子的字向量集(n维向量表示)，$y_i$是类别标签向量集(降维到d维向量表示)，$y_i\in\{C_1,C_2,\ldots,C_k\}$，C表示类别标签。

[0110] 首先，通过如下公式计算类内散度矩阵：

[0111] $$S_w=\sum 0+\sum 1=\sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^T+\sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^T \quad (1)$$

[0112] 上述公式(1)中的S_w表示类内散度矩阵，μ_j(j=0,1)为第j类样本的均值向量，亦即μ_0、μ_1表示两个类别的中心点，T表示矩阵转置，X_0表示数据集D中的句子集，X_1表示数据集D中的类别标签集。

[0113] 其次，通过如下公式计算类间散度矩阵：

[0114] $$S_b=(\mu_0-\mu_1)(\mu_0-\mu_1)^T \quad (2)$$

[0115] 上述公式(2)中的S_b表示类间散度矩阵，μ_0、μ_1表示两个类别的中心点，T表示矩阵转置。

[0116] 基于上述类内散度矩阵S_w和类间散度矩阵S_b计算得到散度矩阵$S^{-1}_wS_b$，并计算$S^{-1}_wS_b$的最大的d个特征值和对应的d个特征向量$(w_1,w_2,\ldots w_d)$，并将上述d个特征向量$(w_1,w_2,\ldots w_d)$进行拼接后得到投影矩阵W^T。

[0117] 基于上述投影矩阵W^T将样本集中的句子样本特征x_i，转化为新的样本$z_i=W^Tx_i$。

[0118] 将上述样本z_i分别带入每一个类别的高斯分布概率密度函数中，分别计算此样本特征属于每一个类别的概率，其中，最大概率值对应的类别即为此样本特征对应的词向量所属的类别。

[0119] 得到输出样本句子集组合$D'=\{(z_1,y_1),(z_2,y_2),\ldots,(z_m,y_m)\}$，其中，$z_i$表示样本句子集，$y_i$表示样本句子集的分类标签。

[0120] 需要说明的是，上述获取语料文本、通过主题分类算法对语料文本进行分类的过程只需完成一次即可，之后的训练过程可以直接获取样本句子集、样本问题以及样本句子集中的样本句子与样本问题之间的关联度矩阵标签对文本关联度权重矩阵进行训练。

[0121] 本实施例通过获取样本句子集、样本问题以及样本句子集中的样本句子与样本问题之间的关联度矩阵标签可以为文本关联度模型的训练做好准备。

[0122] S220、将所述样本句子和所述样本问题进行分词处理，获得至少一个样本句子词单元和至少一个样本问题词单元。

[0123] 具体地，分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。在实际应用中，对每一个样本句子集中的全部样本句子进行分词处理后，获得多个样本句子词单元；对样本问题进行分词处理后，获得多个样本问题词单元。

[0124] 在本实施例中，假设样本句子集包括z_1-z_m共m个样本句子集，每个样本句子集中包括p_1-p_i共i个样本句子，样本问题包括q_1-q_n共n个样本问题，以样本句子集z_1中的样本句子p_1以及样本问题q_1为例，假设样本句子p_1包括“北京是伟大祖国的首都，深受祖国各族人民

的向往”，对上述样本句子p_1进行分词处理，获得[北京、是、伟大、祖国、的、首都、深受、祖国、各族、人民、的、向往]共12个样本句子词单元；样本问题q_1包括“中国的首都是哪座城市?”，对上述样本问题q_1进行分词处理，获得[中国、的、首都、是、哪]共5个样本问题词单元。

[0125] 本实施例通过对样本句子和样本问题进行分词处理，有助于提高文本关联度模型对样本句子、样本问题的语义理解能力，提升文本关联度模型的训练效果。

[0126] S230、分别确定所述样本句子词单元与所述样本问题词单元的词重要度，将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵。

[0127] 具体地，所述步骤S230还可以包括步骤S231至步骤S235。

[0128] S231、分别确定所述样本句子词单元与所述样本问题词单元的词频、词性和情感极性。

[0129] 其中，词频是指词单元在句子中出现的次数，以样本问题“什么是分词?”为例，词单元“分词”在该样本问题中出现了一次，那么词单元“分词”的词频即为1。

[0130] 词性是以语法特征(包括句法功能和形态变化)为主要依据、兼顾词汇意义对词单元进行划分的结果，以样本问题“什么是分词”为例，词单元“什么”的词性为名词，词单元“是”和词单元“分词”的词性均为动词。

[0131] 情感极性是指词单元具有的感情色彩，包括正向、负向、中性等，其中正向词是指褒义积极词、负向词是指贬义消极词。以“粮食都被浪费了”为例，其中词单元“浪费”的情感极性为负向。

[0132] 以样本句子p_1“北京是伟大祖国的首都，深受祖国各族人民的向往”中的词单元“伟大”为例，该词单元的词频为1，词性为形容词，情感极性为正向词。

[0133] 本实施例通过确定词单元的词频、词性和情感极性，可以促进文本关联度模型从不提供方面加深对包含该词单元的句子的理解层次和深度，进而提高文本关联度模型的语义理解能力。

[0134] S232、基于所述样本句子词单元的词频、词性和情感极性分别确定所述样本句子词单元的词重要度，基于所述样本问题词单元的词频、词性和情感极性确定所述样本问题词单元的词重要度。

[0135] 具体地，可以根据词频、词性和情感极性三个特征权重系数分量根据对应的计分规则单独计算分值，并将三个特征权重系数分量的分值之和作为词重要度的总分值，由此可得每个词单元对应的词重要度值计算公式如下所示：

[0136] $V=Ws+Wq+We$ (3)

[0137] 其中，V表示词重要度，以W表示权重系数，Ws表示词性分值，Wq表示词频分值，We表示情感极性分值。

[0138] 需要说明的是，词单元的词频可以直接作为其词频权重系数分量的分值，比如词单元a在样本问题中的词频为3，那么词单元a的词频权重系数分量的分值为3分。在本实施例中副词不统计词频，比如“的”，不统计其词频，因而“的”这样的成分词无词频得分，遇到重复出现的此类副词，跳过即可。

[0139] 词性权重系数分量的分值和情感极性权重系数分量的分值可以视具体情况而定，比如，名词的词性分值为1分、动词、形容词的词性分值为0.5分、副词的词性分值为0分、正向词即褒义积极词的情感极性分值为1分、负向词即贬义消极词的情感极性分值为-1分、中性词的情感极性分值为0分等，本申请对此不做限制。

[0140] 在本实施例中，以样本句子集z_1中的样本句子p_1以及样本问题q_1为例，样本句子p_1的词频分值、词性分值、情感分值和词重要度如表1所示。

[0141] 表1

[0142]

	北京	是	伟大	祖国	的	首都	，	深受
词性	名词	动词	形容词	名词	副词	名词		动词
词性分值	1	0.5	0.5	1	0	1		0.5
词频分值	1	1	1	2	*	1		1
情感分值	0	0	1	1	0	0		0
词重要度	2	1.5	2.5	4	0	2		1.5
	祖国	各族	人民	的	向往	。		
词性	已统计	名词	名词	跳过	动词			
词性分值		1	1		0.5			
词频分值		1	1		1			
情感分值		0	0		1			
词重要度		2	2		2.5			

[0143] 样本问题q_1的词频分值、词性分值、情感分值和词重要度如表2所示。

[0144] 表2

[0145]

	中国	的	首都	是	哪	?
词性	名词	副词	名词	动词	代词	
词性分值	1	0	1	0.5	0	
词频分值	1	*	1	1	1	
情感分值	0	0	0	0	0	
词重要度	2	0	2	1.5	1	

[0146] 本实施例基于词单元的词频、词性和情感极性确定该词单元的词重要度，有助于模型快速准确的了解词单元在句子中所起的作用，进而从细粒度的层面提高模型对于句子的理解能力。

[0147] S233、将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中。

[0148] 在本实施例中，文本关联度模型优选为双向LSTM模型。

[0149] 在实际应用中，可以将样本句子词单元转化为样本句子词单元对应的词向量矩阵，将样本句子词单元的词重要度转化为样本句子词单元对应的词重要度矩阵，将上述样本句子词单元对应的词向量矩阵和词重要度矩阵进行融合，获得样本句子词单元矩阵；将样本问题词单元转化为样本问题词单元对应的词向量矩阵，将样本问题词单元的词重要度转化为样本问题词单元对应的词重要度矩阵，将上述样本问题词单元对应的词向量矩阵和词重要度矩阵进行融合，获得样本问题词单元矩阵，并将上述样本句子词单元矩阵和样本问题词单元矩阵输入至双向LSTM模型中进行处理，其中，融合的方式可以为拼接等，本申请对此不做限制。

[0150] 以样本句子集z_1中的样本句子p_1以及样本问题q_1为例，首先基于样本句子词单元获得每一个样本句子词单元对应的词向量矩阵pa_1-pa_{12}，基于样本句子词单元的词重要度获得每一个样本句子词单元的词重要度矩阵pb_1-pb_{12}，基于样本问题词单元获得每一个样本问题词单元对应的词向量矩阵qa_1-qa_5，基于样本问题词单元的词重要度获得每一个样本问题词单元的词重要度矩阵qb_1-qb_5。其他情况可以此类推，不再赘述。

[0151] 将样本句子p_1中每一个样本句子词单元对应的词向量矩阵pa_1-pa_{12}与词重要度矩阵pb_1-pb_{12}进行融合，获得样本句子词单元矩阵pab_1-pab_{12}，将样本问题q_1中每一个样本问题词单元对应的词向量矩阵qa_1-qa_5与词重要度矩阵qb_1-qb_5进行融合，获得样本问题词单元矩阵qab_1-qab_5，并将上述样本句子词单元矩阵pab_1-pab_{12}和样本问题词单元矩阵qab_1-qab_5输入至双向LSTM模型中。

[0152] 如图3所示，以样本问题中的样本问题词单元[中国、的、首都、是、哪]为例，其中，W_1表示词单元“中国”，W_2表示词单元“的”，W_3表示词单元“首都”，W_4表示词单元“是”，W_5表示词单元“哪”。

[0153] 将上述样本问题的样本问题词单元矩阵[qab_1、qab_2、qab_3、qab_4、qab_5]正向输入至双向LSTM模型中后，得到正向输出矩阵[$Zqab_1$、$Zqab_2$、$Zqab_3$、$Zqab_4$、$Zqab_5$]，将上述样本问题的样本问题词单元矩阵[qab_1、qab_2、qab_3、qab_4、qab_5]反向输入至双向LSTM模型中后，得到反向输出矩阵[$Fqab_5$、$Fqab_4$、$Fqab_3$、$Fqab_2$、$Fqab_1$]，将每个样本问题词单元的正向输出矩阵和反向输出矩阵进行拼接，即得到该样本问题词单元最终的模型输出矩阵即样本问题词单元矩阵，以样本问题词单元“中国”为例，其输入至双向LSTM模型后，最终的模型输出矩阵即为[$Zqab_1$、$Fqab_5$]。其他情况可依次类推，不再赘述。

[0154] 本实施例通过对词单元的词向量矩阵和词重要度矩阵进行融合，获得词单元矩阵，再将词单元矩阵输入至文本关联度模型中进行处理，有助于文本关联度模型同时关注到词单元自身的特征以及词单元融入在句子中的特征，拓展文本关联度模型的关注方面，此外，文本关联度模型选择双向LSTM模型，有助于提高句子理解的深度。

[0155] S234、基于所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度确定所述样本句子词单元与所述样本问题词单元之间的关联度。

[0156] 具体地，可以基于样本问题词单元矩阵与样本句子词单元矩阵计算每一个样本问

题词单元与每一个样本句子词单元之间的余弦相似度，并将上述样本句子词单元与样本问题词单元之间的余弦相似度作为二者之间的关联度，并基于关联度生成每个样本句子与样本问题之间的初始文本关联度矩阵。

[0157] 其中，余弦相似度的计算公式如下所示：

[0158] $$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (4)$$

[0159] $\cos(\theta)$为余弦相似度的数值，x_i表示样本句子词单元的模型输出矩阵，y_i表示样本问题词单元的模型输出矩阵。

[0160] 本实施例通过计算样本问题词单元与样本句子词单元之间的余弦相似度，有助于快速准确的确定样本句子与样本问题之间的关联性，进而快速确定样本句子集中与样本问题之间关联性最大的样本句子，有助于样本问题的准确回答。

[0161] S235、基于所述样本句子词单元与所述样本问题词单元之间的关联度，生成所述样本句子与所述样本问题之间的初始文本关联度权重矩阵。

[0162] 在本实施例中，基于样本句子词单元与样本问题词单元之间的关联度，生成样本句子与样本问题之间的关联度权重矩阵，可以有效提高对样本句子的剖析度，有助于准确的表征样本句子与样本问题之间的关联性。

[0163] S240、基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，直至达到训练停止条件，获得所述文本关联度模型输出的文本关联度权重矩阵。

[0164] 具体地，可以基于所述初始文本关联度权重矩阵与所述关联度矩阵标签确定损失值，并判断所述损失值是否大于预设阈值。

[0165] 若是，则基于所述损失值对所述文本关联度模型进行调整。

[0166] 若否，则结束训练并输出文本关联度权重矩阵。

[0167] 在本实施例中，以样本句子集z_1中的样本句子p_1以及样本问题q_1为例，假设样本句子p_1与样本问题q_1之间的关联度权重矩阵为Z_1，样本句子p_1以及样本问题q_1之间的关联度权重矩阵标签为Z_0，计算Z_1与Z_0之间的损失值，并在损失值大于预设阈值的情况下，反向调整双向LSTM模型的参数，并对双向LSTM模型进行更新，在损失值小于或等于预设阈值的情况下，结束训练并输出最终的文本关联度权重矩阵。

[0168] 具体地，可以采用交叉熵损失函数计算损失值，交叉熵是表示两个概率分布p、q的差异，其中p表示真实分布即关联度权重矩阵标签为Z_0，q表示非真实分布即关联度权重矩阵为Z_1，那么H(p,q)就称为交叉熵，其计算公式如下所示：

[0169] $$H(p,q)=\sum_i p_i\cdot\ln\frac{1}{q_i}=-\sum_i p_i\cdot\ln q_i \quad (5)$$

[0170] 本实施例提供的文本关联度模型的训练方法，通过文本关联度模型对样本句子词单元及其词重要度、样本问题词单元及其词重要度进行处理，生成样本句子和样本问题之

间的初始文本关联度权重矩阵，进而获得样本句子与样本问题之间的语义关联度；再基于初始文本关联度权重矩阵和关联度矩阵标签对文本关联度模型进行迭代训练，初始文本关联度矩阵的权重系数随着训练过程的不断推进而逐步更新，不断在细粒度的层面上学习样本问题与样本句子的语义关联，从而实现初始文本关联度权重矩阵的最优化，即获得用于识别提问意图、提高智能问答准确性的文本关联度权重矩阵，不仅可以为问答系统智能度的提高提供助力，还可以加快训练过程中参数的收敛，提高训练速度。

[0171] 如图4所示，本实施例公开了一种问答方法，包括步骤S410至步骤S430。

[0172] S410、获取待回答问题，对所述待回答问题进行分词处理，获得多个问题词单元。

[0173] 具体地，待回答问题是用户输入至交互界面中的亟待解决的问题，可以是任何领域的问题，比如可以是生活领域的“废旧电池应如何处理”、文学领域的“朱自清的代表作是什么?”、计算机领域的“什么是自然语言处理”等等，本申请对此不做限制。

[0174] 本实施例通过对待回答问题进行分词处理，有助于提高问答系统对于待回答问题的语义理解能力，提高问题回答的准确性。

[0175] S420、确定所述问题词单元中的关键词单元以及所述关键词单元的词重要度，并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵。

[0176] 具体地，可以确定每一个所述问题词单元的词频、词性和/或情感极性，并基于所述问题词单元的词频、词性和/或情感极性确定关键词单元。换而言之，可以基于问题词单元的词频、词性、情感极性中的任意一种或几种确定关键词单元。

[0177] 例如，假设样本问题为“废旧电池应如何处理”，包括样本词单元[废旧、电池、应、如何、处理]，其中每一个样本词单元的词频均为1，情感极性均为中性词，在此种情况下，则可以依据样本词单元的词性确定关键词单元，其中，词单元“电池”的词性为名词，词单元“处理”的词性为动词，那么确定该样本问题中的关键词单元为“电池”和“处理”。

[0178] 具体地，可以确定每一个所述关键词单元的词频、词性和情感极性；基于所述关键词单元的词频、词性和情感极性确定所述关键词单元的词重要度。其中，词重要度的计算公式如公式(3)所示。

[0179] 例如，关键词单元“电池”的词性为名词，则词性分值为1分，情感极性为中性词，则情感极性分值为0分，词频为1，则词频分值为1分，那么关键词单元“电池”的词重要度为2；关键词单元“处理”的词性为动词，则词性分值为0.5分，情感极性为中性词，则情感极性分值为0分，词频为1，则词频分值为1分，那么关键词单元“处理”的词重要度为1.5。

[0180] 具体地，基于所述关键词单元获得词向量矩阵；基于所述关键词单元的词重要度获得词重要度矩阵；将所述词向量矩阵和所述词重要度矩阵进行融合处理，生成关键词矩阵。

[0181] 优选地，将词向量矩阵和词重要度矩阵进行拼接，生成关键词矩阵。

[0182] 本实施例通过将每一个关键词单元的词向量矩阵和词重要度矩阵进行融合处理，生成关键词矩阵，有助于提高问答系统对待回答问题的理解程度，进而提高问题回答的准确性。

[0183] S430、基于所述关键词矩阵与所述文本关联度权重矩阵确定目标句子，并基于所述目标句子生成所述待回答问题的答案。

[0184] 具体地，可以确定所述关键词矩阵与预设句子集的类别标签之间的类别关联度，

并基于所述类别关联度确定目标句子集;基于关键词矩阵与目标句子集中每一个句子的文本关联度权重矩阵确定所述待回答问题与所述目标句子集中每一个句子之间的内容关联度,并基于所述内容关联度确定至少一个目标句子。

[0185] 例如,假设共有10个预设句子集,上述10个预设句子集的类别标签分别为a_1、a_2……a_{10},分别计算关键词单元“电池”与关键词单元“处理”之间的类别关联度,得到关键词单元“电池”与类别标签a_3的关联度最高,关键词单元“处理”与类别标签a_7的关联度最高,那么将类别标签a_3对应的预设句子集a_3以及类别标签a_7对应的预设句子集a_7作为目标句子集。

[0186] 假设预设句子集a_3中包括b_1-b_{80}在内的80个句子,计算关键词单元“电池”与上述80个句子之间的关联度,得到关键词单元“电池”与句子b_{80}之间的关联度最高,那么句子b_{80}即为目标句子;假设预设句子集a_7中包括c_1-c_{120}在内的120个句子,计算关键词单元“处理”与上述120个句子之间的关联度,得到关键词单元“处理”与句子c_{66}之间的关联度最高,那么句子c_{66}即为目标句子。

[0187] 具体地,可以通过语义分析单元提取所述目标句子中的目标词单元;基于所述目标词单元生成所述待回答问题的答案。

[0188] 其中,语义分析单元是一种语义依存分析工具,语义分析单元通过对目标句子进行语义依存分析,进而提取出目标句子中与待回答问题具有紧密关联的主干词单元,将主干词单元重新排列组合后,即生成待回答问题的答案。

[0189] 在实际应用中,还可以通过实体识别单元识别所述目标句子中的时间标签,并基于所述时间标签对所述待回答问题的答案进行更新。

[0190] 其中,实体识别单元是一种NER命名实体识别模块,句子标签可以是任何能够表示时间的词语、短语,比如去年、今年、明年、昨天、今天、明天、庚子年、2020年、周五、三月等等,本申请对此不做限制。

[0191] 例如,假设目标句子为“2022年奥林匹克冬季奥运会将在中国北京举办”,答案句子为“后年北京将举办冬奥会”,通过实体识别单元识别到目标句子中的时间标签“2022年”后,对答案句子中的时间短语“后年”进行更新,更新后的待回答问题的答案即为“2022年北京将举办冬奥会。”

[0192] 通过实体识别单元对待回答问题的答案进行更新,在答案中涉及到时间的情况下,可以有助于明确待回答问题的答案中的时间线,确保时间线清晰不混乱,在答案中不涉及到时间的情况下,便无需通过实体识别单元对待回答问题的答案进行更新。

[0193] 在实际应用中,还可以通过净化单元过滤所述答案中的负面词单元,并对所述答案进行更新。

[0194] 具体地,负面词单元包括净化词典中的反动暴力、色情淫秽、人身攻击、低俗偏激等类型的词语。

[0195] 在实际应用中,可以计算待回答问题答案中的词单元与净化单元中预设负面词单元之间的余弦相似度,并将相似度大于预设阈值的词单元删除,对待回答问题的答案进行更新。

[0196] 其中,余弦相似度的计算公式如下所示:

[0197] $$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (4)$$

[0198] $\cos(\theta)$为余弦相似值，x_i表示待回答问题中的词单元，y_i表示净化词典中的负面词单元。

[0199] 例如，假设答案句子包括“这部电影真垃圾，评分太低”，计算上述答案句子中每一个词单元与净化单元的净化词典中预设负面词单元之间的余弦相似度，得到词单元“垃圾”与预设负面词单元之间的余弦相似度大于预设阈值，则将词单元“垃圾”从答案句子中删除，并对答案句子的结构进行相应调整、更新后得到待回答问题的答案为“这部电影评分太低”。

[0200] 下面结合具体的例子对本实施例进行进一步说明。

[0201] 假设待回答问题为“我们村有多位老人无人赡养，该怎么办?”将上述待回答问题进行分词处理，获得问题词单元[我们、村、有、多、位、老人、无、人、赡养、该、怎么办]。

[0202] 基于上述每一个问题词单元的词性、情感极性确定待回答问题中的关键词单元为[老人、赡养]。

[0203] 确定上述每一个关键词单元的词频、词性和情感极性，基于上述关键词单元的词频、词性和情感极性确定所述关键词单元的词重要度，如表3所示。

[0204] 表3

[0205]

	老人	赡养
词性	名词	动词
词性分值	1	0.5
词频分值	1	1
情感分值	0	1
词重要度	2	2.5

[0206] 基于关键词单元“老人”获得词向量矩阵A_1，基于关键词单元“老人”的词重要度，获得词重要度矩阵B_1，基于关键词单元“赡养”获得词向量矩阵A_2，基于关键词单元“赡养”的词重要度，获得词重要度矩阵B_2。

[0207] 将关键词单元“老人”的词向量矩阵A_1和词重要度矩阵B_1进行拼接，得到关键词矩阵AB_1，将关键词单元“赡养”的词向量矩阵A_2和词重要度矩阵B_2进行拼接，得到关键词矩阵AB_2。

[0208] 假设共包括3个预设句子集，3个预设句子集的类别标签分别为“社会保险”、“医疗服务”、“福利救助”。

[0209] 基于关键词矩阵AB_1计算关键词单元“老人”与3个预设句子集的类别标签之间的类别关联度，得到关键词单元“老人”与类别标签“社会保险”、“医疗服务”、“福利救助”之间

的类别关联度分别为0.55、0.61、0.88;基于关键词矩阵AB_2计算关键词单元“赡养”与3个预设句子集的类别标签之间的类别关联度,得到关键词单元“赡养”与类别标签“社会保险”、“医疗服务”、“福利救助”之间的类别关联度分别为0.30、0.17、0.95。

[0210] 基于上述类别关联度,确定类别标签为“福利救助”的预设句子集为目标句子集。

[0211] 假设“福利救助”目标句子集包括d_1-d_{10}共10个句子。分别计算关键词矩阵AB_1、AB_2与句子d_1-d_{10}之间的内容关联度,得到句子d_3的内容关联度最高,则确定句子d_3“我国《老年人权益保障法》第二十三条规定:“农村老年人,无劳动能力或无生活来源、无赡养人和扶养人的,或者其赡养人和扶养人确无赡养能力或者扶养能力的,由农村集体经济组织负担保吃、保穿、保住、保医、保葬的五保供养,乡、民族乡、镇人民政府负责组织实施”为目标句子。

[0212] 通过语义分析单元提取上述目标句子中的主干词汇生成待回答问题的答案“无人赡养的农村老年人,由农村集体经济组织负担五保供养,乡、民族乡、镇人民政府负责组织实施”。

[0213] 由于上述待回答问题的答案中未涉及到时间,故无需通过实体识别单元更新答案。

[0214] 通过净化单元计算上述答案句子中每一个词单元与预设负面词单元之间的余弦相似度,并无词单元与预设负面词单元之间的余弦相似度大于预设阈值,故无需删除任何词单元,所以,待回答问题“我们村有多位老人无人赡养,该怎么办?”的答案为“无人赡养的农村老年人,由农村集体经济组织负担五保供养,乡、民族乡、镇人民政府负责组织实施”。

[0215] 本实施例提供的问答方法及装置,通过确定待回答问题中的关键词单元及其重要度,获得关键词矩阵,将其与文本关联度权重矩阵一同处理确定目标句子,可以更好的捕捉回答问题与句子之间的语义关联,选取语义关联度高的句子作为目标句子后,再基于目标句子生成待回答问题的答案,可以有效提高问题回答的智能度以及生成答案的准确度和流畅度。

[0216] 此外,本实施例提供的问答方法,可以通过实体识别单元提取目标句子中的时间标签,对答案进行优化,以提高答案与现实时间线之间的匹配度,进而提高答案的准确度;还可以通过净化单元过滤答案中的负面词单元,以实现答案中冗余信息的去除,敏感词、争议词等负面词的过滤。

[0217] 本实施例所述的问答方法,还可以灵活的应用于政务问答、历史问答、常识问答等各种领域。以政务问答为例,本实施例所述的问答方法,能够全面地捕捉用户提问问题与政务文档之间的语义关联,精准地实现政务文本中的时间线匹配,以及敏感词、争议短语等的过滤,保证答案句子生成的准确度和流畅度,用准确、简洁的自然语言回答用户提出的政务领域的问题,满足人们对快速、准确地获取政务信息的需求。

[0218] 如图5所示,本实施例提供一种文本关联度模型的训练装置,包括:

[0219] 样本获取模块510,被配置为获取样本句子集、样本问题以及所述样本句子集中的样本句子与所述样本问题之间的关联度矩阵标签;

[0220] 分词处理模块520,被配置为将所述样本句子和所述样本问题进行分词处理,获得至少一个样本句子词单元和至少一个样本问题词单元;

[0221] 矩阵生成模块530,被配置为分别确定所述样本句子词单元与所述样本问题词单元的词重要度,将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题

词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵；

[0222] 迭代训练模块540，被配置为基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，直至达到训练停止条件，获得所述文本关联度模型输出的文本关联度权重矩阵。

[0223] 可选地，本实施例所述的文本关联度模型的训练装置，还包括：

[0224] 文本分类模块，被配置为获取语料文本，通过主题分类算法对所述语料文本进行分类，获得具有类别标签的多个样本句子集。

[0225] 可选地，所述矩阵生成模块530，进一步被配置为：

[0226] 分别确定所述样本句子词单元与所述样本问题词单元的词频、词性和情感极性；

[0227] 基于所述样本句子词单元的词频、词性和情感极性分别确定所述样本句子词单元的词重要度；

[0228] 基于所述样本问题词单元的词频、词性和情感极性确定所述样本问题词单元的词重要度。

[0229] 可选地，所述矩阵生成模块530，进一步被配置为：

[0230] 将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中；

[0231] 基于所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度确定所述样本句子词单元与所述样本问题词单元之间的关联度；

[0232] 基于所述样本句子词单元与所述样本问题词单元之间的关联度，生成所述样本句子与所述样本问题之间的初始文本关联度权重矩阵。

[0233] 可选地，所述迭代训练模块540，进一步被配置为：

[0234] 基于所述初始文本关联度权重矩阵与所述关联度矩阵标签确定损失值，并判断所述损失值是否大于预设阈值；

[0235] 若是，则基于所述损失值对所述文本关联度模型进行调整；

[0236] 若否，则结束训练并输出文本关联度权重矩阵。

[0237] 本实施例提供的文本关联度模型的训练装置，通过文本关联度模型对样本句子词单元及其词重要度、样本问题词单元及其词重要度进行处理，生成样本句子和样本问题之间的初始文本关联度权重矩阵，进而获得样本句子与样本问题之间的语义关联度；再基于初始文本关联度权重矩阵和关联度矩阵标签对文本关联度模型进行迭代训练，初始文本关联度矩阵的权重系数随着训练过程的不断推进而逐步更新，不断在细粒度的层面上学习样本问题与样本句子的语义关联，从而实现初始文本关联度权重矩阵的最优化，即获得用于识别提问意图、提高智能问答准确性的文本关联度权重矩阵，不仅可以为问答系统智能度的提高提供助力，还可以加快训练过程中参数的收敛，提高训练速度。

[0238] 如图6所示，本实施例提供了一种问答装置，包括：

[0239] 问题分词模块610，被配置为获取待回答问题，对所述待回答问题进行分词处理，获得多个问题词单元；

[0240] 关键词矩阵生成模块620，被配置为确定所述问题词单元中的关键词单元以及所

述关键词单元的词重要度，并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵；

[0241] 答案生成模块630，被配置为基于所述关键词矩阵与所述文本关联度权重矩阵确定目标句子，并基于所述目标句子生成所述待回答问题的答案。

[0242] 可选地，所述关键词矩阵生成模块620，进一步被配置为：

[0243] 确定每一个所述问题词单元的词频、词性和/或情感极性，并基于所述问题词单元的词频、词性和/或情感极性确定关键词单元。

[0244] 可选地，所述关键词矩阵生成模块620，进一步被配置为：

[0245] 确定每一个所述关键词单元的词频、词性和情感极性；

[0246] 基于所述关键词单元的词频、词性和情感极性确定所述关键词单元的词重要度。

[0247] 可选地，所述关键词矩阵生成模块620，进一步被配置为：

[0248] 基于所述关键词单元获得词向量矩阵；

[0249] 基于所述关键词单元的词重要度获得词重要度矩阵；

[0250] 将所述词向量矩阵和所述词重要度矩阵进行融合处理，生成关键词矩阵。

[0251] 可选地，所述答案生成模块630，进一步被配置为：

[0252] 确定所述关键词矩阵与预设句子集的类别标签之间的类别关联度，并基于所述类别关联度确定目标句子集；

[0253] 基于关键词矩阵与目标句子集中每一个句子的文本关联度权重矩阵确定所述待回答问题与所述目标句子集中每一个句子之间的内容关联度，并基于所述内容关联度确定至少一个目标句子。

[0254] 可选地，所述答案生成模块630，进一步被配置为：

[0255] 通过语义分析单元提取所述目标句子中的目标词单元；

[0256] 基于所述目标词单元生成所述待回答问题的答案。

[0257] 可选地，本实施例所述的问答装置，还包括：

[0258] 识别更新模块，被配置为通过实体识别单元识别所述目标句子中的时间标签，并基于所述时间标签对所述待回答问题的答案进行更新。

[0259] 可选地，本实施例所述的问答装置，还包括：

[0260] 净化更新模块，被配置为通过净化单元过滤所述答案中的负面词单元，并对所述答案进行更新。

[0261] 本实施例提供的问答装置，通过确定待回答问题中的关键词单元及其重要度，获得关键词矩阵，将其与文本关联度权重矩阵一同处理确定目标句子，可以更好的捕捉回答问题与句子之间的语义关联，选取语义关联度高的句子作为目标句子后，再基于目标句子生成待回答问题的答案，可以有效提高问题回答的智能度以及生成答案的准确度和流畅度。

[0262] 此外，本实施例提供的问答装置，可以通过实体识别单元提取目标句子中的时间标签，对答案进行优化，以提高答案与现实时间线之间的匹配度，进而提高答案的准确度；还可以通过净化单元过滤答案中的负面词单元，以实现答案中冗余信息的去除，敏感词、争议词等负面词的过滤。

[0263] 本实施例提供的问答装置，还可以灵活的应用于政务问答、历史问答、常识问答等

各种领域。以政务问答为例，本实施例所述的问答方法，能够全面地捕捉用户提问问题与政务文档之间的语义关联，精准地实现政务文本中的时间线匹配，以及敏感词、争议短语等的过滤，保证答案句子生成的准确度和流畅度，用准确、简洁的自然语言回答用户提出的政务领域的问题，满足人们对快速、准确地获取政务信息的需求。

[0264] 本申请一实施例还提供一种计算设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机指令，所述处理器执行所述指令时实现以下步骤：

[0265] 获取样本句子集、样本问题以及所述样本句子集中的样本句子与所述样本问题之间的关联度矩阵标签；

[0266] 将所述样本句子和所述样本问题进行分词处理，获得至少一个样本句子词单元和至少一个样本问题词单元；

[0267] 分别确定所述样本句子词单元与所述样本问题词单元的词重要度，将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵；

[0268] 基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，直至达到训练停止条件，并获得所述文本关联度模型输出的文本关联度权重矩阵。

[0269] 本申请一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现如前所述文本关联度模型的训练方法或问答方法的步骤。

[0270] 上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的文本关联度模型的训练方法或问答方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述文本关联度模型的训练方法或问答方法的技术方案的描述。

[0271] 所述计算机指令包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括电载波信号和电信信号。

[0272] 需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定都是本申请所必须的。

[0273] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其它实施例的相关描述。

[0274] 以上公开的本申请优选实施例只是用于帮助阐述本申请。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书的内容，

可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本申请的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本申请。本申请仅受权利要求书及其全部范围和等效物的限制。

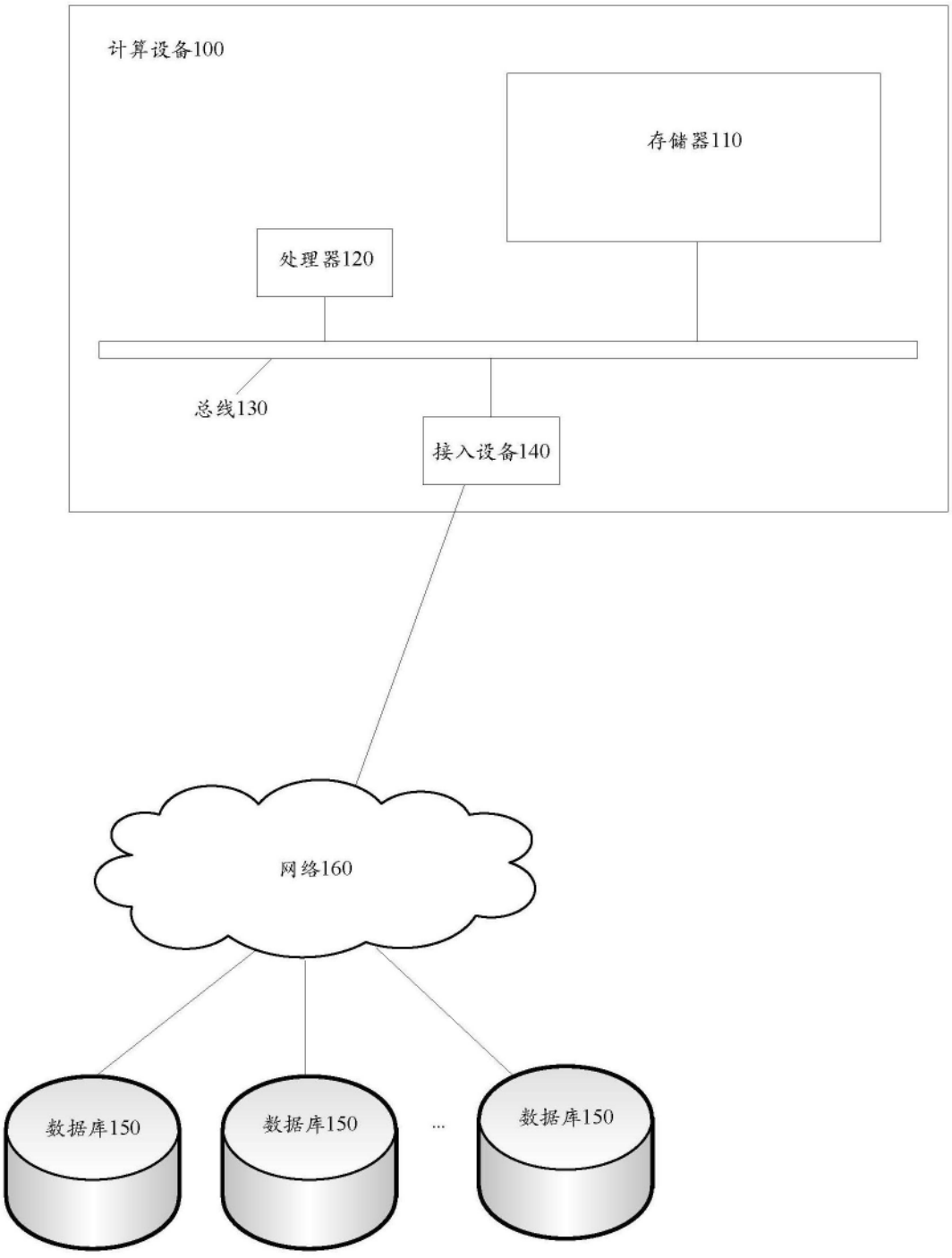

计算设备100
存储器110
处理器120
总线130
接入设备140
网络160
数据库150
数据库150
...
数据库150

图1

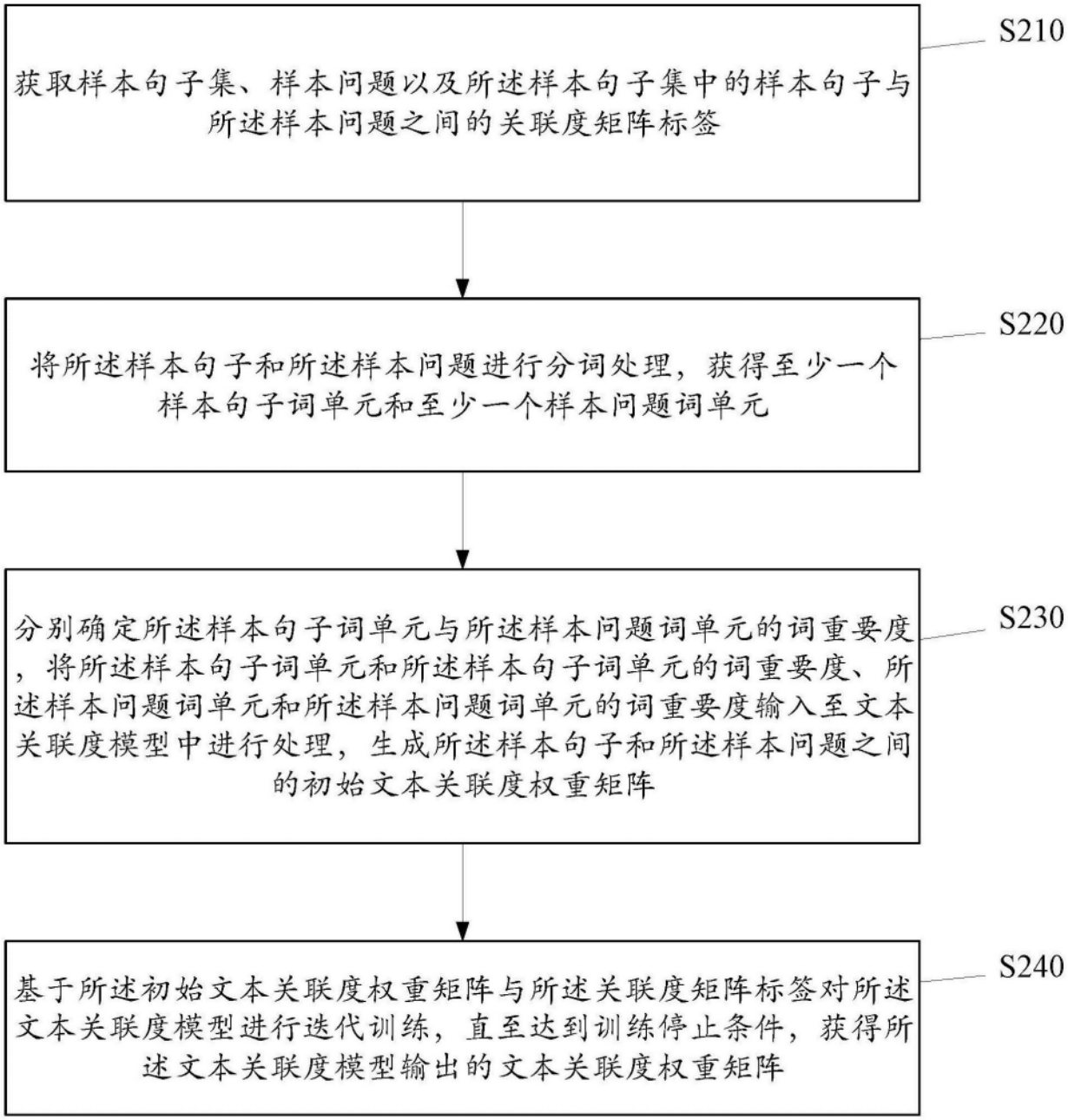

获取样本句子集、样本问题以及所述样本句子集中的样本句子与所述样本问题之间的关联度矩阵标签
S210
将所述样本句子和所述样本问题进行分词处理，获得至少一个样本句子词单元和至少一个样本问题词单元
S220
分别确定所述样本句子词单元与所述样本问题词单元的词重要度，将所述样本句子词单元和所述样本句子词单元的词重要度、所述样本问题词单元和所述样本问题词单元的词重要度输入至文本关联度模型中进行处理，生成所述样本句子和所述样本问题之间的初始文本关联度权重矩阵
S230
基于所述初始文本关联度权重矩阵与所述关联度矩阵标签对所述文本关联度模型进行迭代训练，直至达到训练停止条件，获得所述文本关联度模型输出的文本关联度权重矩阵
S240

图2

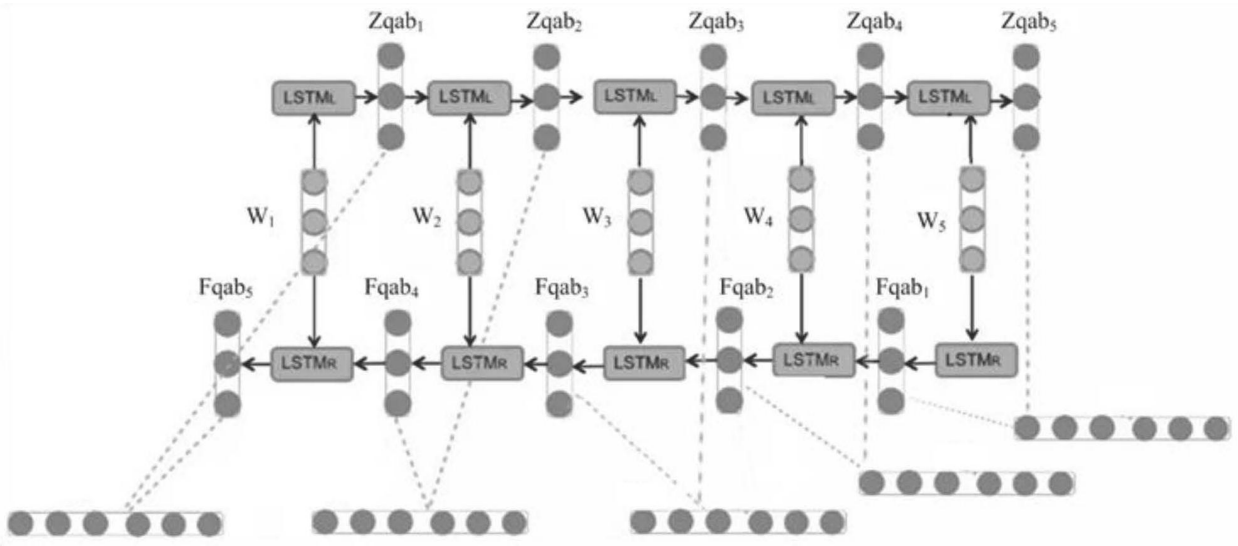

Zqab1
Zqab2
Zqab3
Zqab4
Zqab5
LSTML
W1
W2
W3
W4
W5
Fqab5
Fqab4
Fqab3
Fqab2
Fqab1
LSTMR

图3

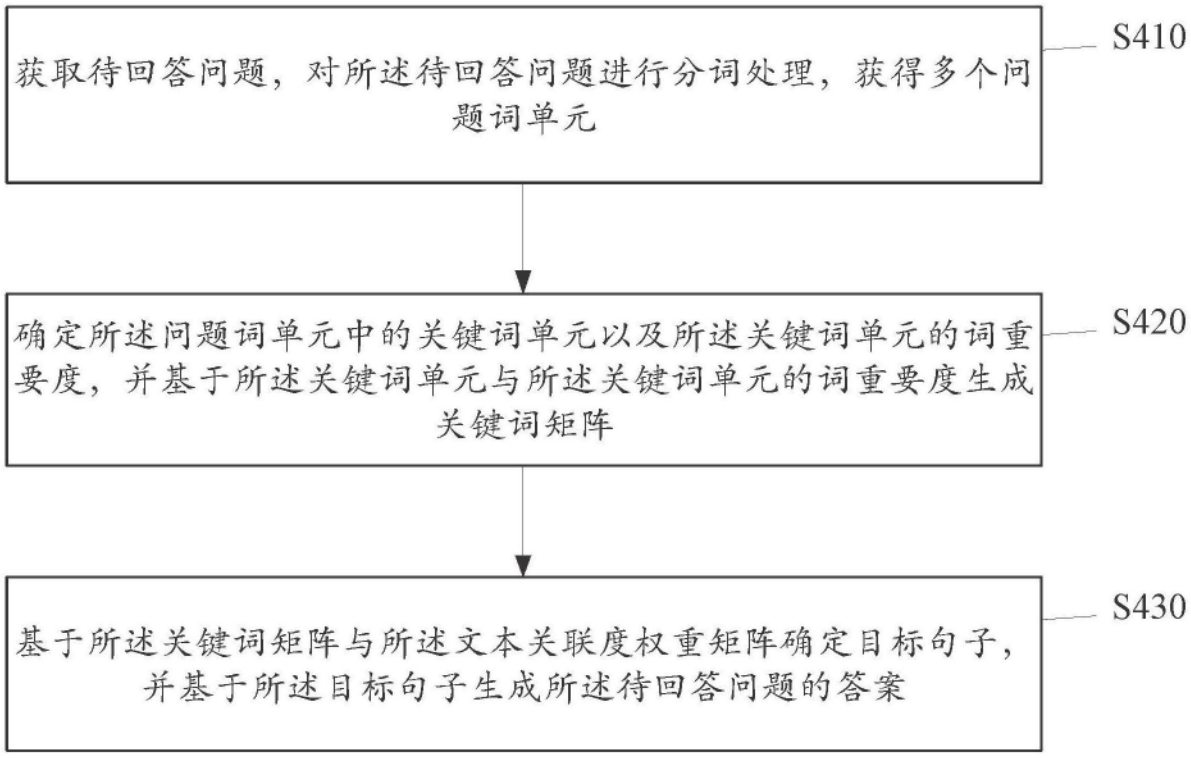

获取待回答问题，对所述待回答问题进行分词处理，获得多个问题词单元
S410
确定所述问题词单元中的关键词单元以及所述关键词单元的词重要度，并基于所述关键词单元与所述关键词单元的词重要度生成关键词矩阵
S420
基于所述关键词矩阵与所述文本关联度权重矩阵确定目标句子，并基于所述目标句子生成所述待回答问题的答案
S430

图4

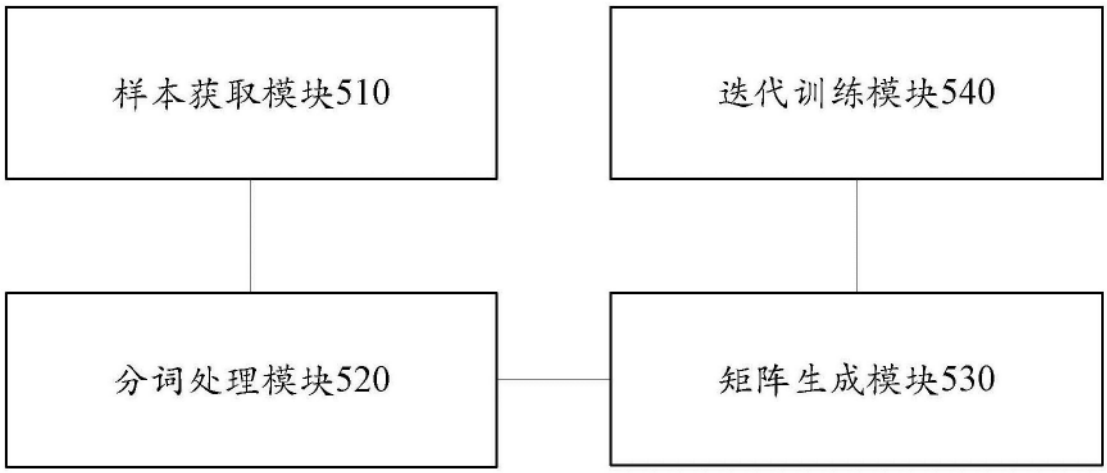

样本获取模块510
迭代训练模块540
分词处理模块520
矩阵生成模块530

图5

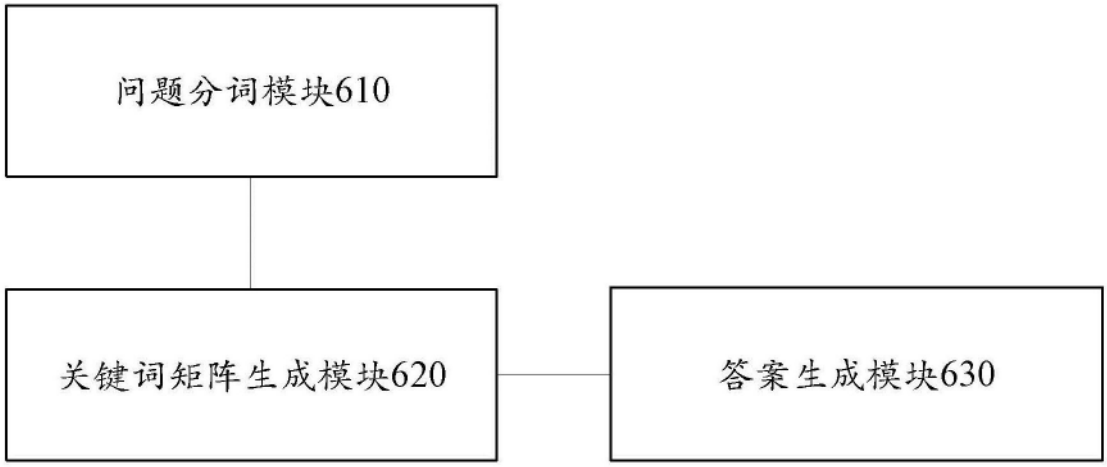

问题分词模块610
关键词矩阵生成模块620
答案生成模块630

图6