(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110781663 B
(45)授权公告日 2023.08.29

(21)申请号 201911031207.4

(22)申请日 2019.10.28

(65)同一申请的已公布的文献号
申请公布号 CN 110781663 A

(43)申请公布日 2020.02.11

(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
专利权人 成都金山互动娱乐科技有限公司

(72)发明人 陈楠 唐剑波 李长亮

(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 40/216(2020.01)
G06F 16/35(2019.01)
G06F 16/33(2019.01)
G06F 16/332(2019.01)

(56)对比文件
CN 110377710 A,2019.10.25
US 2012077178 A1,2012.03.29
CN 110222152 A,2019.09.10
CN 109766418 A,2019.05.17
CN 110347802 A,2019.10.18
CN 110348535 A,2019.10.18
CN 109816111 A,2019.05.28
CN 109558477 A,2019.04.02
CN 109933792 A,2019.06.25
WO 2012040356 A1,2012.03.29
US 2019311064 A1,2019.10.10
陈楠.基于深度学习的情感分类与评价对象判别.《中国优秀硕士学位论文全文数据库信息科技辑》.2019,全文.

审查员 冯晨露

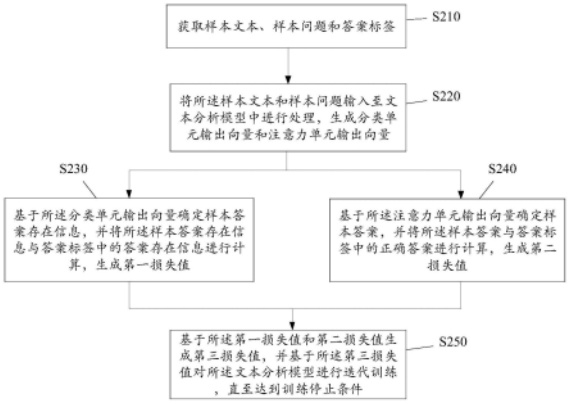
权利要求书3页 说明书12页 附图5页

(54)发明名称

文本分析模型的训练方法及装置、文本分析方法及装置

(57)摘要

本申请提供文本分析模型的训练方法及装置、文本分析方法及装置。其中,所述训练方法包括:获取样本文本、样本问题和答案标签;将样本文本和样本问题输入至文本分析模型中进行处理,生成分类单元输出向量和注意力单元输出向量;基于分类单元输出向量确定样本答案存在信息,并将样本答案存在信息与答案标签中的答案存在信息进行计算,生成第一损失值,基于注意力单元输出向量确定样本答案,并将样本答案与答案标签中的正确答案进行计算,生成第二损失值;基于第一损失值和第二损失值生成第三损失值,并基于第三损失值对文本分析模型进行迭代训练,直至达到训练停止条件。本申请所述的方法可以有效提高文本分析模型的准确率。



CN 110781663 B

1.一种文本分析模型的训练方法,其特征在于,包括:

获取样本文本、样本问题和答案标签,其中,所述答案标签包括与样本文本、样本问题相对应的答案存在信息以及正确答案;

将所述样本文本和样本问题输入至文本分析模型中进行处理,生成分类单元输出向量和注意力单元输出向量;

基于所述分类单元输出向量确定样本答案存在信息,并将所述样本答案存在信息与答案标签中的答案存在信息进行计算,生成第一损失值;

基于所述注意力单元输出向量确定样本答案,并将所述样本答案与答案标签中的正确答案进行计算,生成第二损失值;

基于所述第一损失值和第二损失值生成第三损失值,并基于所述第三损失值对所述文本分析模型进行迭代训练,直至达到训练停止条件。

2.根据权利要求1所述的文本分析模型的训练方法,其特征在于,在所述获取样本文本、样本问题和答案标签之后,还包括:

将样本文本和样本问题进行分词处理,获得词单元集合;

所述将所述样本文本和样本问题输入至文本分析模型中进行处理,生成分类单元输出向量和注意力单元输出向量, 包括:

将所述词单元集合输入至文本分析模型中进行处理,生成首个词单元的分类单元输出向量和每一个词单元的注意力单元输出向量。

3.根据权利要求2所述的文本分析模型的训练方法,其特征在于,所述将所述词单元集合输入至文本分析模型中进行处理,包括:

将所述词单元集合输入至文本分析模型的注意力单元中进行处理,生成每一个词单元的注意力单元输出向量;

将首个词单元的注意力单元输出向量输入至分类单元中进行处理,生成首个词单元的分类单元输出向量。

4.根据权利要求2所述的文本分析模型的训练方法,其特征在于,所述基于所述分类单元输出向量确定样本答案存在信息,包括:

S11、基于所述首个词单元的分类单元输出向量判断所述样本文本中是否存在所述样本问题的答案,若是,则执行步骤S12,若否,则执行步骤S13;

S12、生成存在答案标签,并将所述存在答案标签作为样本答案存在信息;

S13、生成不存在答案标签,并将所述不存在答案标签作为样本答案存在信息。

5.根据权利要求2所述的文本分析模型的训练方法,其特征在于,所述基于所述注意力单元输出向量确定样本答案,包括:

将每一个所述词单元的注意力单元输出向量进行线性与非线性处理,获得每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率;

基于所述每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率确定样本答案。

6.根据权利要求1所述的文本分析模型的训练方法,其特征在于,所述基于所述第一损失值和第二损失值生成第三损失值,包括:

确定所述第一损失值的权重值和所述第二损失值的权重值;

基于所述第一损失值的权重值以及第二损失值的权重值进行加权求和处理，生成第三损失值。

7.根据权利要求1所述的文本分析模型的训练方法，其特征在于，所述基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件，包括：

S21、判断所述第三损失值是否处于稳定状态，若是，则执行步骤S22，若否，则执行步骤S23；

S22、基于所述第三损失值对所述文本分析模型进行更新；

S23、停止训练。

8.一种文本分析方法，其特征在于，包括：

获取待分析文本和待回答问题；

将所述待分析文本和待回答问题输入至文本分析模型中进行处理，确定答案存在信息并确定所述待回答问题的答案；

其中所述文本分析模型是通过上述权利要求1-7中任意一项所述的训练方法训练得到的。

9.根据权利要求8所述的文本分析方法，其特征在于，所述确定答案存在信息，包括：

S31、判断所述待分析文本中是否存在所述待回答问题的答案，若是，则执行步骤S32，若否，则执行步骤S33；

S32、生成存在答案标签，并将所述存在答案标签作为答案存在信息；

S33、生成不存在答案标签，并将所述不存在答案标签作为答案存在信息。

10.一种文本分析模型的训练装置，其特征在于，包括：

样本获取模块，被配置为获取样本文本、样本问题和答案标签，其中，所述答案标签包括与样本文本、样本问题相对应的答案存在信息以及正确答案；

样本处理模块，被配置为将所述样本文本和样本问题输入至文本分析模型中进行处理，生成分类单元输出向量和注意力单元输出向量；

第一计算模块，被配置为基于所述分类单元输出向量确定样本答案存在信息，并将所述样本答案存在信息与答案标签中的答案存在信息进行计算，生成第一损失值；

第二计算模块，被配置为基于所述注意力单元输出向量确定样本答案，并将所述样本答案与答案标签中的正确答案进行计算，生成第二损失值；

迭代训练模块，被配置为基于所述第一损失值和第二损失值生成第三损失值，并基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件。

11.一种文本分析装置，其特征在于，包括：

获取模块，被配置为获取待分析文本和待回答问题；

处理模块，被配置为将所述待分析文本和待回答问题输入至文本分析模型中进行处理，确定答案存在信息并确定所述待回答问题的答案；

其中所述文本分析模型是通过上述权利要求1-7中任意一项所述的训练方法训练得到的。

12.一种计算设备，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1-7或者8-9任意一项所述方法的步骤。

13.一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1-7或者8-9任意一项所述方法的步骤。

文本分析模型的训练方法及装置、文本分析方法及装置

技术领域

[0001] 本申请涉及自然语言处理技术领域，特别涉及文本分析模型的训练方法及装置、文本分析方法及装置、计算设备及计算机可读存储介质。

背景技术

[0002] 自然语言处理(Natural Language Processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向，它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。

[0003] 对于自然语言处理任务，通常选用双向注意力神经网络模型模型(Bidirectional Encoder Representation from Transformers,BERT)进行处理。现有的BERT模型在进行阅读理解任务时，仅仅通过对答案起点位置和答案终点位置做位置分类，来确定待分析文本中是否存在答案以及答案具体是什么，准确性有待提高。

发明内容

[0004] 有鉴于此，本申请实施例提供了文本分析模型的训练方法及装置、文本分析方法及装置、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。

[0005] 本申请实施例公开了一种文本分析模型的训练方法，包括：

[0006] 获取样本文本、样本问题和答案标签；

[0007] 将所述样本文本和样本问题输入至文本分析模型中进行处理，生成分类单元输出向量和注意力单元输出向量；

[0008] 基于所述分类单元输出向量确定样本答案存在信息，并将所述样本答案存在信息与答案标签中的答案存在信息进行计算，生成第一损失值；

[0009] 基于所述注意力单元输出向量确定样本答案，并将所述样本答案与答案标签中的正确答案进行计算，生成第二损失值；

[0010] 基于所述第一损失值和第二损失值生成第三损失值，并基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件。

[0011] 进一步地，在所述获取样本文本、样本问题和答案标签之后，还包括：

[0012] 将样本文本和样本问题进行分词处理，获得词单元集合；

[0013] 所述将所述样本文本和样本问题输入至文本分析模型中进行处理，生成分类单元输出向量和注意力单元输出向量，包括：

[0014] 将所述词单元集合输入至文本分析模型中进行处理，生成首个词单元的分类单元输出向量和每一个词单元的注意力单元输出向量。

[0015] 进一步地，所述将所述词单元集合输入至文本分析模型中进行处理，包括：

[0016] 将所述词单元集合输入至文本分析模型的注意力单元中进行处理，生成每一个词单元的注意力单元输出向量；

[0017] 将首个词单元的注意力单元输出向量输入至分类单元中进行处理，生成首个词单

元的分类单元输出向量。

[0018] 进一步地，所述基于所述分类单元输出向量确定样本答案存在信息，包括：

[0019] S11、基于所述首个词单元的分类单元输出向量判断所述样本文本中是否存在所述样本问题的答案，若是，则执行步骤S12，若否，则执行步骤S13；

[0020] S12、生成存在答案标签，并将所述存在答案标签作为样本答案存在信息；

[0021] S13、生成不存在答案标签，并将所述不存在答案标签作为样本答案存在信息。

[0022] 进一步地，所述基于所述注意力单元输出向量确定样本答案，包括：

[0023] 将每一个所述词单元的注意力单元输出向量进行线性与非线性处理，获得每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率；

[0024] 基于所述每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率确定样本答案。

[0025] 进一步地，所述基于所述第一损失值和第二损失值生成第三损失值，包括：

[0026] 确定所述第一损失值的权重值和所述第二损失值的权重值；

[0027] 基于所述第一损失值的权重值以及第二损失值的权重值进行加权求和处理，生成第三损失值。

[0028] 进一步地，所述基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件，包括：

[0029] S21、判断所述第三损失值是否处于稳定状态，若是，则执行步骤S22，若否，则执行步骤S23；

[0030] S22、基于所述第三损失值对所述文本分析模型进行更新；

[0031] S23、停止训练。

[0032] 本申请还提供一种文本分析方法，包括：

[0033] 获取待分析文本和待回答问题；

[0034] 将所述待分析文本和待回答问题输入至文本分析模型中进行处理，确定答案存在信息并确定所述待回答问题的答案；

[0035] 其中所述文本分析模型是通过上述的训练方法训练得到的。

[0036] 进一步地，所述确定答案存在信息，包括：

[0037] S31、判断所述待分析文本中是否存在所述待回答问题的答案，若是，则执行步骤S32，若否，则执行步骤S33；

[0038] S32、生成存在答案标签，并将所述存在答案标签作为答案存在信息；

[0039] S33、生成不存在答案标签，并将所述不存在答案标签作为答案存在信息。

[0040] 本申请还提供一种文本分析模型的训练装置，包括：

[0041] 样本获取模块，被配置为获取样本文本、样本问题和答案标签；

[0042] 样本处理模块，被配置为将所述样本文本和样本问题输入至文本分析模型中进行处理，生成分类单元输出向量和注意力单元输出向量；

[0043] 第一计算模块，被配置为基于所述分类单元输出向量确定样本答案存在信息，并将所述样本答案存在信息与答案标签中的答案存在信息进行计算，生成第一损失值；

[0044] 第二计算模块，被配置为基于所述注意力单元输出向量确定样本答案，并将所述样本答案与答案标签中的正确答案进行计算，生成第二损失值；

[0045] 迭代训练模块，被配置为基于所述第一损失值和第二损失值生成第三损失值，并基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件。

[0046] 本申请还提供一种文本分析装置，包括：

[0047] 获取模块，被配置为获取待分析文本和待回答问题；

[0048] 处理模块，被配置为将所述待分析文本和待回答问题输入至文本分析模型中进行处理，确定答案存在信息并确定所述待回答问题的答案；

[0049] 其中所述文本分析模型是通过上述的训练方法训练得到的。

[0050] 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分析模型的训练方法或文本分析方法的步骤。

[0051] 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分析模型的训练方法或文本分析方法的步骤。

[0052] 本申请提供的文本分析模型的训练方法及装置，一方面通过在文本分析模型的注意力单元后设置分类单元，并利用分类单元生成样本答案存在信息，判断样本文本中是否存在样本问题的答案，再与答案标签中的答案存在信息进行对比和计算，得到第一损失值，另一方面将文本分析模型生成的样本答案与正确答案进行对比和计算，得到第二损失值，最后基于上述两个损失值加权求和得到的第三损失值对文本分析模型进行迭代训练，在对样本文本、样本问题进行特征提取、特征分析、寻找问题答案的基础上，进一步地关注了样本文本中是否存在样本问题的答案，且将判断“有无答案”以及“答案是什么”两部分特征相结合对文本分析模型进行训练，可以有效提高文本分析模型的准确率。

[0053] 本申请提供的文本分析方法及装置，在对待分析文本进行分析以寻找待回答问题答案的基础上，增加了对于待分析文本中是否存在待回答问题的答案的判断，可以有效提高阅读理解问答的准确率和效率，避免在待分析文本中不存在待回答问题的答案的情况下，依旧生成错误答案的问题形成误导。

附图说明

[0054] 图1是本申请实施例的计算设备的结构示意图；

[0055] 图2是本申请实施例的文本分析模型的训练方法的步骤流程示意图；

[0056] 图3是本申请实施例的文本分析模型的训练方法的步骤流程示意图；

[0057] 图4是本申请实施例的文本分析模型的训练方法的步骤流程示意图；

[0058] 图5是本申请实施例的文本分析方法的步骤流程示意图；

[0059] 图6是本申请实施例的文本分析方法的步骤流程示意图；

[0060] 图7是本申请实施例的文本分析模型的训练装置的结构示意图；

[0061] 图8是本申请实施例的文本分析装置的结构示意图。

具体实施方式

[0062] 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。

[0063] 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。

[0064] 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”。

[0065] 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。

[0066] 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、数字或字母等语言单元，这些单元被称为词单元。对于英文文本，字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，最小的字单元可以是一个词语、一个字、一个标点符号、一个数字等。

[0067] BERT模型：一种双向注意力神经网络模型。BERT模型可以通过左、右两侧上下文来预测当前词和通过当前句子预测下一个句子。BERT模型的目标是利用大规模无标注语料训练、获得文本的包含丰富语义信息的文本的语义表示，然后将文本的语义表示在特定NLP任务中作微调，最终应用于该NLP任务。

[0068] F1值：以单词为单位，统计预测的答案和标准答案之间的准确率和召回率，再通过下面公式技术F1值。F1＝2*R*P/(R+P)

[0069] 准确率(Precision)：P＝TP/(TP+FP)。通俗地讲，就是预测正确的正例数据占预测为正例数据的比例。

[0070] 召回率(Recall)：R＝TP/(TP+FN)。通俗地讲，就是预测为正例的数据占实际为正例数据的比例。

[0071] 在本申请中，提供了一种文本分析模型的训练方法及装置、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。

[0072] 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。

[0073] 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。

[0074] 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。

[0075] 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。

[0076] 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分析模型的训练方法的示意性流程图，包括步骤S210至步骤S250。

[0077] S210、获取样本文本、样本问题和答案标签。

[0078] 具体地，样本文本为包含有一定信息内容的书面文本，其可以是一句话、一段文字、多段文字、一篇文章或多篇文章等各种篇幅的文本，也可以是中文文本、英文文本、俄文文本等各种语言文本，本申请对此不做限制。

[0079] 样本问题为要求回答或解释的题目，既可以是与样本文本中的信息内容相关联的问题，也可以是与样本文本中的信息内容无关联的问题，本申请对此不做限制。

[0080] 答案标签包括与样本文本、样本问题相对应的答案存在信息以及正确答案。其中，答案存在信息是用于标识样本文本中是否存在样本问题的答案的信息，其可以是任何能够区分有答案或答案的标识，比如，可以为“有答案”/“无答案”或“存在答案”/“不存在答案”，也可以以“(1,0)”标识有答案/“(0,1)”标识无答案，或是其他方式均可，本申请对此不做限制。正确答案是样本问题的正确答案，需要说明的是，正确答案通常为样本文本中的内容，在样本文本中不存在样本问题的答案的情况下，正确答案为空，但是在样本问题可以根据公知常识得到正确答案的情况下，正确答案可以为根据公知常识得到的正确答案，本申请对此不做限制。

[0081] 例如，若样本文本包括“‘落霞与孤鹜齐飞，秋水共长天一色’出自王勃所作的《滕王阁序》”，样本问题包括“《滕王阁序》的作者是谁?”则答案标签包括答案存在信息“有答案”和正确答案“王勃”。若样本文本包括“《唐诗三百首》是一部流传很广的唐诗选集”，样本问题包括“李白是哪朝诗人?”则答案标签包括“无答案”，正确答案可以为空，也可以为“唐朝”。

[0082] S220、将所述样本文本和样本问题输入至文本分析模型中进行处理，生成分类单元输出向量和注意力单元输出向量。

[0083] 具体地，所述文本分析模型为BERT模型，且文本分析模型中依次包括注意力单元和分类单元。

[0084] 进一步地，将样本文本和样本问题进行分词处理，获得词单元集合，再将所述词单元集合输入至文本分析模型中进行处理，生成首个词单元的分类单元输出向量和每一个词单元的注意力单元输出向量。

[0085] 更进一步地，将所述词单元集合输入至文本分析模型的注意力单元中进行处理，生成每一个词单元的注意力单元输出向量；再将首个词单元的注意力单元输出向量输入至分类单元中进行处理，生成首个词单元的分类单元输出向量。

[0086] 具体地，注意力单元中既可以仅仅包括一个注意力层，也可以包括两个或多个注意力层，注意力单元的输出向量即为注意力单元中最后一个注意力层的输出向量。例如，假

设注意力单元包括12个注意力层，那么注意力单元的输出向量即为第12个注意力层的输出向量。

[0087] 分类单元包括一个二分类层，用于判断样本文本中是否存在样本问题的答案，在样本文本中存在样本问题的答案的情况下，其输出为(1,0)，在样本文本中不存在样本问题的情况下，其输出为(0,1)。

[0088] 例如，假设文本分析模型中的注意力单元包括3个注意力层，样本文本包括“‘落霞与孤鹜齐飞，秋水共长天一色’出自王勃所作的《滕王阁序》”，样本问题包括“《滕王阁序》的作者是谁?”对上述样本文本和样本问题进行分词处理，得到词单元集合[CLS、SEP、落、霞、……是、谁、SEP]，其中，CLS为句首标志符号，SEP为分句标志符号，将上述词单元集合进行嵌入处理后输入至文本分析模型注意力单元的第一注意力层中进行特征提取，生成第一注意力层的输出向量[A_{11}、A_{12}、A_{13}、A_{14}……A_{142}、A_{143}]，将上述第一注意力层的输出向量输入至第二注意力层中进行特征提取，生成第二注意力层的输出向量[A_{21}、A_{22}、A_{23}、A_{24}……A_{242}、A_{243}]，将上述第二注意力层的输出向量输入至第三注意力层中进行特征提取，生成第三注意力层的输出向量[A_{31}、A_{32}、A_{33}、A_{34}……A_{342}、A_{343}]，并将上述第三注意力层的输出向量作为注意力单元的输出向量。将词单元集合中首个词单元“CLS”第三注意力层的输出向量A_{31}输入至分类单元中进行处理，得到分类单元的输出向量B_1。

[0089] 分类单元及二分类层的设置，可以对样本文本中是否存在样本问题进行准确的判断，辅助提升模型的准确率。

[0090] S230、基于所述分类单元输出向量确定样本答案存在信息，并将所述样本答案存在信息与答案标签中的答案存在信息进行计算，生成第一损失值。

[0091] 具体地，样本答案存在信息是基于分类单元的输出向量得到的用于标识样本文本中是否存在样本问题的答案的信息，其可以是任何能够区分有答案或答案的标识，本申请对此不做限制。

[0092] 具体地，可以将所述样本答案存在信息与答案标签中的样本答案存在信息进行对比，通过损失函数计算损失值，并将上述损失值作为第一损失值。

[0093] 在实际应用中，损失函数可以为如分类交叉熵、最大熵函数等，本申请对此不做限制。

[0094] 通过损失函数计算第一损失值，可以在训练过程中明确模型分析得出的有无答案的情况与真实的有无答案的情况之间的差异，并根据差异调整模型以提高模型的准确率。

[0095] 进一步地，所述步骤S230包括步骤S310至步骤S330，如图3所示。

[0096] S310、基于所述首个词单元的分类单元输出向量判断所述样本文本中是否存在所述样本问题的答案，若是，则执行步骤S320，若否，则执行步骤S330。

[0097] S320、生成存在答案标签，并将所述存在答案标签作为样本答案存在信息。

[0098] S330、生成不存在答案标签，并将所述不存在答案标签作为样本答案存在信息。

[0099] 具体地，首个词单元的分类单元输出向量包括(1,0)和(0,1)两种，那么在首个词单元的分类单元输出向量为(1,0)的情况下，样本文本中存在样本问题的答案，则可以生成存在答案标签“有答案”，并将上述存在答案标签作为样本答案存在信息，在首个词单元的分类单元输出向量为(0,1)的情况下，样本文本中不存在样本问题的答案，则可以生成不存在答案标签“无答案”，并将上述不存在答案标签作为样本答案存在信息。

[0100] S240、基于所述注意力单元输出向量确定样本答案，并将所述样本答案与答案标签中的正确答案进行计算，生成第二损失值。

[0101] 进一步地，可以将每一个所述词单元的注意力单元输出向量进行线性与非线性处理，获得每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率；再基于所述每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率确定样本答案。

[0102] 具体地，在获得每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率之后，以作为样本答案开始位置概率最高的词单元与作为样本答案结束位置概率最高的词单元之间的内容作为样本答案。

[0103] 例如，假设样本文本“‘落霞与孤鹜齐飞，秋水共长天一色’出自王勃所作的《滕王阁序》”中每一个词单元作为答案开始位置的概率分别为$[x_1, x_2, x_3 \cdots\cdots x_{30}]$，每一个词单元作为答案结束位置的概率分别为$[y_1, y_2, y_3 \cdots\cdots y_{30}]$，其中，答案开始位置的概率中，$x_{19}$概率值最大，答案结束位置的概率中，$y_{20}$概率值最大，则样本答案为“王勃”。

[0104] 具体地，可以将所述样本答案与答案标签中的正确答案进行对比，通过损失函数计算损失值，并将上述损失值作为第二损失值。

[0105] 在实际应用中，损失函数可以为如分类交叉熵、最大熵函数等，本申请对此不做限制。

[0106] 例如，假设样本答案为“王勃”，答案标签中的正确答案为“作者是王勃”，将上述样本答案和答案标签中的正确答案进行最大熵损失函数的计算，得到损失值为0.1，则0.1即为第二损失值。

[0107] 通过损失函数计算第二损失值，可以在训练过程中明确模型分析得出的答案与标准答案之间的差异，并根据差异调整模型以提高模型的准确率。

[0108] S250、基于所述第一损失值和第二损失值生成第三损失值，并基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件。

[0109] 进一步地，确定所述第一损失值的权重值和所述第二损失值的权重值；基于所述第一损失值的权重值以及第二损失值的权重值进行加权求和处理，生成第三损失值。

[0110] 需要说明的是，对于第一损失值与第二损失值的权重值可以通过训练得到，且第一损失值的权重值与第二损失值的权重值之和为1。

[0111] 更进一步地，所述步骤S250还包括步骤S410至步骤S430，如图4所示。

[0112] S410、判断所述第三损失值是否处于稳定状态，若是，则执行步骤S420，若否，则执行步骤S430。

[0113] S420、基于所述第三损失值对所述文本分析模型进行更新。

[0114] S430、停止训练。

[0115] 具体地，判断第三损失值是否处于稳定状态的条件可以为判断第三损失值是否趋于稳定，若第三损失值的波动仍然较大，则基于上述第三损失值对文本分析模型进行更新，若第三损失值已趋于稳定，则停止训练。

[0116] 更为具体地，可以将当前次训练得到的第三损失值与上一次训练得到的第三损失值相比，若当前次训练得到的第三损失值与上一次训练得到的第三损失值之间的差值大于预设差值，则基于当前次训练得到的第三损失值对文本分析模型进行更新，若当前次训练

得到的第三损失值与上一次训练得到的第三损失值之间的差值小于预设差值，则停止训练。

[0117] 例如，假设将包括有多个样本文本、样本问题和答案标签的样本集合输入至文本分析模型中进行训练，预先设置在相邻两次训练得到的第三损失值之间的差值小于0.1的情况下，停止训练。将样本集合输入至文本分析模型中后，第一次训练得到的第三损失值为0.60，第二次训练得到的损失值为0.40，与第一次训练得到的第三损失值之间的差值为0.20，大于0.10，继续训练，第三次训练得到的第三损失值为0.30，与第二次训练得到的第三损失值之间的差值为0.10，继续训练，第四次训练得到的第三损失值为0.25，与第三次训练得到的第三损失值之间的差值为0.05，大于0.10，停止训练。

[0118] 另外，判断第三损失值是否处于稳定状态的条件还可以为判断第三损失值是否小于预设损失值阈值，若第三损失值大于或等于预设损失值阈值，则判断第三损失值未处于稳定状态，并基于第三损失值对文本分析模型进行更新和训练，若第三损失值小于预设损失值阈值，则判断第三损失值处于稳定状态，停止更新和训练。或是以其他方式判断第三损失值是否存于稳定状态均可，本申请对此不做限制。

[0119] 下面结合具体的例子对本实施例进行进一步说明。

[0120] 例如，假设样本文本包括“故宫是中国明清两代的皇家宫殿，占地七十二万平方米”，样本问题包括“故宫占地面积为多少?”，答案标签包括答案存在信息“有答案”和正确答案“七十二万平方米”。

[0121] 将样本文本和样本问题进行分词处理，获得词单元集合[CLS、故、宫、是、中、国……多、少、SEP]。

[0122] 假设文本分析模型中的注意力单元包括6层注意力层，将上述词单元集合输入至文本分析模型中，首先经过注意力单元的处理后，生成词单元集合中每一个词单元的注意力单元输出向量[C_1、C_2、C_3、C_4……C_{35}、C_{36}]，将首个词单元的注意力单元输出向量C_1输入至分类单元中进行处理，生成分类单元输出向量(1,0)。

[0123] 基于上述首个词单元的分类单元输出向量(1,0)，得到样本答案存在信息为“有答案”，将其与答案标签中的答案存在信息进行损失函数的计算，得到第一损失值为0.05。

[0124] 将上述每个词单元的注意力单元输出向量进行线性映射与非线性变换处理，获得每一个词单元作为答案开始位置的概率[0.10,0.33.0.25.0.19,0.15,0.21,0.42,0.13,0.32,0.11,0.22,0.23,0.13,0.16,0.20,0.19,0.67,0.39,0.54,0.03.0.20.0.19,0.12,0.21,0.43,0.13,0.32,0.17,0.27,0.23,0.23,0.10,0.24,0.19,0.08,0.02]和每一个词单元作为答案结束位置的概率[0.05,0.13.0.25.0.24,0.10,0.13,0.12,0.23,0.30,0.11,0.14,0.19,0.14,0.28,0.20,0.11,0.17,0.27,0.33,0.09.0.15.0.49,0.32,0.28,0.70,0.42,0.22,0.07,0.25,0.23,0.22,0.09,0.16,0.16,0.10,0.10]。

[0125] 可以看出，第17个词单元作为答案开始位置的概率最高，第25个词单元作为答案结束位置的概率最高，则样本答案为“占地七十二万平方米”。

[0126] 将样本答案“占地七十二万平方米”，与正确答案“七十二万平方米”进行损失函数计算，得到第二损失值为0.2。

[0127] 假设第一损失值与第二损失值的权重值均为0.5，则第三损失值为0.1×0.5+0.2×0.5＝0.15。

[0128] 假设预设损失值阈值为0.10，第三损失值大于预设损失值阈值，对文本分析模型进行更新并迭代训练。

[0129] 本实施例提供的文本分析模型的训练方法，一方面通过在文本分析模型的注意力单元后设置分类单元，并利用分类单元生成样本答案存在信息，判断样本文本中是否存在样本问题的答案，再与答案标签中的答案存在信息进行对比和计算，得到第一损失值，另一方面将文本分析模型生成的样本答案与正确答案进行对比和计算，得到第二损失值，最后基于上述两个损失值加权求和得到的第三损失值对文本分析模型进行迭代训练，在对样本文本、样本问题进行特征提取、特征分析、寻找问题答案的基础上，进一步地关注了样本文本中是否存在样本问题的答案，且将判断“有无答案”以及“答案是什么”两部分相结合对文本分析模型进行训练，对于base版的文本分析模型，F1值提高了2.6%，对于large版的文本分析模型，F1值提高了0.7%，可以有效提高文本分析模型的准确率。

[0130] 如图5所示，一种文本分析方法，其特征在于，包括步骤S510至步骤S520。

[0131] S510、获取待分析文本和待回答问题。

[0132] S520、将所述待分析文本和待回答问题输入至文本分析模型中进行处理，确定答案存在信息并确定所述待回答问题的答案。

[0133] 其中所述文本分析模型是通过上述实施例所述的训练方法训练得到的。

[0134] 进一步地，所述步骤S520，还包括步骤S610至S630，如图6所示。

[0135] S610、判断所述待分析文本中是否存在所述待回答问题的答案，若是，则执行步骤S620，若否，则执行步骤S630。

[0136] S620、生成存在答案标签，并将所述存在答案标签作为答案存在信息。

[0137] S630、生成不存在答案标签，并将所述不存在答案标签作为答案存在信息。

[0138] 下面结合具体的例子对本实施例进行进一步说明。

[0139] 例如，假设获取到待分析文本包括“丝绸之路是古代中国与外国交通贸易和文化交往的通道”，待回答问题包括“丝绸之路起源于哪个朝代?”。

[0140] 将待分析文本和待回答问题进行分词处理，生成词单元集合[CLS、丝、绸……朝、代、SEP]。

[0141] 将上述词单元集合输入至文本分析模型中，经过注意力单元的处理，生成每一个词单元的注意力单元输出向量[E_1、E_2、E_3、E_4……E_{37}、E_{38}]，将上述每一个词单元的注意力单元输出向量进行线性映射与非线性变换处理，得到第7个词单元作为答案开始位置的概率最高，第10个词单元作为答案结束位置的概率最高，生成答案“古代中国”。将首个词单元的注意力单元输出向量E_1输入至分类单元中进行处理，生成分类单元输出向量(0,1)，得到答案存在信息为“无答案”。

[0142] 本实施例提供的文本分析方法及装置，在对待分析文本进行分析寻找待回答问题答案的基础上，增加了对于待分析文本中是否存在待回答问题的答案的判断，可以有效提高阅读理解问答的准确率和效率，避免在待分析文本中不存在待回答问题的答案的情况下，依旧生成错误答案的问题。

[0143] 如图7所示，一种文本分析模型的训练装置，包括：

[0144] 样本获取模块710，被配置为获取样本文本、样本问题和答案标签；

[0145] 样本处理模块720，被配置为将所述样本文本和样本问题输入至文本分析模型中

进行处理，生成分类单元输出向量和注意力单元输出向量；

[0146] 第一计算模块730，被配置为基于所述分类单元输出向量确定样本答案存在信息，并将所述样本答案存在信息与答案标签中的答案存在信息进行计算，生成第一损失值；

[0147] 第二计算模块740，被配置为基于所述注意力单元输出向量确定样本答案，并将所述样本答案与答案标签中的正确答案进行计算，生成第二损失值；

[0148] 迭代训练模块750，被配置为基于所述第一损失值和第二损失值生成第三损失值，并基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件。

[0149] 可选地，所述文本分析模型的训练装置，还包括：

[0150] 样本分词模块，将样本文本和样本问题进行分词处理，获得词单元集合；

[0151] 所述样本处理模块720，进一步被配置为：

[0152] 将所述词单元集合输入至文本分析模型中进行处理，生成首个词单元的分类单元输出向量和每一个词单元的注意力单元输出向量。

[0153] 可选地，所述样本处理模块720，进一步被配置为：

[0154] 将所述词单元集合输入至文本分析模型的注意力单元中进行处理，生成每一个词单元的注意力单元输出向量；

[0155] 将首个词单元的注意力单元输出向量输入至分类单元中进行处理，生成首个词单元的分类单元输出向量。

[0156] 可选地，所述第一计算模块730，进一步被配置为：

[0157] 第一判断模块，被配置为基于所述首个词单元的分类单元输出向量判断所述样本文本中是否存在所述样本问题的答案，若是，则执行第一生成模块，若否，则执行第二生成模块；

[0158] 第一生成模块，被配置为生成存在答案标签，并将所述存在答案标签作为样本答案存在信息；

[0159] 第二生成模块，被配置为生成不存在答案标签，并将所述不存在答案标签作为样本答案存在信息。

[0160] 可选地，所述第二计算模块740，进一步被配置为：

[0161] 将每一个所述词单元的注意力单元输出向量进行线性与非线性处理，获得每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率；

[0162] 基于所述每一个词单元作为样本答案开始位置的概率和作为样本答案结束位置的概率确定样本答案。

[0163] 可选地，所述迭代训练模块750，进一步被配置为：

[0164] 确定所述第一损失值的权重值和所述第二损失值的权重值；

[0165] 基于所述第一损失值的权重值以及第二损失值的权重值进行加权求和处理，生成第三损失值。

[0166] 可选地，所述迭代训练模块750，进一步被配置为：

[0167] 第二判断模块，被配置为判断所述第三损失值是否处于稳定状态，若是，则执行更新模块，若否，则执行停止模块；

[0168] 更新模块，被配置为基于所述第三损失值对所述文本分析模型进行更新；

[0169] 停止模块，被配置为停止训练。

[0170] 本实施例提供的文本分析模型的训练装置，一方面通过在文本分析模型的注意力单元后设置分类单元，并利用分类单元生成样本答案存在信息，判断样本文本中是否存在样本问题的答案，再与答案标签中的答案存在信息进行对比和计算，得到第一损失值，另一方面将文本分析模型生成的样本答案与正确答案进行对比和计算，得到第二损失值，最后基于上述两个损失值加权求和得到的第三损失值对文本分析模型进行迭代训练，在对样本文本、样本问题进行特征提取、特征分析、寻找问题答案的基础上，进一步地关注了样本文本中是否存在样本问题的答案，且将判断“有无答案”以及“答案是什么”两部分相结合对文本分析模型进行训练，可以有效提高文本分析模型的准确率。

[0171] 如图8所示，一种文本分析装置，包括：

[0172] 获取模块810，被配置为获取待分析文本和待回答问题；

[0173] 处理模块820，被配置为将所述待分析文本和待回答问题输入至文本分析模型中进行处理，确定答案存在信息并确定所述待回答问题的答案；

[0174] 其中所述文本分析模型是通过上述的训练方法训练得到的。

[0175] 可选地，所述处理模块820，进一步被配置为：

[0176] 第三判断模块，被配置为判断所述待分析文本中是否存在所述待回答问题的答案，若是，则执行第三生成模块，若否，则执行第四生成模块；

[0177] 第三生成模块，被配置为生成存在答案标签，并将所述存在答案标签作为答案存在信息；

[0178] 第四生成模块，被配置为生成不存在答案标签，并将所述不存在答案标签作为答案存在信息。

[0179] 本实施例提供的文本分析装置，在对待分析文本进行分析寻找待回答问题答案的基础上，增加了对于待分析文本中是否存在待回答问题的答案的判断，可以有效提高阅读理解问答的准确率和效率，避免在待分析文本中不存在待回答问题的答案的情况下，依旧生成错误答案的问题。

[0180] 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：

[0181] 获取样本文本、样本问题和答案标签；

[0182] 将所述样本文本和样本问题输入至文本分析模型中进行处理，生成分类单元输出向量和注意力单元输出向量；

[0183] 基于所述分类单元输出向量确定样本答案存在信息，并将所述样本答案存在信息与答案标签中的答案存在信息进行计算，生成第一损失值；

[0184] 基于所述注意力单元输出向量确定样本答案，并将所述样本答案与答案标签中的正确答案进行计算，生成第二损失值；

[0185] 基于所述第一损失值和第二损失值生成第三损失值，并基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件。

[0186] 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分析模型的训练方法或文本分析方法的步骤。

[0187] 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分析模型的训练方法或文本分析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分析模型的训练方法或文本分析方法的技术方案的描述。

[0188] 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。

[0189] 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。

[0190] 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。

[0191] 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。

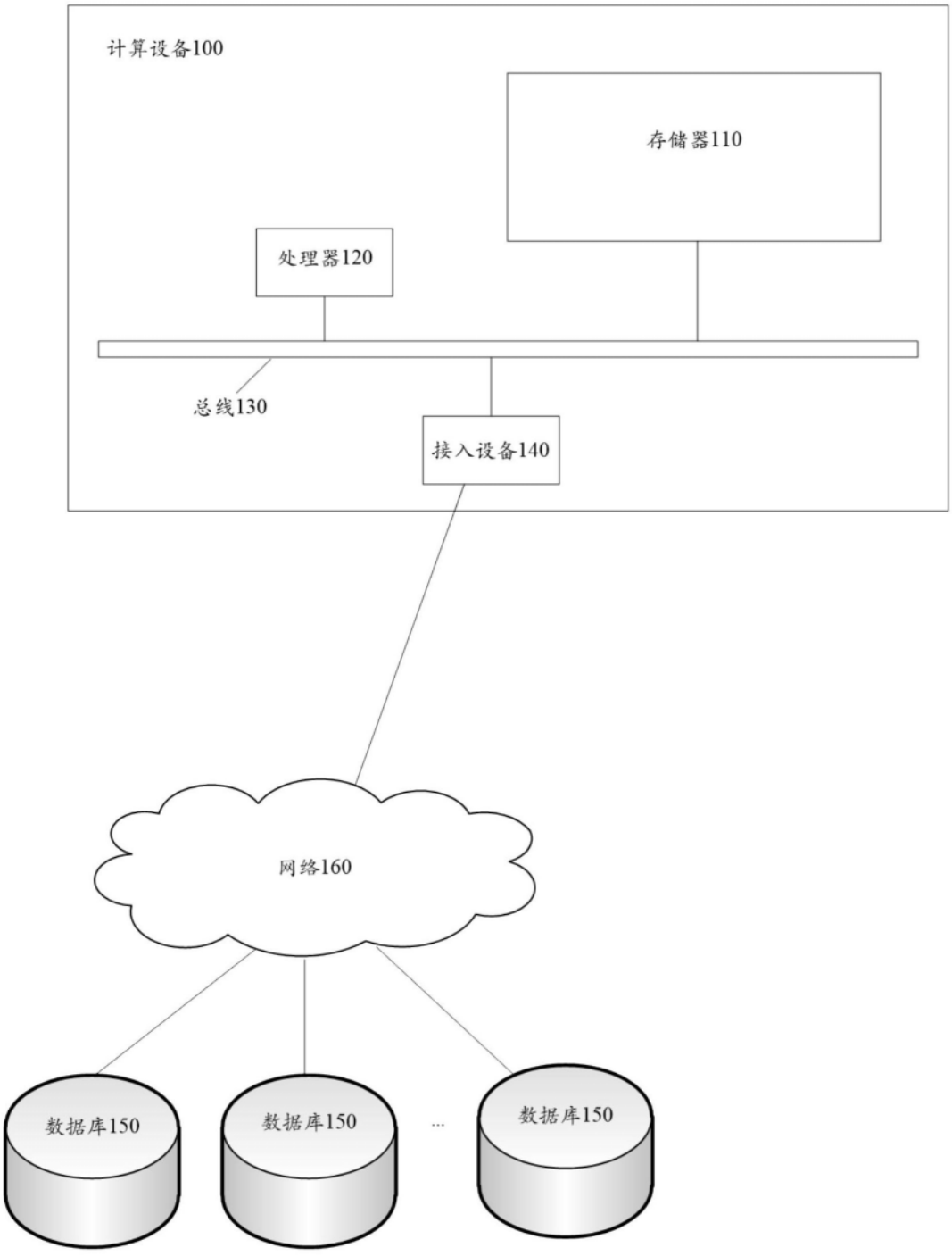

计算设备100
存储器110
处理器120
总线130
接入设备140
网络160
数据库150
数据库150
...
数据库150

图1

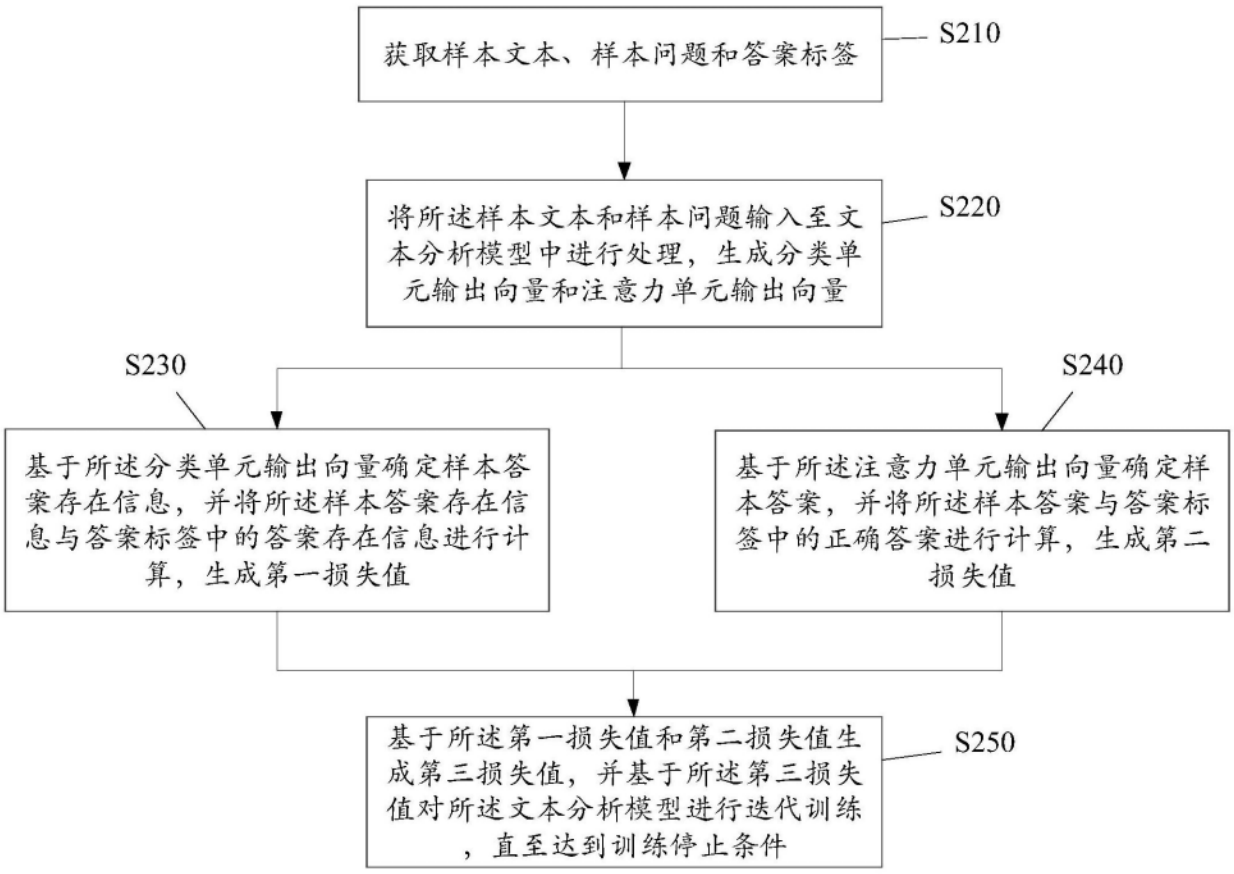

获取样本文本、样本问题和答案标签
S210
将所述样本文本和样本问题输入至文本分析模型中进行处理，生成分类单元输出向量和注意力单元输出向量
S220
S230
S240
基于所述分类单元输出向量确定样本答案存在信息，并将所述样本答案存在信息与答案标签中的答案存在信息进行计算，生成第一损失值
基于所述注意力单元输出向量确定样本答案，并将所述样本答案与答案标签中的正确答案进行计算，生成第二损失值
基于所述第一损失值和第二损失值生成第三损失值，并基于所述第三损失值对所述文本分析模型进行迭代训练，直至达到训练停止条件
S250

图2

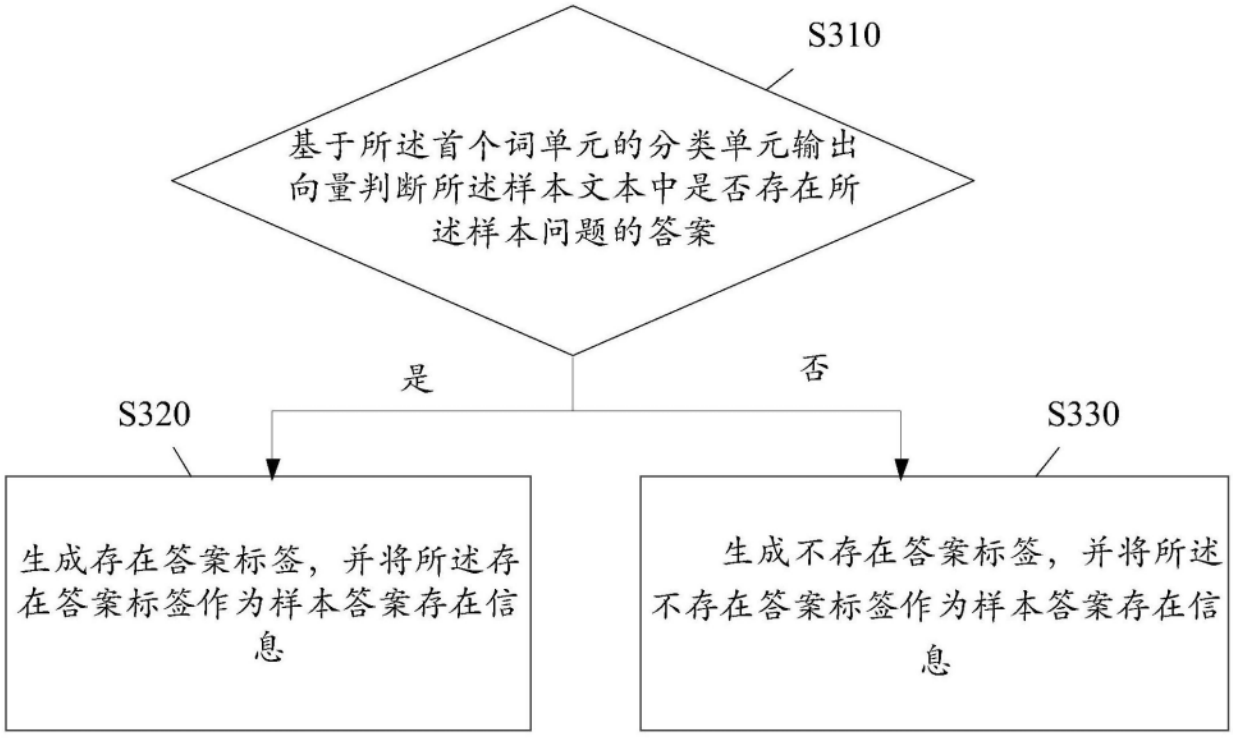

S310
基于所述首个词单元的分类单元输出向量判断所述样本文本中是否存在所述样本问题的答案
是
否
S320
S330
生成存在答案标签，并将所述存在答案标签作为样本答案存在信息
生成不存在答案标签，并将所述不存在答案标签作为样本答案存在信息

图3

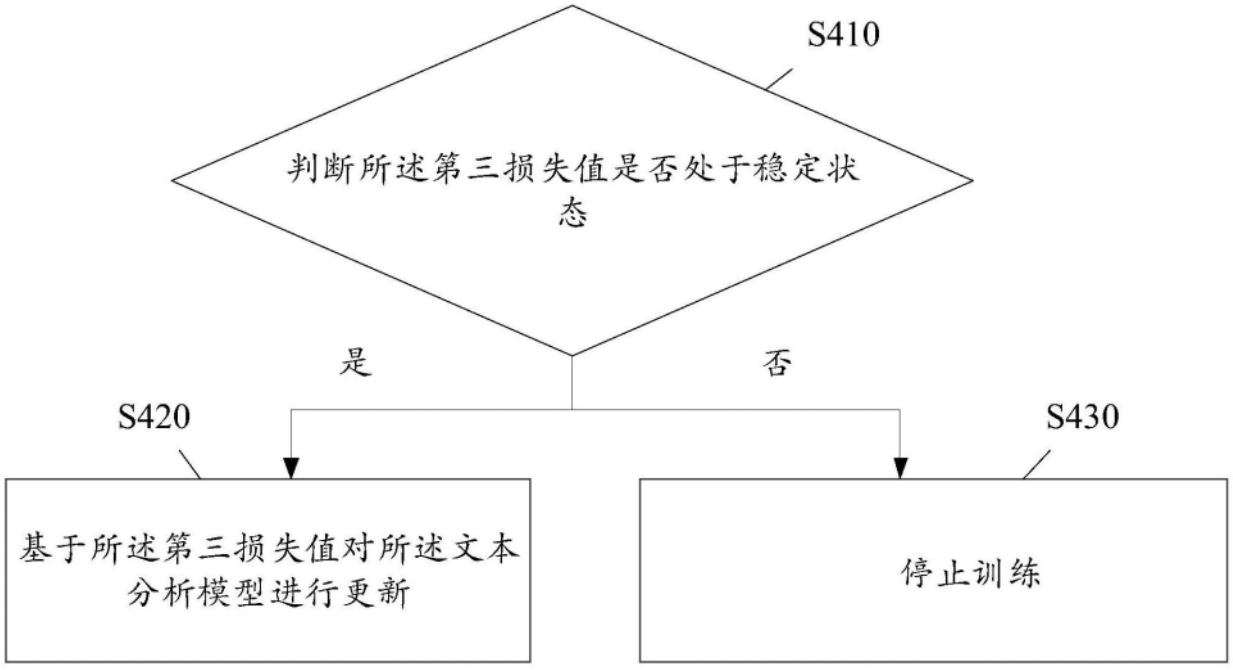

S410
判断所述第三损失值是否处于稳定状态
是
否
S420
S430
基于所述第三损失值对所述文本分析模型进行更新
停止训练

图4

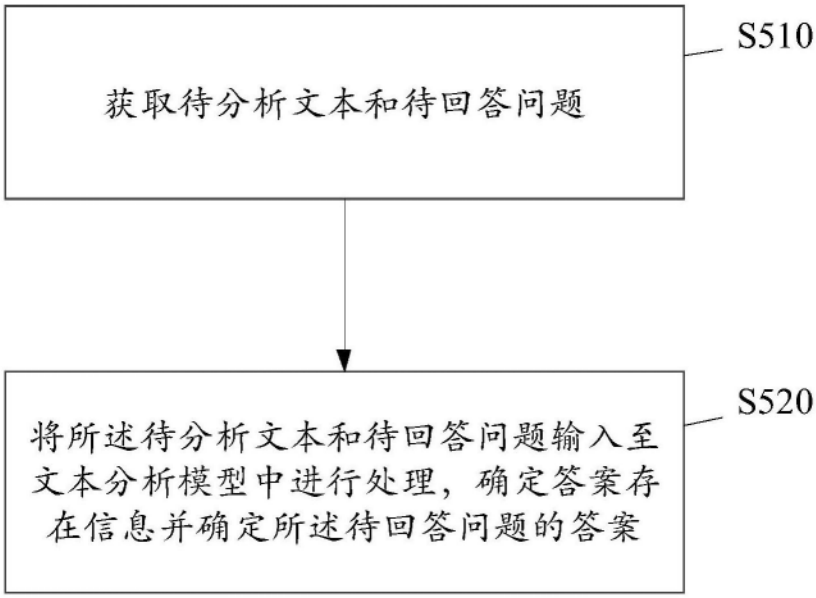

S510
获取待分析文本和待回答问题
S520
将所述待分析文本和待回答问题输入至文本分析模型中进行处理，确定答案存在信息并确定所述待回答问题的答案

图5

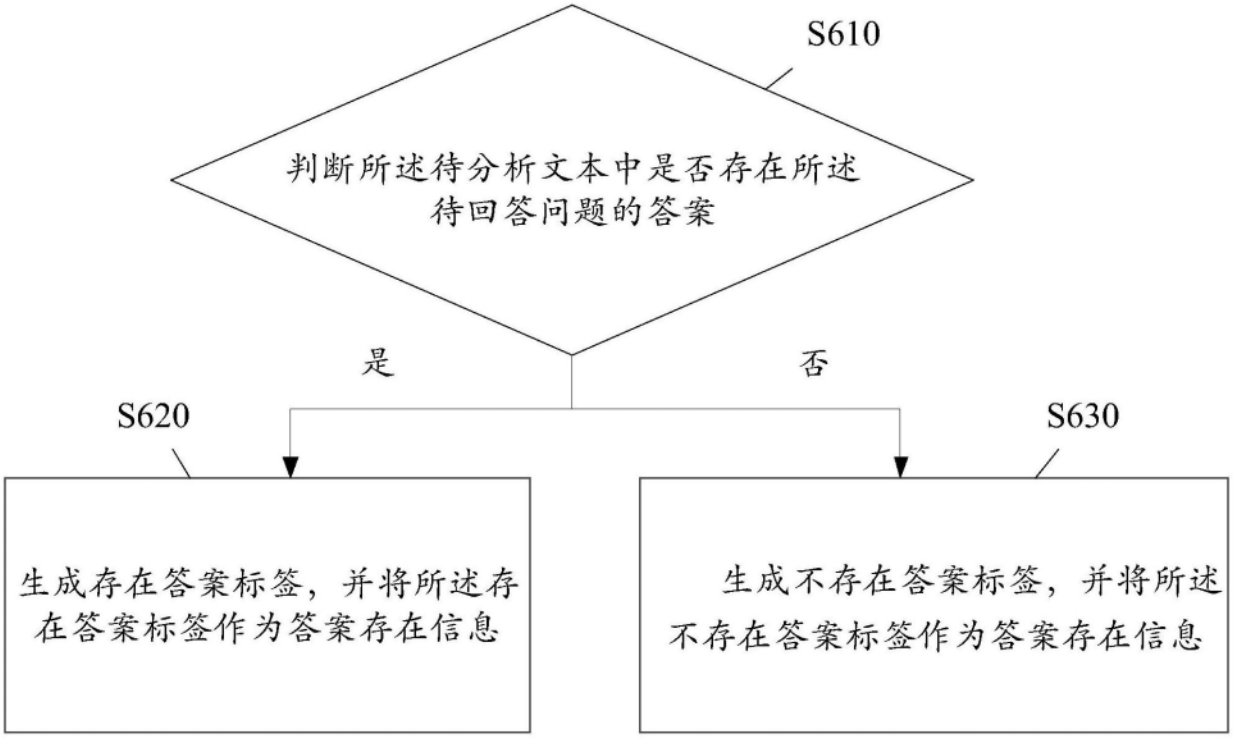

S610
判断所述待分析文本中是否存在所述待回答问题的答案
是
否
S620
生成存在答案标签，并将所述存在答案标签作为答案存在信息
S630
生成不存在答案标签，并将所述不存在答案标签作为答案存在信息

图6

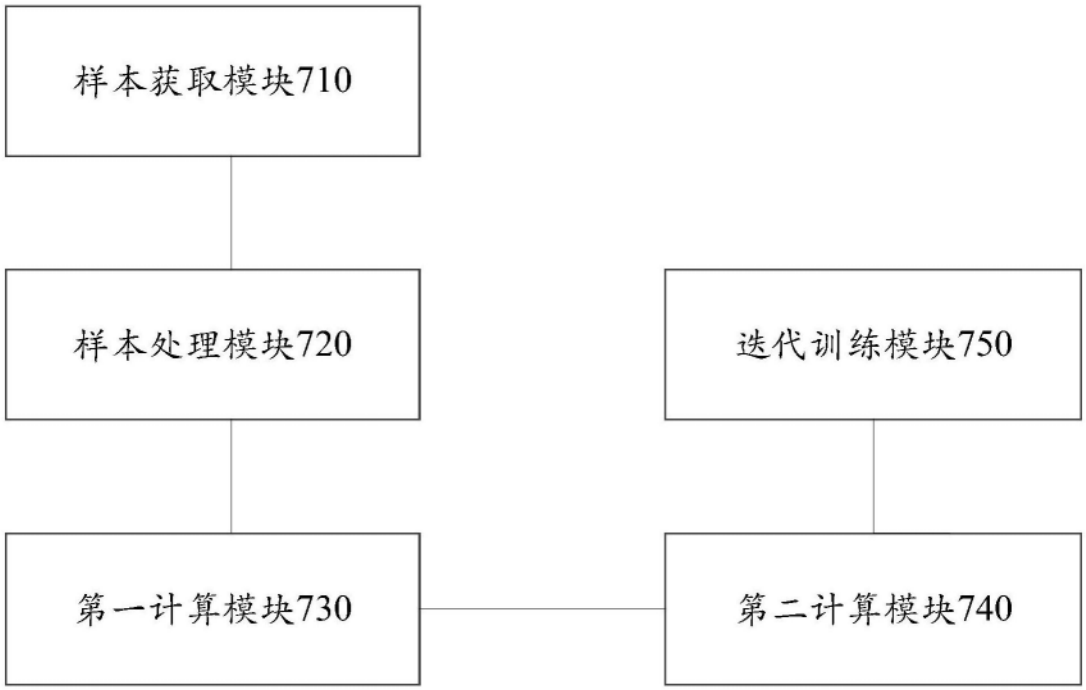

样本获取模块710
样本处理模块720
迭代训练模块750
第一计算模块730
第二计算模块740

图7


获取模块810
处理模块820

图8