(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118779647 A
(43)申请公布日 2024.10.15

(21)申请号 202310347623.5

(22)申请日 2023.04.03

(71)申请人 株式会社理光
地址 日本东京都

(72)发明人 李宏宇 董滨 姜珊珊

(74)专利代理机构 北京银龙知识产权代理有限公司 11243
专利代理师 姜精斌

(51)Int.Cl.
G06F 18/214(2023.01)
G06N 3/0455(2023.01)
G06N 3/08(2023.01)
G06F 16/33(2019.01)
G06F 16/332(2019.01)

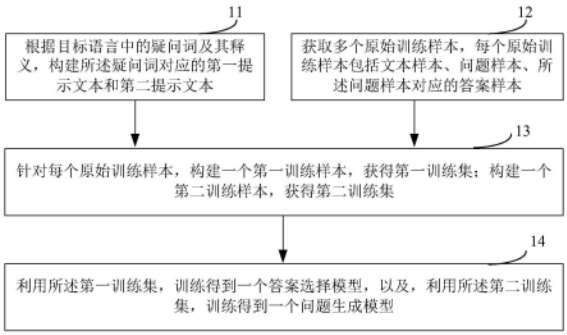
权利要求书2页 说明书14页 附图5页

(54)发明名称

模型训练方法、装置及存储介质

(57)摘要

本申请提供了一种模型训练方法、装置及存储介质。本申请实施例在模型训练过程中，将疑问词对应的第一提示文本和文本样本作为答案选择模型的输入，使得答案选择模型输出特定类型的潜在答案，另外，还将所述疑问词对应的第二提示文本、文本样本和答案样本作为问题生成模型的输入，从而通过包含有同一疑问词及其释义的提示文本，将答案选择模型和问题生成模型关联起来，使得问题生成模型能够生成与所选择的潜在答案更相关的问题，进而能够改善上述模型所生成的问题答案对的性能。



CN 118779647 A

1.一种模型训练方法,其特征在于,包括:

根据目标语言中的疑问词及其释义,构建所述疑问词对应的第一提示文本和第二提示文本,其中,所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案,所述第二提示文本用于提示生成包含所述疑问词的问题;

获取多个原始训练样本,每个原始训练样本包括文本样本、问题样本、所述问题样本对应的答案样本,所述问题样本包含有疑问词;

针对每个原始训练样本,利用所述原始训练样本的文本样本、答案样本、问题样本中的疑问词对应的第一提示文本,构建一个第一训练样本,获得包括多个所述第一训练样本的第一训练集;以及,利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本,构建一个第二训练样本,获得包括多个所述第二训练样本的第二训练集;

利用所述第一训练集,训练得到一个答案选择模型,以及,利用所述第二训练集,训练得到一个问题生成模型。

2.如权利要求1所述的方法,其特征在于,所述第一训练集还包括至少一个第三训练样本,所述第三训练样本按照以下方式构建:

针对第一文本样本,确定所述第一文本样本所对应的所有问题样本中均不包含的第一疑问词;

利用所述第一疑问词对应的第一提示文本、空白答案和所述第一文本样本,构建一个所述第三训练样本。

3.如权利要求1或2所述的方法,其特征在于,所述利用所述第一训练集,训练得到一个答案选择模型,以及,利用所述第二训练集,训练得到一个问题生成模型,包括:

将所述第一训练集中每个训练样本的第一提示文本和文本样本输入至答案选择模型,以所述答案选择模型输出对应的答案样本为目标,对所述答案选择模型进行训练,获得训练好的所述答案选择模型;

将所述第二训练集中每个训练样本的第二提示文本、文本样本和答案样本输入至问题生成模型,以所述问题生成模型输出对应的问题样本为目标,对所述问题生成模型进行训练,获得训练好的所述问题生成模型。

4.如权利要求1所述的方法,其特征在于,还包括:

针对目标文本,从所述目标语言中的疑问词中选择出第二疑问词;

将所述第二疑问词对应的第一提示文本和所述目标文本,输入至所述答案选择模型,获得所述答案选择模型输出的目标答案;

将所述第二疑问词对应的第二提示文本、所述目标文本和所述目标答案,输入至所述问题生成模型,获得所述问题生成模型输出的目标问题。

5.如权利要求1所述的方法,其特征在于,

所述疑问词对应的第一提示文本和第二提示文本,均包含有所述疑问词和所述疑问词的释义。

6.如权利要求1所述的方法,其特征在于,所述答案选择模型为自然语言理解模型,所述问题生成模型为自然语言生成模型。

7.一种模型训练装置,其特征在于,包括:

第一构建模块，用于根据目标语言中的疑问词及其释义，构建所述疑问词对应的第一提示文本和第二提示文本，其中，所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案，所述第二提示文本用于提示生成包含所述疑问词的问题；

第一获取模块，用于获取多个原始训练样本，每个原始训练样本包括文本样本、问题样本、所述问题样本对应的答案样本，所述问题样本包含有疑问词；

第二构建模块，用于针对每个原始训练样本，利用所述原始训练样本的文本样本、答案样本、问题样本中的疑问词对应的第一提示文本，构建一个第一训练样本，获得包括多个所述第一训练样本的第一训练集；以及，利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本，构建一个第二训练样本，获得包括多个所述第二训练样本的第二训练集；

训练模块，用于利用所述第一训练集，训练得到一个答案选择模型，以及，利用所述第二训练集，训练得到一个问题生成模型。

8.如权利要求7所述的装置，其特征在于，所述第一训练集还包括至少一个第三训练样本，所述装置还包括：

第三构建模块，用于按照以下方式构建所述第三训练样本：

针对第一文本样本，确定所述第一文本样本所对应的所有问题样本中均不包含的第一疑问词；

利用所述第一疑问词对应的第一提示文本、空白答案和所述第一文本样本，构建一个所述第三训练样本。

9.如权利要求7或8所述的装置，其特征在于，所述训练模块，还用于：

将所述第一训练集中每个训练样本的第一提示文本和文本样本输入至答案选择模型，以所述答案选择模型输出对应的答案样本为目标，对所述答案选择模型进行训练，获得训练好的所述答案选择模型；

将所述第二训练集中每个训练样本的第二提示文本、文本样本和答案样本输入至问题生成模型，以所述问题生成模型输出对应的问题样本为目标，对所述问题生成模型进行训练，获得训练好的所述问题生成模型。

10.如权利要求7所述的装置，其特征在于，还包括：

模型应用模块，用于针对目标文本，从所述目标语言中的疑问词中选择出第二疑问词；将所述第二疑问词对应的第一提示文本和所述目标文本，输入至所述答案选择模型，获得所述答案选择模型输出的目标答案；将所述第二疑问词对应的第二提示文本、所述目标文本和所述目标答案，输入至所述问题生成模型，获得所述问题生成模型输出的目标问题。

11.如权利要求7所述的装置，其特征在于，

所述疑问词对应的第一提示文本和第二提示文本，均包含有所述疑问词和所述疑问词的释义。

12.如权利要求7所述的装置，其特征在于，所述答案选择模型为自然语言理解模型，所述问题生成模型为自然语言生成模型。

13.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时，实现如权利要求1至6中任一项所述的模型训练方法的步骤。

模型训练方法、装置及存储介质

技术领域

[0001] 本发明涉及机器学习与自然语言处理(NLP,Natural Language Processing)技术领域,具体涉及一种模型训练方法、装置及存储介质。

背景技术

[0002] 问题生成技术是自然语言处理领域的一项重要的技术。问题生成的目标是:对于一篇用户指定的文章,生成若干与文章相关的问题并在文章中指出这些问题的答案。问题生成技术被广泛应用于问答系统和搜索引擎中,用以自动生成问题和答案的组合,即问题答案对(本文中有时候也简称为问答对)。问答系统和搜索引擎需要大量问题答案对。在问答系统中,其中一种自动问答方法是:使用相似度算法,将用户问题与数据库内的预构建的问题答案对相匹配从而获取答案;另一方面,搜索引擎使用机器阅读理解模型,针对用户的问题从检索出的文章中找出准确答案,然而机器阅读理解模型需要大量问答对进行训练。构建这些问答对需要大量的时间和人力,而且有时需要标注人员具有一定的领域专业性,而问题生成技术能够通过自动生成问答对,大幅度降低问答系统和搜索引擎的构建成本。

[0003] 目前,主流的问题生成系统往往包括一个答案选择模型和一个问题生成模型,其中,答案选择模型能够从文本中选取出潜在答案,问题生成模型能根据文本和答案选择模型选择出的潜在答案生成相关的问题。现有技术的一种问题生成方案,通过训练一个答案选择模型(sequence-to-sequence模型)来生成若干关键短语(潜在答案)。该答案选择模型从一个大规模问答数据集(SQuAD)的人工选择的答案中学习到了给潜在的答案分配更高的概率。另外,该方案提供了一个问题生成模型,该问题生成模型根据选择的潜在答案进行问题生成。该方案存在以下确定:训练过程中,答案选择模型学习在同一时刻选择文章中所有类型的潜在答案,这会导致模型难以捕捉提问价值的特征。另外,在选择答案后,问题生成模型没有从答案选择模型得到任何额外信息(例如答案类型)来帮助问题生成,这有可能会导致问题生成模型生成与所选答案不相关的问题。

[0004] 因此,亟需一种能够提高所选答案和生成的问题之间相关性的问题生成方案。

发明内容

[0005] 本申请的至少一个实施例提供了一种模型训练方法、装置及存储介质,能够提高答案选择模型所选择的答案与问题生成模型生成的问题的相关性。

[0006] 提高语义匹配任务的准确率。

[0007] 为了解决上述技术问题,本申请是这样实现的:

[0008] 第一方面,本申请实施例提供了一种模型训练方法,包括:

[0009] 根据目标语言中的疑问词及其释义,构建所述疑问词对应的第一提示文本和第二提示文本,其中,所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案,所述第二提示文本用于提示生成包含所述疑问词的问题;

[0010] 获取多个原始训练样本,每个原始训练样本包括文本样本、问题样本、所述问题样

本对应的答案样本，所述问题样本包含有疑问词；

[0011] 针对每个原始训练样本，利用所述原始训练样本的文本样本、答案样本、问题样本中的疑问词对应的第一提示文本，构建一个第一训练样本，获得包括多个所述第一训练样本的第一训练集；以及，利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本，构建一个第二训练样本，获得包括多个所述第二训练样本的第二训练集；

[0012] 利用所述第一训练集，训练得到一个答案选择模型，以及，利用所述第二训练集，训练得到一个问题生成模型。

[0013] 可选的，所述第一训练集还包括至少一个第三训练样本，所述第三训练样本按照以下方式构建：

[0014] 针对第一文本样本，确定所述第一文本样本所对应的所有问题样本中均不包含的第一疑问词；

[0015] 利用所述第一疑问词对应的第一提示文本、空白答案和所述第一文本样本，构建一个所述第三训练样本。

[0016] 可选的，所述利用所述第一训练集，训练得到一个答案选择模型，以及，利用所述第二训练集，训练得到一个问题生成模型，包括：

[0017] 将所述第一训练集中每个训练样本的第一提示文本和文本样本输入至答案选择模型，以所述答案选择模型输出对应的答案样本为目标，对所述答案选择模型进行训练，获得训练好的所述答案选择模型；

[0018] 将所述第二训练集中每个训练样本的第二提示文本、文本样本和答案样本输入至问题生成模型，以所述问题生成模型输出对应的问题样本为目标，对所述问题生成模型进行训练，获得训练好的所述问题生成模型。

[0019] 可选的，上述方法还包括：

[0020] 针对目标文本，从所述目标语言中的疑问词中选择出第二疑问词；

[0021] 将所述第二疑问词对应的第一提示文本和所述目标文本，输入至所述答案选择模型，获得所述答案选择模型输出的目标答案；

[0022] 将所述第二疑问词对应的第二提示文本、所述目标文本和所述目标答案，输入至所述问题生成模型，获得所述问题生成模型输出的目标问题。

[0023] 可选的，所述疑问词对应的第一提示文本和第二提示文本，均包含有所述疑问词和所述疑问词的释义。

[0024] 可选的，所述答案选择模型为自然语言理解模型，所述问题生成模型为自然语言生成模型。

[0025] 第二方面，本申请实施例提供了一种模型训练装置，包括：

[0026] 第一构建模块，用于根据目标语言中的疑问词及其释义，构建所述疑问词对应的第一提示文本和第二提示文本，其中，所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案，所述第二提示文本用于提示生成包含所述疑问词的问题；

[0027] 第一获取模块，用于获取多个原始训练样本，每个原始训练样本包括文本样本、问题样本、所述问题样本对应的答案样本，所述问题样本包含有疑问词；

[0028] 第二构建模块，用于针对每个原始训练样本，利用所述原始训练样本的文本样本、

答案样本、问题样本中的疑问词对应的第一提示文本，构建一个第一训练样本，获得包括多个所述第一训练样本的第一训练集；以及，利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本，构建一个第二训练样本，获得包括多个所述第二训练样本的第二训练集；

[0029] 训练模块，用于利用所述第一训练集，训练得到一个答案选择模型，以及，利用所述第二训练集，训练得到一个问题生成模型。

[0030] 可选的，所述第一训练集还包括至少一个第三训练样本，所述装置还包括：

[0031] 第三构建模块，用于按照以下方式构建所述第三训练样本：

[0032] 针对第一文本样本，确定所述第一文本样本所对应的所有问题样本中均不包含的第一疑问词；

[0033] 利用所述第一疑问词对应的第一提示文本、空白答案和所述第一文本样本，构建一个所述第三训练样本。

[0034] 可选的，所述训练模块，还用于：

[0035] 将所述第一训练集中每个训练样本的第一提示文本和文本样本输入至答案选择模型，以所述答案选择模型输出对应的答案样本为目标，对所述答案选择模型进行训练，获得训练好的所述答案选择模型；

[0036] 将所述第二训练集中每个训练样本的第二提示文本、文本样本和答案样本输入至问题生成模型，以所述问题生成模型输出对应的问题样本为目标，对所述问题生成模型进行训练，获得训练好的所述问题生成模型。

[0037] 可选的，上述装置还包括：

[0038] 模型应用模块，用于针对目标文本，从所述目标语言中的疑问词中选择出第二疑问词；将所述第二疑问词对应的第一提示文本和所述目标文本，输入至所述答案选择模型，获得所述答案选择模型输出的目标答案；将所述第二疑问词对应的第二提示文本、所述目标文本和所述目标答案，输入至所述问题生成模型，获得所述问题生成模型输出的目标问题。

[0039] 可选的，所述疑问词对应的第一提示文本和第二提示文本，均包含有所述疑问词和所述疑问词的释义。

[0040] 可选的，所述答案选择模型为自然语言理解模型，所述问题生成模型为自然语言生成模型。

[0041] 第三方面，本申请实施例提供了一种计算机可读存储介质，所述计算机可读存储介质上存储有程序，所述程序被处理器执行时，实现如上所述的方法的步骤。

[0042] 与现有技术相比，本申请实施例提供的模型训练方法及装置，在模型训练过程中，将疑问词对应的第一提示文本和文本样本作为答案选择模型的输入，使得答案选择模型输出特定类型的潜在答案，另外，还将所述疑问词对应的第二提示文本、文本样本和答案样本作为问题生成模型的输入，从而通过包含有同一疑问词及其释义的提示文本，将答案选择模型和问题生成模型关联起来，提升了所选答案和生成的问题的相关性，进而能够改善模型所生成的问题答案对的性能。

附图说明

[0043] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本申请的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0044] 图1为本申请实施例的模型训练方法的一种流程图；

[0045] 图2为本申请实施例的构建提示文本的示例图；

[0046] 图3为采用提示模板的一个示例图；

[0047] 图4为本申请实施例针对问题样本构建提示文本的示例图；

[0048] 图5为本申请实施例利用第一训练样本训练答案选择模型的一个示例图；

[0049] 图6为本申请实施例利用第二训练样本训练问题生成模型的一个示例图；

[0050] 图7为本申请实施例的模型训练装置的结构示意图；

[0051] 图8为本申请实施例的模型训练装置的另一结构示意图。

具体实施方式

[0052] 为使本申请要解决的技术问题、技术方案和优点更加清楚，下面将结合附图及具体实施例进行详细描述。在下面的描述中，提供诸如具体的配置和组件的特定细节仅仅是为了帮助全面理解本申请的实施例。因此，本领域技术人员应该清楚，可以对这里描述的实施例进行各种改变和修改而不脱离本申请的范围和精神。另外，为了清楚和简洁，省略了对已知功能和构造的描述。

[0053] 应理解，说明书通篇中提到的“一个实施例”或“一实施例”意味着与实施例有关的特定特征、结构或特性包括在本申请的至少一个实施例中。因此，在整个说明书各处出现的“在一个实施例中”或“在一实施例中”未必一定指相同的实施例。此外，这些特定的特征、结构或特性可以任意适合的方式结合在一个或多个实施例中。本申请的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，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、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。说明书以及权利要求中“和/或”表示所连接对象的至少其中之一。

[0054] 在本申请的各种实施例中，应理解，下述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

[0055] 问题生成系统通常包括一个答案选择模型和一个问题生成模型。相关研究尝试在以下两个方面提升问题生成的性能。其一是在答案选择模型选择的答案是否更具有提问价值方面，另一方面是问题生成模型生成的问题是否与选择的答案具有相关性方面。例如，一种相关技术方案使用外部解析器来解析出文本中的实体(如时间、人物等)及它们之间的关系，并且选择与其他实体关系数最多的实体作为潜在答案。接下来，这些潜在答案和它们的实体类型被一同输入到问题生成模型来生成问题。上述方案中，实体类型可以用来帮助问

题生成模型来生成与作为答案的实体更相关的问题，但是这种方案高度依赖外部解析器的性能并且无法选择命名实体之外范围内的词作为潜在答案，例如，以英语为例，无法选择“take a nap”这种并非命名实体的动词短语作为潜在答案。另外还有一些相关方案，通过训练一个基于神经网络的答案选择模型来选择潜在答案。这些方案虽然可以选择短语作为答案，但是答案选择模型并不会将任何额外信息传递给后续的问题生成模型。另外，从各种各样类型的潜在答案中捕捉提问价值的特征，这对于问题生成模型来说是困难的。

[0056] 为提高所选答案和生成的问题的相关性，改善所生成的问题答案对的性能，本申请实施例提供了一种模型训练方法，如图1所示，该方法包括：

[0057] 步骤11，根据目标语言中的疑问词及其释义，构建所述疑问词对应的第一提示文本和第二提示文本，其中，所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案，所述第二提示文本用于提示生成包含所述疑问词的问题。

[0058] 这里，目标语言可以是中文、英文或日文等自然语言，本文中将主要以英文为例进行说明。在自然语言中，通常存在着疑问词，以英文为例，疑问词包括“who”、“when”、“where”、“how much”、“how long”等等；以中文为例，疑问词包括：“谁”、“什么时候”、“什么地方”、“为什么”等等。各个疑问词的释义可以通过查询相关词典(如牛津词典)或百科(如维基百科)获取，本申请实施例对此不做具体限定。

[0059] 具体的，可以获取预先通过人工方式收集的目标语言中所有的疑问词，将所有的疑问词与包括多个原始训练样本的原始训练集中的问题样本进行字符串匹配，保留在原始训练集的问题样本中出现过的疑问词。然后，对保留的每一个疑问词，通过查词典或者百科，获取它们的释义。例如“when”的释义为“the time at something happens”。

[0060] 针对每个疑问词，可以根据所述疑问词及其释义，构建所述疑问词对应的第一提示文本(答案选择提示文本)和第二提示文本(问题生成提示文本)，其中，所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案。具体的，所述疑问词对应的第一提示文本，包含所述疑问词和所述疑问词的释义。所述第二提示文本用于提示生成包含所述疑问词的问题。所述疑问词对应的第二提示文本，也包含所述疑问词和所述疑问词的释义。这样，本申请实施例可以针对每个疑问词，生成同一疑问词所对应的第一提示文本和第二提示文本。

[0061] 下面以英文为例，提供构建提示文本的具体示例。

[0062] 本示例中预先构建第一提示模板和第二提示模板。本文中，第一提示模板有时也称作答案选择提示(prompt)模板，第二提示模板有时也称作问题生成提示(prompt)模板。

[0063] 其中，第一提示模板为一段提示文本，该提示文本有两处空白位置，分别对应于疑问词和疑问词的释义。该提示文本用于提示针对包含所述疑问词的问题选择潜在答案。在生成某个疑问词对应的第一提示文本时，将该疑问词及其释义分别填入第一提示模板中对应的空白位置处，从而获得第一提示文本。

[0064] 第二提示模板也是一段提示文本，该提示文本也有两处空白位置，分别对应于疑问词和疑问词的释义。该提示文本用于提示生成包含所述疑问词的问题。在生成某个疑问词对应的第二提示文本时，将该疑问词及其释义分别填入第二提示模板中对应的空白位置处，从而获得第二提示文本。

[0065] 例如，第一提示模板(答案选择prompt模板)具体可以为：

[0066] Find a text to answer a question which asks about.

[0067] 其中,第一个下划线对应的空白位置用于填入疑问词,第二个下划线对应的空白位置用于填入疑问词的释义。

[0068] 第二提示模板(问题生成prompt模板)具体可以为:

[0069] Ask a question which asks about.

[0070] 以疑问词“When”为例,“When”的释义为“the time at something happens”,将“When”及其释义填入上面的模板,可以得到如图2所示的第一提示文本和第二提示文本,这两个提示文本均为疑问词“When”对应的提示文本,因此这两个提示文本之间存在一一对应关系。

[0071] 另外需要说明的是,以上仅是本申请实施例可以采用的提示模板/提示文本的一种示例,本申请实施例还可以采用其他形式的提示模板/提示文本,例如,第一提示模板还可以是以下任一形式:

[0072] Select an answer to question which is about.

[0073] Choose an answer to question which is about.

[0074] 第二提示模板还可以是以下任一形式:

[0075] Generate a question which is about.

[0076] Provide a question which is about.

[0077] 步骤12,获取多个原始训练样本,每个原始训练样本包括文本样本、问题样本、所述问题样本对应的答案样本,所述问题样本包含有疑问词。

[0078] 这里,获取原始数据集,所述原始数据集包括多个原始训练样本。通常,训练样本包括:一段文本样本,一个或多个问题,以及每个问题对应的答案。为了便于处理,对于包含有多个问题及答案的训练样本,可以将其拆分为多个原始训练样本,使得每个原始训练样本仅包括一个问题及其答案。这样,每个原始训练样本包括有一个文本样本、一个问题样本和该问题样本对应的答案样本。文本样本可以是一段文章,问题样本是针对该文章提供的一个问题,答案样本则是该问题的答案,通常,答案样本是该文章中的部分文字,即文章文本的子字符串。

[0079] 本申请实施例中,原始训练样本是指任何机器阅读理解的数据集,比如由斯坦福大学公开的SQuAD1.1数据集。数据集中每一个训练数据都由人工标注,每一个训练数据包括:一篇文章,数个与文章相关的问题以及其对应的答案(答案均为文章的子字符串)。一篇文章、一个与该文章相关的问题以及该问题对应的答案,可以作为答案选择模型和问题生成模型的一个训练样本。这些文章、问题及答案在答案选择模型和问题生成模型的训练过程中都可以被获取,但是却以不同的方式输入至模型:对于答案选择模型的训练,答案被用作训练的目标,而文章和问题用作输入;对于问题生成模型的训练,问题被用作训练的目标,而文章和答案用作输入。

[0080] 步骤13,针对每个原始训练样本,利用所述原始训练样本的文本样本、答案样本、问题样本中的疑问词对应的第一提示文本,构建一个第一训练样本,获得包括多个所述第一训练样本的第一训练集;以及,利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本,构建一个第二训练样本,获得包括多个所述第二训练样本的第二训练集。

[0081] 很多预训练语言模型,例如基于Transformers的双向编码器(Bidirectional Encoder Representation from Transformer,BERT)模型、T5模型等,在大多数自然语言处理任务上性能优秀,并且可以接收一个句子作为提示(prompt),来引导模型完成指定的任务。如图3所示,将prompt语句“translate English to German”(即,将英语翻译成德语)与要翻译的目标英文语句相连接后,输入到T5模型后,T5模型能够输出目标英文语句的德语翻译。基于此,本申请实施例提出一种prompt驱动的问题生成相关模型的训练方法,答案选择模型和问题生成模型都使用了prompt机制,具体的,通过引入提示文本(prompt文本),在原始训练样本的基础上重新构建新的训练样本,以训练相关模型。

[0082] 具体的,本申请实施例在提示文本中融入疑问词及其释义的信息,疑问词及其释义可以反映出潜在答案的类型信息,例如,对于疑问词“when”来说,其潜在答案是时间类型,对于疑问词“where”来说,其潜在答案是地点类型,本申请实施例通过在提示文本中融入了潜在答案的类型信息,一方面,提示文本将答案类型信息作为提示输入到答案选择模型,使其每次只输出特定类型的潜在答案,这使得模型能够更好的捕捉“提问价值”特征;另一方面,提示文本也将同样的答案类型信息传递给问题生成模型,从而将两个模型联系起来,使得问题生成模型能够生成与所选择的潜在答案更相关的问题。

[0083] 本申请实施例在构建训练样本时,针对每个原始训练样本,可以构建得到该原始训练样本对应的一个第一训练样本,具体的,可以利用所述原始训练样本的文本样本、答案样本、问题样本中的疑问词对应的第一提示文本,构建一个第一训练样本,这样,针对多个原始训练样本,可以获得多个第一训练样本,从而得到第一训练集。类似的,针对每个原始训练样本,可以构建得到该原始训练样本对应的一个第二训练样本,具体的,可以利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本,构建一个第二训练样本,这样,针对多个原始训练样本,可以获得多个第二训练样本,从而得到第二训练集。

[0084] 假设某个原始训练样本中的问题样本为“When were the Normans in Normandy?”,其中的疑问词为“When”,那么,针对该疑问词“When”,可以构建如图4所示的两个提示文本,进而得到一个第一训练样本和一个第二训练样本,其中,该第一训练样本包括图4中的第一提示文本、该原始训练样本的文本样本和答案样本,该第二训练样本包括图4中的第二提示文本、该原始训练样本的文本样本、问题样本和答案样本。

[0085] 步骤14,利用所述第一训练集,训练得到一个答案选择模型,以及,利用所述第二训练集,训练得到一个问题生成模型。

[0086] 本申请实施例中,所述答案选择模型为自然语言理解模型,具体可以是以下模型中的任一种:BERT、RoBERTa、ALBERT、ERNIE、ELECTRA等模型。答案选取模型可以使用BERT之类的预训练语言模型作为基础架构。例如,BERT之类的预训练语言模型是在Transformer模型架构的基础上由大规模语料预训练而得,能够实现诸如答案选取等的自然语言处理任务。所述问题生成模型为自然语言生成模型,具体可以是以下模型中的任一种:T5模型、GPT、BART等模型。例如,问题生成模型可以使用T5模型作为基础架构,T5模型是基于Transformer的编码器-解码器架构预训练而得,具有文本生成能力。

[0087] 这里,在答案选择模型训练时,将所述第一训练集中每个第一训练样本的第一提示文本和文本样本输入至答案选择模型,所述第一提示文本中包含了问题样本中的疑问

词,并以所述答案选择模型输出对应的答案样本为目标,对所述答案选择模型进行训练,获得训练好的所述答案选择模型。

[0088] 图5提供了利用第一训练样本训练答案选择模型的一个示例,此时答案选择模型的输入为第一训练样本中的第一提示文本和文本样本,训练目标是模型输出的第一训练样本中的答案样本。图5中的答案生成模型是BERT之类的预训练语言模型,将第一提示文本和文本样本配对后,以图5的形式构造输入,并输入至答案选择模型,即用特殊分隔符“<SEP>”,将第一提示文本与文本样本连接后作为答案选择模型的输入,这种输入方式是由BERT模型预训练时的输入方式决定的。以图5为例,

[0089] 第一提示文本为:Find a text to answer a“when”question which asks about the time at something happens.

[0090] 文本样本为:The Normans(Norman:Nourmands;French:Normands;Latin:Normanni)were the people who in the 10th and 11th centuries gave their name to Normandy,a region in France…The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century,and it continued to evolve over the succeeding centuries.

[0091] 答案样本为:in the first half of the 10th century或in the 10th and 11th centuries.

[0092] 也就是说,在答案选择模型训练过程中,将第一训练样本中的第一提示文本和文本样本输入至答案选择模型,答案选择模型从文本样本选择出一段文本作为答案文本并输出,计算答案选择模型输出的答案文本与该第一训练样本中的答案样本的相似性,并基于所述相似性对答案选择模型的模型参数进行优化,以使答案选择模型输出的答案文本与所述答案样本更加接近。通过上述优化过程,最终获得训练好的答案选择模型。

[0093] 在问题生成模型训练时,将所述第二训练集中每个第二训练样本的第二提示文本、文本样本和答案样本输入至问题生成模型,以所述问题生成模型输出对应的问题样本为目标,对所述问题生成模型进行训练,获得训练好的所述问题生成模型。

[0094] 图6提供了利用第二训练样本训练问题生成模型的一个示例,图6中使用答案样本、第二提示文本和文本样本构成三元组,以图6所示的输入方式构造输入,并输入至问题生成模型。这种输入方式是由作为问题生成模型基础架构的T5模型决定的。其中,使用“Prompt:”加到第二提示文本前作为前缀,使用“Paragraph:”加到文本样本前作为前缀,之后再将二者相连。而答案样本则被隐式输入,即使用XML形式(<answer>与</answer>)将答案标记在文本样本中(如图中6的下划线文字部分)。以图6为例,

[0095] 第二提示文本为:Ask a“when”question which asks about the time at something happens.

[0096] 文本样本为:The Normans(Norman:Nourmands;French:Normands;Latin:Normanni)were the people who in the 10th and 11th centuries gave their name to Normandy,a region in France…The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century,and it continued to evolve over the succeeding centuries.

[0097] 答案样本为:in the 10th and 11th centuries.

[0098] 问题样本为:When were the Normans in Normandy?

[0099] 也就是说,在问题生成模型训练过程中,将第二训练样本中的第二提示文本、文本样本和问题样本输入至问题生成模型,问题生成模型生成一段文本作为问题文本并输出,计算问题生成模型输出的问题文本与该第二训练样本中的问题样本的相似性,并基于所述相似性对问题生成模型的模型参数进行优化,以使问题生成模型输出的问题文本与所述问题样本更加接近。通过上述优化过程,最终获得训练好的问题生成模型。

[0100] 通过以上步骤,本申请实施例在模型训练过程中,将疑问词对应的第一提示文本和文本样本作为答案选择模型的输入,使得答案选择模型输出特定类型的潜在答案,另外,还将所述疑问词对应的第二提示文本、文本样本和答案样本作为问题生成模型的输入,从而通过包含有同一疑问词及其释义的提示文本,将答案选择模型和问题生成模型关联起来,使得问题生成模型能够生成与所选择的潜在答案更相关的问题,进而能够改善上述模型所生成的问题答案对的性能。

[0101] 上述步骤13中所生成的第一训练样本为正训练样本,即文本样本中存在问题的答案。考虑到对于某些带有某些疑问词的问题,文本样本中可能不存在这些问题的答案,因此,本申请实施例进一步生成第一训练集时,在该第一训练集中增加至少一个第三训练样本(负训练样本)。具体的,针对原始训练样本中的某个文本样本(为了便于描述,这里称之为第一文本样本),确定所述第一文本样本所对应的所有问题样本中均不包含的第一疑问词。这里,所述第一疑问词可以是存在于所述目标语言的所有疑问词中,但不存在于所述第一文本样本所对应的所有问题样本中。所述第一疑问词还可以是存在于原始训练集的问题样本中,但不存在于所述第一文本样本所对应的所有问题样本中。然后,利用所述第一疑问词对应的第一提示文本、空白答案和所述第一文本样本,构建一个所述第三训练样本。例如,空白答案为“None”(空,作为答案)。

[0102] 例如,假设所有原始训练样本中共有以下3个原始训练样本均包括同一个文本样本x,具体如下:

[0103] 原始训练样本a:文本样本x,问题1(包含疑问词1),答案1;

[0104] 原始训练样本b:文本样本x,问题2(包含疑问词2),答案2;

[0105] 原始训练样本c:文本样本x,问题3(包含疑问词3),答案3;

[0106] 假设疑问词一共有5个,分别为疑问词1～5,则可以看出,同一文本样本x所对应的所有问题样本中均不包含的疑问词为疑问词4和疑问词5,因此,可以生成以下两个第三训练样本(负训练样本):

[0107] 第三训练样本1:文本样本x,疑问词4对应的第一提示文本,答案(None);

[0108] 第三训练样本2:文本样本x,疑问词5对应的第一提示文本,答案(None)。

[0109] 第一训练集中的正负样本的比例可以通过使用不同比例训练多个答案选择模型,根据答案选择模型的性能,选取性能指标最优的答案选择模型所使用的比例作为最终比例。

[0110] 通过上述步骤14,本申请实施例能够获得训练好的答案选择模型和问题生成模型,在此之后,本申请实施例还可以利用上述模型为目标文本生成问题答案对。所述目标文本可以是用户输入的一段文本。具体的,本申请实施例可以述目标语言中的疑问词中选择出一个疑问词(为了便于描述,称之为第二疑问词)。然后,将所述第二疑问词对应的所述第

一提示文本和所述目标文本，输入至所述答案选择模型，获得所述答案选择模型输出的答案(为便于描述，这里称之为目标答案)。然后，将所述第二疑问词对应的所述第二提示文本、所述目标文本和所述目标答案，输入至所述问题生成模型，获得所述问题生成模型输出的目标问题，从而获得由所述目标问题和目标答案组成的一组问题答案对。上述过程中，第一提示文本和第二提示文本的生成方式与模型训练过程中相同。上述过程也可以参考图5-图6，此时图5和图6中的文本样本替换为上述目标文本。通过以上方式，利用一个或多个疑问词，可以获得所述目标文本的一组或多组问题答案对。

[0111] 基于以上方法，本申请实施例还提供了实施上述方法的装置，请参考图7，本申请实施例提供了一种模型训练装置，包括：

[0112] 第一构建模块71，用于根据目标语言中的疑问词及其释义，构建所述疑问词对应的第一提示文本和第二提示文本，其中，所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案，所述第二提示文本用于提示生成包含所述疑问词的问题；

[0113] 第一获取模块72，用于获取多个原始训练样本，每个原始训练样本包括文本样本、问题样本、所述问题样本对应的答案样本，所述问题样本包含有疑问词；

[0114] 第二构建模块73，用于针对每个原始训练样本，利用所述原始训练样本的文本样本、答案样本、问题样本中的疑问词对应的第一提示文本，构建一个第一训练样本，获得包括多个所述第一训练样本的第一训练集；以及，利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本，构建一个第二训练样本，获得包括多个所述第二训练样本的第二训练集；

[0115] 训练模块74，用于利用所述第一训练集，训练得到一个答案选择模型，以及，利用所述第二训练集，训练得到一个问题生成模型。

[0116] 通过以上模块，本申请实施例能够提升训练得到的模型所选答案和生成的问题的相关性。

[0117] 可选的，所述第一训练集还包括至少一个第三训练样本，上述装置还包括：

[0118] 第三构建模块，用于按照以下方式构建所述第三训练样本：

[0119] 针对第一文本样本，确定所述第一文本样本所对应的所有问题样本中均不包含的第一疑问词；

[0120] 利用所述第一疑问词对应的第一提示文本、空白答案和所述第一文本样本，构建一个所述第三训练样本。

[0121] 可选的，所述训练模块74，还用于：

[0122] 将所述第一训练集中每个训练样本的第一提示文本和文本样本输入至答案选择模型，以所述答案选择模型输出对应的答案样本为目标，对所述答案选择模型进行训练，获得训练好的所述答案选择模型；

[0123] 将所述第二训练集中每个训练样本的第二提示文本、文本样本和答案样本输入至问题生成模型，以所述问题生成模型输出对应的问题样本为目标，对所述问题生成模型进行训练，获得训练好的所述问题生成模型。

[0124] 可选的，上述装置还包括：

[0125] 模型应用模块，用于针对目标文本，从所述目标语言中的疑问词中选择出第二疑问词；将所述第二疑问词对应的第一提示文本和所述目标文本，输入至所述答案选择模型，

获得所述答案选择模型输出的目标答案;将所述第二疑问词对应的第二提示文本、所述目标文本和所述目标答案,输入至所述问题生成模型,获得所述问题生成模型输出的目标问题。

[0126] 可选的,所述疑问词对应的第一提示文本和第二提示文本,均包含有所述疑问词和所述疑问词的释义。

[0127] 可选的,所述答案选择模型为自然语言理解模型,所述问题生成模型为自然语言生成模型。

[0128] 请参考图8,本申请实施例还提供了模型训练装置的一种硬件结构框图,如图8所示,该模型训练装置800包括:

[0129] 处理器802;和

[0130] 存储器804,在所述存储器804中存储有计算机程序指令,

[0131] 其中,在所述计算机程序指令被所述处理器运行时,使得所述处理器802执行以下步骤:

[0132] 根据目标语言中的疑问词及其释义,构建所述疑问词对应的第一提示文本和第二提示文本,其中,所述第一提示文本用于提示针对包含所述疑问词的问题选择潜在答案,所述第二提示文本用于提示生成包含所述疑问词的问题;

[0133] 获取多个原始训练样本,每个原始训练样本包括文本样本、问题样本、所述问题样本对应的答案样本,所述问题样本包含有疑问词;

[0134] 针对每个原始训练样本,利用所述原始训练样本的文本样本、答案样本、问题样本中的疑问词对应的第一提示文本,构建一个第一训练样本,获得包括多个所述第一训练样本的第一训练集;以及,利用所述原始训练样本的文本样本、问题样本、答案样本、问题样本中的疑问词对应的第二提示文本,构建一个第二训练样本,获得包括多个所述第二训练样本的第二训练集;

[0135] 利用所述第一训练集,训练得到一个答案选择模型,以及,利用所述第二训练集,训练得到一个问题生成模型。

[0136] 进一步地,如图8所示,该模型训练装置800还包括网络接口801、输入设备803、硬盘805、和显示设备806。

[0137] 上述各个接口和设备之间可以通过总线架构互连。总线架构可以是可以包括任意数量的互联的总线和桥。具体由处理器802代表的一个或者多个中央处理器(CPU)和/或图形处理器(GPU),以及由存储器804代表的一个或者多个存储器的各种电路连接在一起。总线架构还可以将诸如外围设备、稳压器和功率管理电路等之类的各种其它电路连接在一起。可以理解,总线架构用于实现这些组件之间的连接通信。总线架构除包括数据总线之外,还包括电源总线、控制总线和状态信号总线,这些都是本领域所公知的,因此本文不再对其进行详细描述。

[0138] 所述网络接口801,可以连接至网络(如因特网、局域网等),从网络中接收原始训练样本等数据,并可以将接收到的数据保存在硬盘805中。

[0139] 所述输入设备803,可以接收操作人员输入的各种指令,并发送给处理器802以供执行。所述输入设备803可以包括键盘或者点击设备(例如,鼠标,轨迹球(trackball)、触感板或者触摸屏等。

[0140] 所述显示设备806,可以将处理器802执行指令获得的结果进行显示,例如显示模型训练进度等。

[0141] 所述存储器804,用于存储操作系统运行所必须的程序和数据,以及处理器802计算过程中的中间结果等数据。

[0142] 可以理解,本申请实施例中的存储器804可以是易失性存储器或非易失性存储器,或可包括易失性和非易失性存储器两者。其中,非易失性存储器可以是只读存储器(ROM)、可编程只读存储器(PROM)、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)或闪存。易失性存储器可以是随机存取存储器(RAM),其用作外部高速缓存。本文描述的装置和方法的存储器804旨在包括但不限于这些和任意其它适合类型的存储器。

[0143] 在一些实施方式中,存储器804存储了如下的元素,可执行模块或者数据结构,或者他们的子集,或者他们的扩展集:操作系统8041和应用程序8042。

[0144] 其中,操作系统8041,包含各种系统程序,例如框架层、核心库层、驱动层等,用于实现各种基础业务以及处理基于硬件的任务。应用程序8042,包含各种应用程序,例如浏览器(Browser)等,用于实现各种应用业务。实现本申请实施例方法的程序可以包含在应用程序8042中。

[0145] 本申请上述实施例揭示的方法可以应用于处理器802中,或者由处理器802实现。处理器802可能是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器802中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器802可以是通用处理器、数字信号处理器(DSP)、专用集成电路(ASIC)、现场可编程门阵列(FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件,可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器,闪存、只读存储器,可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器804,处理器802读取存储器804中的信息,结合其硬件完成上述方法的步骤。

[0146] 可以理解的是,本文描述的这些实施例可以用硬件、软件、固件、中间件、微码或其组合来实现。对于硬件实现,处理单元可以实现在一个或多个专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理设备(DSPD)、可编程逻辑设备(PLD)、现场可编程门阵列(FPGA)、通用处理器、控制器、微控制器、微处理器、用于执行本申请所述功能的其它电子单元或其组合中。

[0147] 对于软件实现,可通过执行本文所述功能的模块(例如过程、函数等)来实现本文所述的技术。软件代码可存储在存储器中并通过处理器执行。存储器可以在处理器中或在处理器外部实现。

[0148] 具体地,所述第一训练集还包括至少一个第三训练样本,所述计算机程序被处理器802执行时还可实现如下步骤:

[0149] 按照以下方式构建所述第三训练样本:

[0150] 针对第一文本样本,确定所述第一文本样本所对应的所有问题样本中均不包含的第一疑问词;

[0151] 利用所述第一疑问词对应的第一提示文本、空白答案和所述第一文本样本，构建一个所述第三训练样本。

[0152] 具体地，所述计算机程序被处理器802执行时还可实现如下步骤：

[0153] 将所述第一训练集中每个训练样本的第一提示文本和文本样本输入至答案选择模型，以所述答案选择模型输出对应的答案样本为目标，对所述答案选择模型进行训练，获得训练好的所述答案选择模型；

[0154] 将所述第二训练集中每个训练样本的第二提示文本、文本样本和答案样本输入至问题生成模型，以所述问题生成模型输出对应的问题样本为目标，对所述问题生成模型进行训练，获得训练好的所述问题生成模型。

[0155] 具体地，所述计算机程序被处理器802执行时还可实现如下步骤：

[0156] 针对目标文本，从所述目标语言中的疑问词中选择出第二疑问词；

[0157] 将所述第二疑问词对应的第一提示文本和所述目标文本，输入至所述答案选择模型，获得所述答案选择模型输出的目标答案；

[0158] 将所述第二疑问词对应的第二提示文本、所述目标文本和所述目标答案，输入至所述问题生成模型，获得所述问题生成模型输出的目标问题。

[0159] 可选的，所述疑问词对应的第一提示文本和第二提示文本，均包含有所述疑问词和所述疑问词的释义。

[0160] 可选的，所述答案选择模型为自然语言理解模型，所述问题生成模型为自然语言生成模型。

[0161] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0162] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0163] 在本申请所提供的实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置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，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的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，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0164] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者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者全部单元来实现本申请实施例方案的目的。

[0165] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[0166] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以

存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0167] 以上所述，仅为本申请的具体实施方式，但本申请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申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

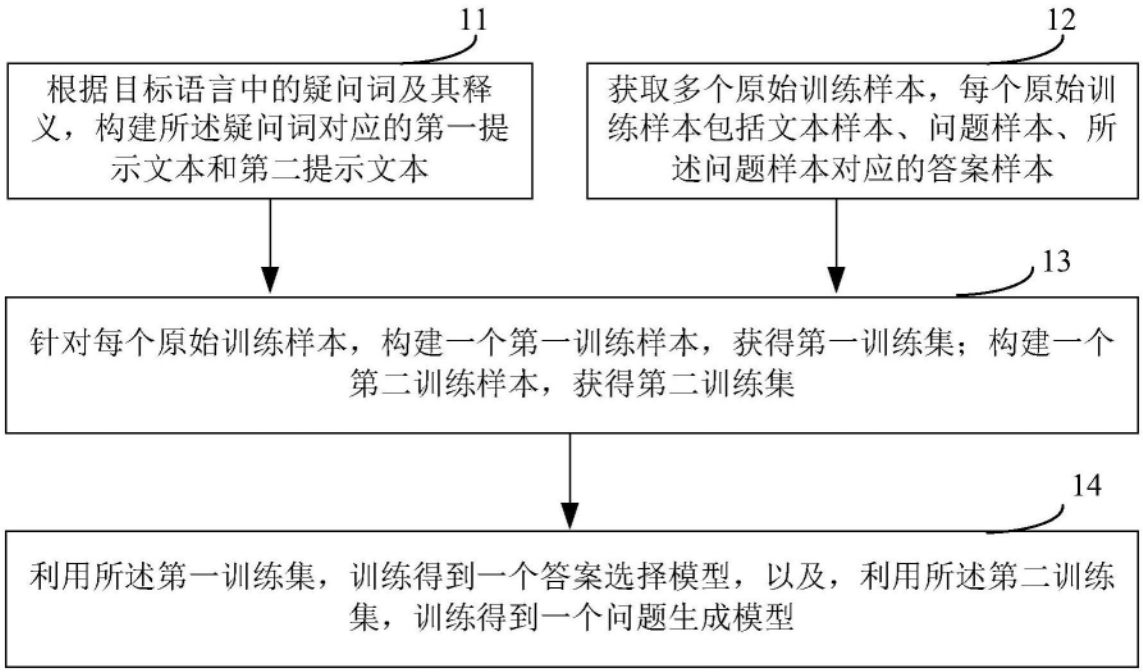

11
根据目标语言中的疑问词及其释义，构建所述疑问词对应的第一提示文本和第二提示文本
12
获取多个原始训练样本，每个原始训练样本包括文本样本、问题样本、所述问题样本对应的答案样本
13
针对每个原始训练样本，构建一个第一训练样本，获得第一训练集；构建一个第二训练样本，获得第二训练集
14
利用所述第一训练集，训练得到一个答案选择模型，以及，利用所述第二训练集，训练得到一个问题生成模型

图1

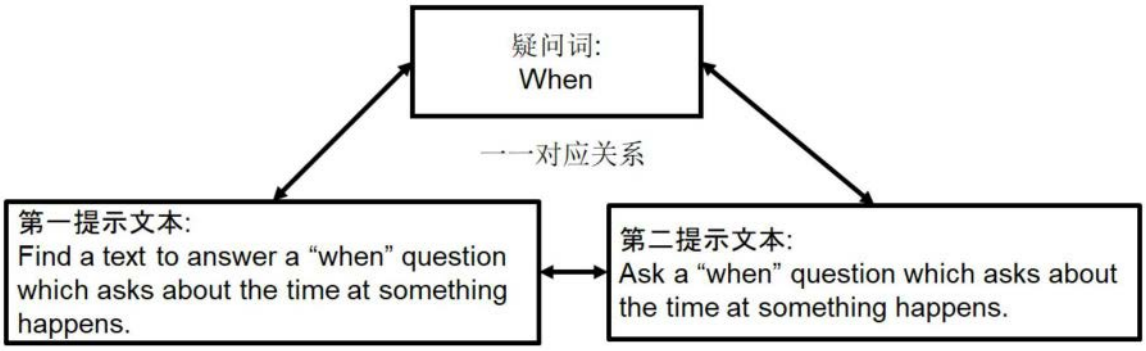

疑问词:
When
一一对应关系
第一提示文本:
Find a text to answer a “when” question which asks about the time at something happens.
第二提示文本:
Ask a “when” question which asks about the time at something happens.

图2

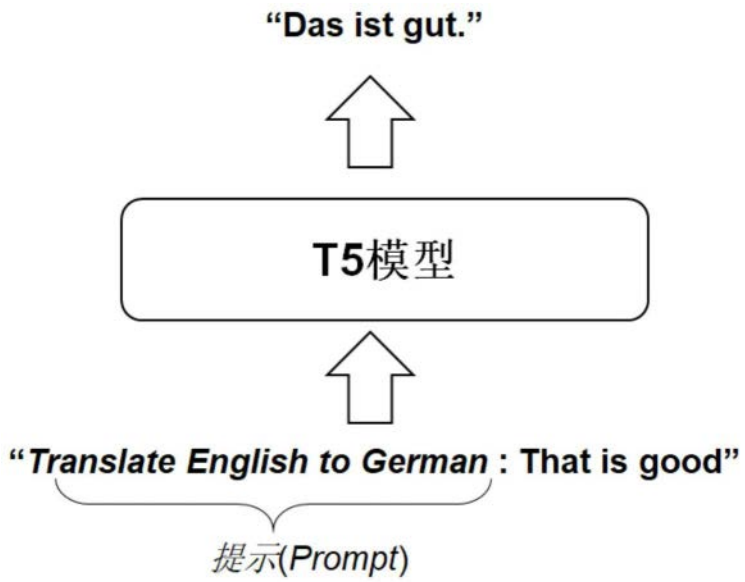

“Das ist gut.”
T5模型
“Translate English to German : That is good”
提示(Prompt)

图3

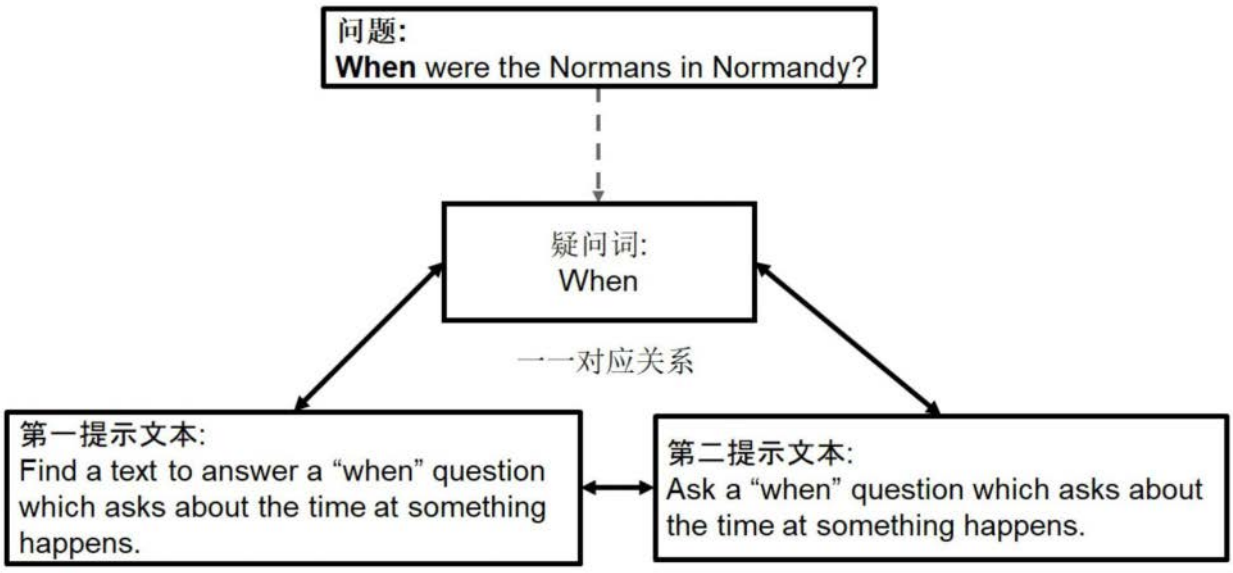

问题:
When were the Normans in Normandy?
疑问词:
When
一一对应关系
第一提示文本:
Find a text to answer a “when” question which asks about the time at something happens.
第二提示文本:
Ask a “when” question which asks about the time at something happens.

图4

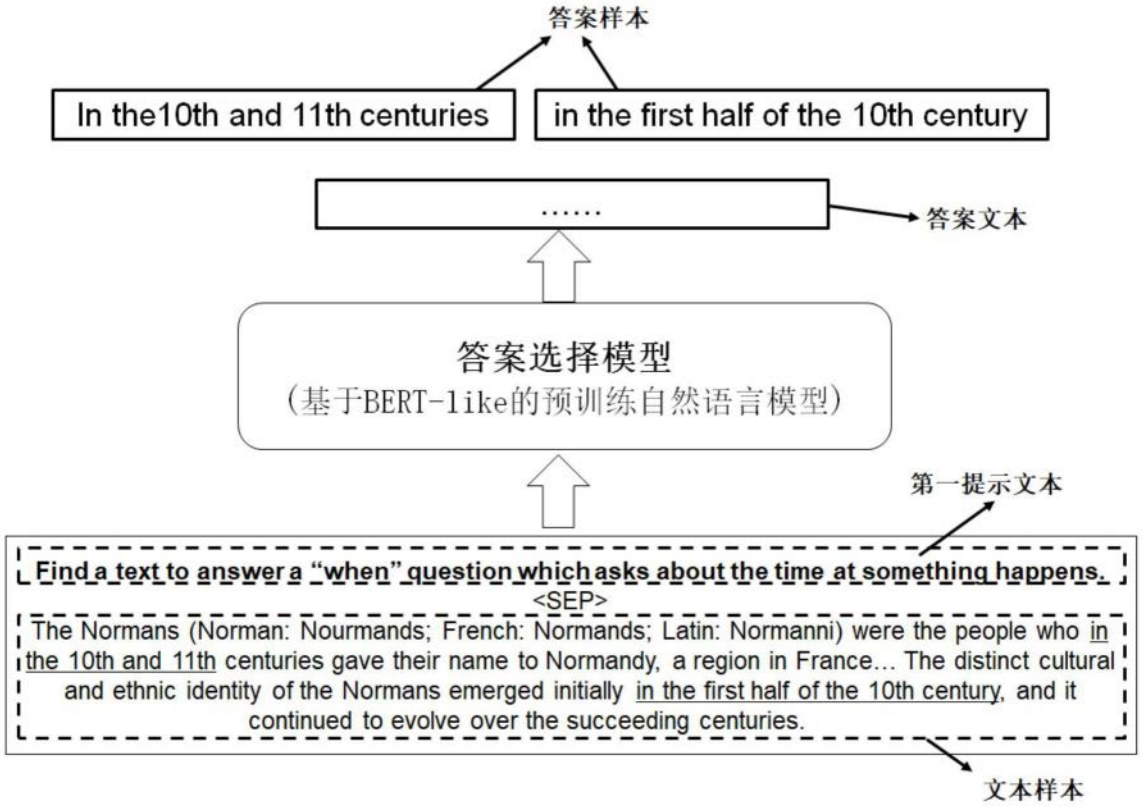

答案样本
In the10th and 11th centuries
in the first half of the 10th century
.....
答案文本
答案选择模型
(基于BERT-like的预训练自然语言模型)
第一提示文本
Find a text to answer a "when" question which asks about the time at something happens.
<SEP>
The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France... The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.
文本样本

图5

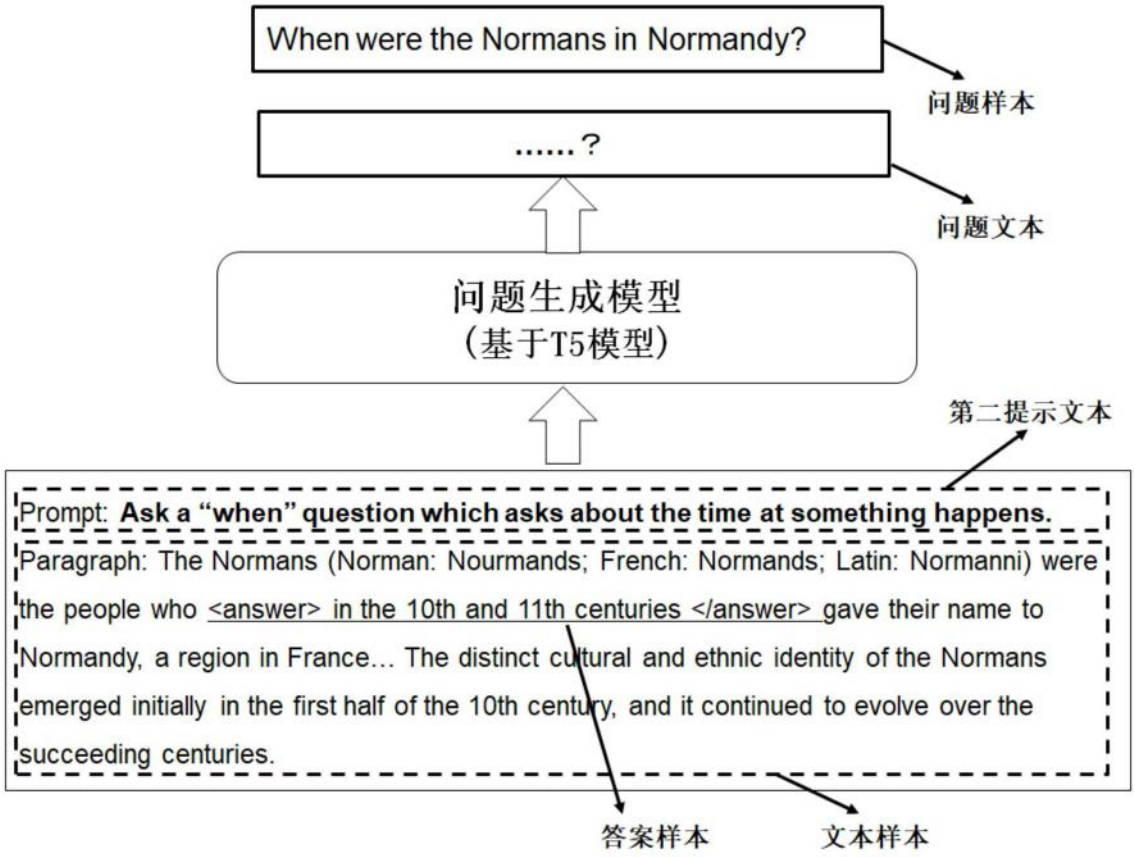

When were the Normans in Normandy?
问题样本
.....?
问题文本
问题生成模型
（基于T5模型）
第二提示文本
Prompt: Ask a “when” question which asks about the time at something happens.
Paragraph: The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who <answer> in the 10th and 11th centuries </answer> gave their name to Normandy, a region in France... The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.
答案样本
文本样本

图6

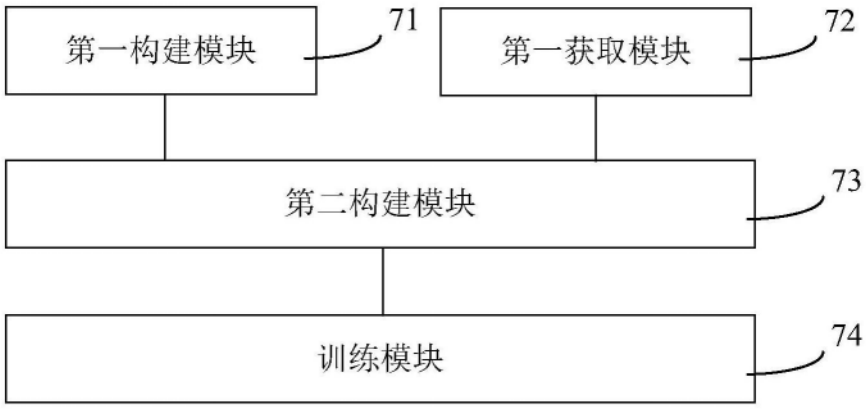

第一构建模块
71
第一获取模块
72
第二构建模块
73
训练模块
74

图7

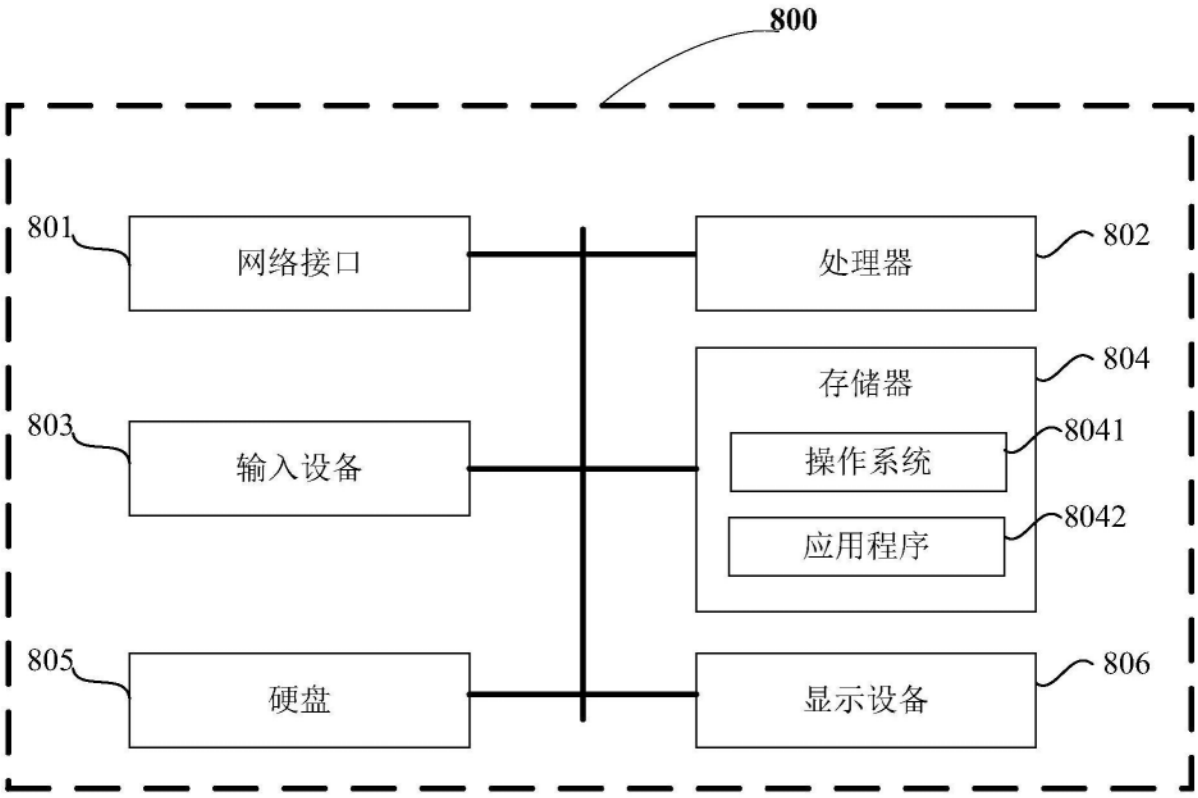

800
801
网络接口
802
处理器
804
存储器
803
输入设备
8041
操作系统
8042
应用程序
805
硬盘
806
显示设备

图8