(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109726274 B
(45)授权公告日 2021.04.30

(21)申请号 201811641895.1

(22)申请日 2018.12.29

(65)同一申请的已公布的文献号
申请公布号 CN 109726274 A

(43)申请公布日 2019.05.07

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 孙兴武 刘璟

(74)专利代理机构 北京市铸成律师事务所 11313
代理人 杨瑾瑾 陈建焕

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 16/35(2019.01)
G06F 40/211(2020.01)

(56)对比文件
CN 108363743 A,2018.08.03
CN 108846130 A,2018.11.20
US 2016371277 A1,2016.12.22
Xinya Du 等.Learning to Ask: Neural Question Generation for Reading Comprehension.《ACL》.2017,第1342-1352页.
Junwei Bao 等.Question Generation With Doubly Adversarial Nets.《IEEE》.2018,第2230-2239页.

审查员 王璐

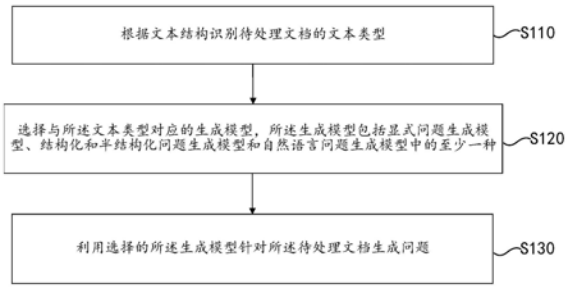
权利要求书4页 说明书16页 附图9页

(54)发明名称

问题生成方法、装置及存储介质

(57)摘要

本发明实施例提出一种问题生成方法、装置及计算机可读存储介质。其中问题生成方法包括:根据文本结构识别待处理文档的文本类型;选择与所述文本类型对应的生成模型,所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种;利用选择的所述生成模型针对所述待处理文档生成问题。本发明实施例针对不同的文本类型的特点,对于整篇文档而言,或者对整篇文档的各部分文本而言,都选择最适用的生成模型,提高了生成问题的准确率。



CN 109726274 B

1.一种问题生成方法,其特征在于,包括:

根据文本结构识别待处理文档的文本类型,所述文本类型包括显式FAQ文本类型、结构化和半结构化文本类型,以及自然语言文本类型中的至少一种;

选择与所述文本类型对应的生成模型,所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种;在所述文本类型为显式FAQ文本类型的情况下,将所述显式问题生成模型作为与所述文本类型对应的生成模型;在所述文本类型为结构化和半结构化文本类型的情况下,将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型;在所述文本类型为自然语言文本类型的情况下,将所述自然语言问题生成模型作为与所述文本类型对应的生成模型;

利用选择的所述生成模型针对所述待处理文档生成问题。

2.根据权利要求1所述的方法,其特征在于,

根据文本结构识别待处理文档的文本类型,包括:识别所述待处理文档的文本结构中是否有问答结构;

选择与所述文本类型对应的生成模型,包括:若所述待处理文档的文本结构中有问答结构,则将所述显式问题生成模型作为与所述文本类型对应的生成模型;

利用选择的所述生成模型针对所述待处理文档生成问题,包括:利用所述显式问题生成模型针对所述待处理文档生成问题。

3.根据权利要求2所述的方法,其特征在于,利用所述显式问题生成模型针对所述待处理文档生成问题,包括:

判断所述问答结构中的问题部分和对应的回答部分是否匹配,将匹配成功的所述问答结构对应的部分文本作为候选文本筛选出来;

利用第一循环神经网络模型,对筛选出的所述候选文本进行分类,以从所述候选文本中识别出显式问题;

将所述显式问题作为针对所述待处理文档生成的问题。

4.根据权利要求1所述的方法,其特征在于,

根据文本结构识别待处理文档的文本类型,包括:识别所述待处理文档的文本结构中是否有标题结构,所述标题结构包括标题或表格;

选择与所述文本类型对应的生成模型,包括:若所述待处理文档的文本结构中有标题结构,则将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型;

利用选择的所述生成模型针对所述待处理文档生成问题,包括:利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题。

5.根据权利要求4所述的方法,其特征在于,利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题,包括:

在所述待处理文档的文本结构中有标题的情况下,获取与所述标题相关的属性复述;

根据所述属性复述生成问题。

6.根据权利要求5所述的方法,其特征在于,获取与所述标题相关的属性复述,包括:

获取与所述标题相关的搜索点击展现日志;

对所述搜索点击展现日志进行数据挖掘,得到与所述标题相关的属性复述;

将所述属性复述存入属性复述表中。

7.根据权利要求6所述的方法,其特征在于,根据所述属性复述生成问题,包括:

根据所述属性复述,利用第一编码器-解码器模型生成问题;或者,

从所述属性复述表中查询与所述标题相关的属性复述,并根据查询到的所述属性复述生成问题。

8.根据权利要求1所述的方法,其特征在于,

根据文本结构识别待处理文档的文本类型,包括:识别所述待处理文档的文本结构中是否有问答结构和标题结构,所述标题结构包括标题或表格;

选择与所述文本类型对应的生成模型,包括:若所述待处理文档的文本结构中没有问答结构且没有标题结构,则将所述自然语言问题生成模型作为与所述文本类型对应的生成模型;

利用选择的所述生成模型针对所述待处理文档生成问题,包括:利用所述自然语言问题生成模型针对所述待处理文档生成问题。

9.根据权利要求8所述的方法,其特征在于,利用所述自然语言问题生成模型针对所述待处理文档生成问题,包括:

利用第二循环神经网络模型,从所述待处理文档中筛选出目标句子,所述目标句子包括语义完整的句子;

利用第三循环神经网络模型,从所述目标句子中选择候选答案片段;

根据所述候选答案片段,利用第二编码器-解码器模型生成问题。

10.根据权利要求1-9中任一项所述的方法,其特征在于,还包括:针对生成的所述问题进行答案边界定位。

11.根据权利要求10所述的方法,其特征在于,针对生成的所述问题进行答案边界定位,包括:

利用双向注意流网络预测所述问题对应的答案片段的起止位置;

利用学习排序模型将所述答案片段排序,根据排序结果对所述问题进行答案边界定位,其中,所述学习排序模型的特征包括所述答案片段的起止位置。

12.一种问题生成装置,其特征在于,包括:

文本类型识别单元,用于根据文本结构识别待处理文档的文本类型,所述文本类型包括显式FAQ文本类型、结构化和半结构化文本类型,以及自然语言文本类型中的至少一种;

生成模型选择单元,用于选择与所述文本类型对应的生成模型,所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种;在所述文本类型为显式FAQ文本类型的情况下,将所述显式问题生成模型作为与所述文本类型对应的生成模型;在所述文本类型为结构化和半结构化文本类型的情况下,将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型;在所述文本类型为自然语言文本类型的情况下,将所述自然语言问题生成模型作为与所述文本类型对应的生成模型;

问题生成单元,用于利用选择的所述生成模型针对所述待处理文档生成问题。

13.根据权利要求12所述的装置,其特征在于,

所述文本类型识别单元包括第一识别子单元,所述第一识别子单元用于:识别所述待处理文档的文本结构中是否有问答结构;

所述生成模型选择单元包括第一选择子单元，所述第一选择子单元用于：若所述待处理文档的文本结构中有问答结构，则将所述显式问题生成模型作为与所述文本类型对应的生成模型；

所述问题生成单元包括第一生成子单元，所述第一生成子单元用于：利用所述显式问题生成模型针对所述待处理文档生成问题。

14.根据权利要求13所述的装置，其特征在于，所述第一生成子单元还用于：

判断所述问答结构中的问题部分和对应的回答部分是否匹配，将匹配成功的所述问答结构对应的部分文本作为候选文本筛选出来；

利用第一循环神经网络模型，对筛选出的所述候选文本进行分类，以从所述候选文本中识别出显式问题；

将所述显式问题作为针对所述待处理文档生成的问题。

15.根据权利要求12所述的装置，其特征在于，

所述文本类型识别单元包括第二识别子单元，所述第二识别子单元用于：识别所述待处理文档的文本结构中是否有标题结构，所述标题结构包括标题或表格；

所述生成模型选择单元包括第二选择子单元，所述第二选择子单元用于：若所述待处理文档的文本结构中有标题结构，则将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型；

所述问题生成单元包括第二生成子单元，所述第二生成子单元用于：利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题。

16.根据权利要求15所述的装置，其特征在于，所述第二生成子单元包括：

复述获取子单元，用于在所述待处理文档的文本结构中有标题的情况下，获取与所述标题相关的属性复述；

复述问题生成子单元，用于根据所述属性复述生成问题。

17.根据权利要求16所述的装置，其特征在于，所述复述获取子单元还用于：

获取与所述标题相关的搜索点击展现日志；

对所述搜索点击展现日志进行数据挖掘，得到与所述标题相关的属性复述；

将所述属性复述存入属性复述表中。

18.根据权利要求17所述的装置，其特征在于，所述复述问题生成子单元还用于：

根据所述属性复述，利用第一编码器-解码器模型生成问题；或者，

从所述属性复述表中查询与所述标题相关的属性复述，并根据查询到的所述属性复述生成问题。

19.根据权利要求12所述的装置，其特征在于，

所述文本类型识别单元包括第三识别子单元，所述第三识别子单元用于：识别所述待处理文档的文本结构中是否有问答结构和标题结构，所述标题结构包括标题或表格；

所述生成模型选择单元包括第三选择子单元，所述第三选择子单元用于：若所述待处理文档的文本结构中没有问答结构且没有标题结构，则将所述自然语言问题生成模型作为与所述文本类型对应的生成模型；

所述问题生成单元包括第三生成子单元，所述第三生成子单元用于：利用所述自然语言问题生成模型针对所述待处理文档生成问题。

20.根据权利要求19所述的装置,其特征在于,所述第三生成子单元还用于:

利用第二循环神经网络模型,从所述待处理文档中筛选出目标句子,所述目标句子包括语义完整的句子;

利用第三循环神经网络模型,从所述目标句子中选择候选答案片段;

根据所述候选答案片段,利用第二编码器-解码器模型生成问题。

21.根据权利要求12-20中任一项所述的装置,其特征在于,还包括答案边界定位单元,用于针对生成的所述问题进行答案边界定位。

22.根据权利要求21所述的装置,其特征在于,所述答案边界定位单元还用于:

利用双向注意流网络预测所述问题对应的答案片段的起止位置;

利用学习排序模型将所述答案片段排序,根据排序结果对所述问题进行答案边界定位,其中,所述学习排序模型的特征包括所述答案片段的起止位置。

23.一种问题生成装置,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序;

当所述一个或多个程序被所述一个或多个处理器执行时,使得所述一个或多个处理器实现如权利要求1-11中任一所述的方法。

24.一种计算机可读存储介质,其存储有计算机程序,其特征在于,该程序被处理器执行时实现如权利要求1-11中任一所述的方法。

问题生成方法、装置及存储介质

技术领域

[0001] 本发明涉及信息技术领域，尤其涉及一种问题生成方法、装置及计算机可读存储介质。

背景技术

[0002] FAQ(Frequently Asked Questions，问答系统)是当前网络上提供在线帮助的主要手段，通过事先组织好一些可能的常问的问答对，发布在网页上为用户提供咨询服务。

[0003] 现有技术的FAQ实现方式主要包括以下几种：

[0004] (1)通用问答系统，基于检索或者基于知识的问答服务。

[0005] (2)定制化检索，对文档内容分段分词创建索引；或者，通过文档结构化或人工筛选的方法从而得到问答对。

[0006] (3)基于词匹配或同义词匹配的问题检索。

[0007] 现有技术的缺陷主要包括以下几个方面：

[0008] (1)基于检索或者基于知识的通用问答系统不能解决定制化的需求。

[0009] (2)对于通过对文档内容创建索引实现问答的方式，首先并非所有的内容都是问答内容，因此通篇存储会造成存储空间的浪费；其次是这种方式生成问题的准确率低，因为词命中不意味着当前内容是答案；还有无法判断答案边界和无法形成可视化FAQ文档。所谓可视化是指将对文本内容深度的阅读理解，从而提取出若干问答对，方便用户查找问题检索答案。现在的技术无法对篇章深度理解或者对文本生成好的问题。

[0010] (3)基于同义词匹配或词匹配的问题检索的泛化能力差且召回率低。

发明内容

[0011] 本发明实施例提供一种问题生成方法、装置及计算机可读存储介质，以至少解决现有技术中的一个或多个技术问题。

[0012] 第一方面，本发明实施例提供了一种问题生成方法，包括：

[0013] 根据文本结构识别待处理文档的文本类型；

[0014] 选择与所述文本类型对应的生成模型，所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种；

[0015] 利用选择的所述生成模型针对所述待处理文档生成问题。

[0016] 在一种实施方式中，根据文本结构识别待处理文档的文本类型，包括：识别所述待处理文档的文本结构中是否有问答结构；

[0017] 选择与所述文本类型对应的生成模型，包括：若所述待处理文档的文本结构中有问答结构，则将所述显式问题生成模型作为与所述文本类型对应的生成模型；

[0018] 利用选择的所述生成模型针对所述待处理文档生成问题，包括：利用所述显式问题生成模型针对所述待处理文档生成问题。

[0019] 在一种实施方式中，利用所述显式问题生成模型针对所述待处理文档生成问题，

包括：

[0020] 判断所述问答结构中的问题部分和对应的回答部分是否匹配，将匹配成功的所述问答结构对应的部分文本作为候选文本筛选出来；

[0021] 利用第一循环神经网络模型，对筛选出的所述候选文本进行分类，以从所述候选文本中识别出显式问题；

[0022] 将所述显式问题作为针对所述待处理文档生成的问题。

[0023] 在一种实施方式中，根据文本结构识别待处理文档的文本类型，包括：识别所述待处理文档的文本结构中是否有标题结构，所述标题结构包括标题或表格；

[0024] 选择与所述文本类型对应的生成模型，包括：若所述待处理文档的文本结构中有标题结构，则将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型；

[0025] 利用选择的所述生成模型针对所述待处理文档生成问题，包括：利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题。

[0026] 在一种实施方式中，利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题，包括：

[0027] 在所述待处理文档的文本结构中有标题的情况下，获取与所述标题相关的属性复述；

[0028] 根据所述属性复述生成问题。

[0029] 在一种实施方式中，获取与所述标题相关的属性复述，包括：

[0030] 获取与所述标题相关的搜索点击展现日志；

[0031] 对所述搜索点击展现日志进行数据挖掘，得到与所述标题相关的属性复述；

[0032] 将所述属性复述存入属性复述表中。

[0033] 在一种实施方式中，根据所述属性复述生成问题，包括：

[0034] 根据所述属性复述，利用第一编码器-解码器模型生成问题；或者，

[0035] 从所述属性复述表中查询与所述标题相关的属性复述，并根据查询到的所述属性复述生成问题。

[0036] 在一种实施方式中，根据文本结构识别待处理文档的文本类型，包括：识别所述待处理文档的文本结构中是否有问答结构和标题结构，所述标题结构包括标题或表格；

[0037] 选择与所述文本类型对应的生成模型，包括：若所述待处理文档的文本结构中没有问答结构且没有标题结构，则将所述自然语言问题生成模型作为与所述文本类型对应的生成模型；

[0038] 利用选择的所述生成模型针对所述待处理文档生成问题，包括：利用所述自然语言问题生成模型针对所述待处理文档生成问题。

[0039] 在一种实施方式中，利用所述自然语言问题生成模型针对所述待处理文档生成问题，包括：

[0040] 利用第二循环神经网络模型，从所述待处理文档中筛选出目标句子，所述目标句子包括语义完整的句子；

[0041] 利用第三循环神经网络模型，从所述目标句子中选择候选答案片段；

[0042] 根据所述候选答案片段，利用第二编码器-解码器模型生成问题。

[0043] 在一种实施方式中，所述方法还包括：针对生成的所述问题进行答案边界定位。

[0044] 在一种实施方式中，针对生成的所述问题进行答案边界定位，包括：

[0045] 利用双向注意流网络预测所述问题对应的答案片段的起止位置；

[0046] 利用学习排序模型将所述答案片段排序，根据排序结果对所述问题进行答案边界定位，其中，所述学习排序模型的特征包括所述答案片段的起止位置。

[0047] 第二方面，本发明实施例提供了一种问题生成装置，包括：

[0048] 文本类型识别单元，用于根据文本结构识别待处理文档的文本类型；

[0049] 生成模型选择单元，用于选择与所述文本类型对应的生成模型，所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种；

[0050] 问题生成单元，用于利用选择的所述生成模型针对所述待处理文档生成问题。

[0051] 在一种实施方式中，所述文本类型识别单元包括第一识别子单元，所述第一识别子单元用于：识别所述待处理文档的文本结构中是否有问答结构；

[0052] 所述生成模型选择单元包括第一选择子单元，所述第一选择子单元用于：若所述待处理文档的文本结构中有问答结构，则将所述显式问题生成模型作为与所述文本类型对应的生成模型；

[0053] 所述问题生成单元包括第一生成子单元，所述第一生成子单元用于：利用所述显式问题生成模型针对所述待处理文档生成问题。

[0054] 在一种实施方式中，所述第一生成子单元还用于：

[0055] 判断所述问答结构中的问题部分和对应的回答部分是否匹配，将匹配成功的所述问答结构对应的部分文本作为候选文本筛选出来；

[0056] 利用第一循环神经网络模型，对筛选出的所述候选文本进行分类，以从所述候选文本中识别出显式问题；

[0057] 将所述显式问题作为针对所述待处理文档生成的问题。

[0058] 在一种实施方式中，所述文本类型识别单元包括第二识别子单元，所述第二识别子单元用于：识别所述待处理文档的文本结构中是否有标题结构，所述标题结构包括标题或表格；

[0059] 所述生成模型选择单元包括第二选择子单元，所述第二选择子单元用于：若所述待处理文档的文本结构中有标题结构，则将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型；

[0060] 所述问题生成单元包括第二生成子单元，所述第二生成子单元用于：利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题。

[0061] 在一种实施方式中，所述第二生成子单元包括：

[0062] 复述获取子单元，用于在所述待处理文档的文本结构中有标题的情况下，获取与所述标题相关的属性复述；

[0063] 复述问题生成子单元，用于根据所述属性复述生成问题。

[0064] 在一种实施方式中，所述复述获取子单元还用于：

[0065] 获取与所述标题相关的搜索点击展现日志；

[0066] 对所述搜索点击展现日志进行数据挖掘，得到与所述标题相关的属性复述；

[0067] 将所述属性复述存入属性复述表中。

[0068] 在一种实施方式中，所述复述问题生成子单元还用于：

[0069] 根据所述属性复述，利用第一编码器-解码器模型生成问题；或者，

[0070] 从所述属性复述表中查询与所述标题相关的属性复述，并根据查询到的所述属性复述生成问题。

[0071] 在一种实施方式中，所述文本类型识别单元包括第三识别子单元，所述第三识别子单元用于：识别所述待处理文档的文本结构中是否有问答结构和标题结构，所述标题结构包括标题或表格；

[0072] 所述生成模型选择单元包括第三选择子单元，所述第三选择子单元用于：若所述待处理文档的文本结构中没有问答结构且没有标题结构，则将所述自然语言问题生成模型作为与所述文本类型对应的生成模型；

[0073] 所述问题生成单元包括第三生成子单元，所述第三生成子单元用于：利用所述自然语言问题生成模型针对所述待处理文档生成问题。

[0074] 在一种实施方式中，所述第三生成子单元还用于：

[0075] 利用第二循环神经网络模型，从所述待处理文档中筛选出目标句子，所述目标句子包括语义完整的句子；

[0076] 利用第三循环神经网络模型，从所述目标句子中选择候选答案片段；

[0077] 根据所述候选答案片段，利用第二编码器-解码器模型生成问题。

[0078] 在一种实施方式中，所述装置还包括答案边界定位单元，用于针对生成的所述问题进行答案边界定位。

[0079] 在一种实施方式中，所述答案边界定位单元还用于：

[0080] 利用双向注意流网络预测所述问题对应的答案片段的起止位置；

[0081] 利用学习排序模型将所述答案片段排序，根据排序结果对所述问题进行答案边界定位，其中，所述学习排序模型的特征包括所述答案片段的起止位置。

[0082] 第三方面，本发明实施例提供了一种问题生成装置，所述装置的功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括一个或多个与上述功能相对应的模块。

[0083] 在一个可能的设计中，所述装置的结构中包括处理器和存储器，所述存储器用于存储支持所述装置执行上述方法的程序，所述处理器被配置为用于执行所述存储器中存储的程序。所述装置还可以包括通信接口，用于与其他设备或通信网络通信。

[0084] 第四方面，本发明实施例提供了一种计算机可读存储介质，其存储有计算机程序，该程序被处理器执行时实现上述第一方面中任一所述的方法。

[0085] 上述技术方案中的一个技术方案具有如下优点或有益效果：针对不同的文本类型的特点，对于整篇文档而言，或者对整篇文档的各部分文本而言，都选择最适用的生成模型，提高了生成问题的准确率。

[0086] 上述技术方案中的另一个技术方案具有如下优点或有益效果：通过问答技术能够得到问题对应答案的准确边界，进一步提高了生成的FAQ文档的准确性。

[0087] 上述概述仅仅是为了说明书的目的，并不意图以任何方式进行限制。除上述描述的示意性的方面、实施方式和特征之外，通过参考附图和以下的详细描述，本发明进一步的

方面、实施方式和特征将会是容易明白的。

附图说明

[0088] 在附图中，除非另外规定，否则贯穿多个附图相同的附图标记表示相同或相似的部件或元素。这些附图不一定是按照比例绘制的。应该理解，这些附图仅描绘了根据本发明公开的一些实施方式，而不应将其视为是对本发明范围的限制。

[0089] 图1为本发明实施例提供的问题生成方法的流程图。

[0090] 图2为本发明实施例提供的问题生成方法的FAQ挖掘目标文档样例示意图。

[0091] 图3为本发明实施例提供的问题生成方法的文本类型的示意图。

[0092] 图4为本发明实施例提供的问题生成方法的利用显式问题生成模型生成问题的流程图。

[0093] 图5为本发明实施例提供的问题生成方法的利用显式问题生成模型生成问题的流程图。

[0094] 图6为本发明实施例提供的问题生成方法的利用结构化和半结构化问题生成模型生成问题的流程图。

[0095] 图7为本发明实施例提供的问题生成方法的利用结构化和半结构化问题生成模型生成问题的流程图。

[0096] 图8为本发明实施例提供的问题生成方法的利用结构化和半结构化问题生成模型生成问题的流程图。

[0097] 图9为本发明实施例提供的问题生成方法的利用自然语言问题生成模型生成问题的流程图。

[0098] 图10为本发明实施例提供的问题生成方法的利用自然语言问题生成模型生成问题的流程图。

[0099] 图11为本发明实施例提供的问题生成方法的流程图。

[0100] 图12为本发明实施例提供的问题生成方法的在线部分的示意图。

[0101] 图13为本发明实施例提供的问题生成方法的答案边界定位的流程图。

[0102] 图14为本发明实施例提供的问题生成方法的离线部分的示意图。

[0103] 图15为本发明实施例提供的问题生成装置的结构框图。

[0104] 图16为本发明实施例提供的问题生成装置的结构框图。

[0105] 图17为本发明实施例提供的问题生成装置的第二生成子单元的结构框图。

[0106] 图18为本发明实施例提供的问题生成装置的结构框图。

[0107] 图19为本发明实施例提供的问题生成装置的结构框图。

具体实施方式

[0108] 在下文中，仅简单地描述了某些示例性实施例。正如本领域技术人员可认识到的那样，在不脱离本发明的精神或范围的情况下，可通过各种不同方式修改所描述的实施例。因此，附图和描述被认为本质上是示例性的而非限制性的。

[0109] 图1为本发明实施例提供的问题生成方法的流程图。如图1所示，本发明实施例的问题生成方法包括：

[0110] 步骤S110,根据文本结构识别待处理文档的文本类型;

[0111] 步骤S120,选择与所述文本类型对应的生成模型,所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种;

[0112] 步骤S130,利用选择的所述生成模型针对所述待处理文档生成问题。

[0113] 在很多网站上都使用FAQ为用户提供咨询服务。在FAQ中列出了一些用户常见的问题,是一种在线帮助形式。用户在利用一些网站的功能或者服务时往往会遇到一些看似很简单、但不经过说明可能很难搞清楚的问题。有时甚至会因为这些细节问题的影响而失去用户。在很多情况下,只要经过简单的解释就可以解决这些问题,这就是FAQ的价值。

[0114] FAQ设计的问题和解答都必须是用户经常问到和遇到的。例如在网络营销中,FAQ被认为是一种常用的在线顾客服务手段,一个好的FAQ系统,应该至少可以回答用户80%以上的一般问题和常见问题。通过FAQ的使用,不仅方便了用户,也大大减轻了网站工作人员的压力,节省了大量的顾客服务成本,并且增加了用户的满意度。因此,在FAQ设计中,利用科学合理的方法生成高准确率的问题是至关重要的。

[0115] 图2为本发明实施例提供的问题生成方法的FAQ挖掘目标文档样例示意图。可从图2所示的指定文档中自动生成问答对,期望生成的问答对举例如下:

[0116] Q:吸二手烟对身体的危害有哪些?

[0117] A:1、增加肺癌几率…;2、对记忆力的危害,烟雾中的尼古丁…3、引发儿童哮喘病、肺炎、耳部炎症…。

[0118] Q:吸二手烟的人患肺癌的概率是正常人的多少倍?

[0119] A:2.6倍至6倍。

[0120] 以上,Q(question)表示问题,A(answer)表示问题对应的答案。

[0121] 本发明实施例通过定制化检索实现智能问答,可针对指定文档进行处理,以实现定制化检索的目的,使生成的智能问答更有针对性。通过进行阅读理解、问题生成技术和问答技术对待处理文档进行分析,提取出用户可能提问的问题,实现定制化问答服务。

[0122] 具体地,可将待处理的指定文档的文本分为三种文本类型:显式FAQ文本类型、结构化和半结构化文本类型、自然语言文本类型。图3为本发明实施例提供的问题生成方法的文本类型的表格示意图。图3中的“Q”表示各个文本样例对应的生成的问题。

[0123] 图3的表格中第一行所示的是显式FAQ文本类型,其对应的生成的问题是:“98元不限量可以定购本地、国内流量包吗?”

[0124] 图3的表格中第二行所示的是结构化和半结构化文本类型,其对应的生成的问题是:“百度女神卡办理途径有哪些?”其中,“结构化”可包括文档中有表格结构的情形,“半结构化”可包括文档中有标题的情形。生成的问题中的“百度女神卡”可根据文档的总标题或文档相关的词条、关键词获得。

[0125] 图3的表格中第三行所示的是自然语言文本类型,其对应的生成的问题是:“4G福卡新用户入网如何计费?”其中,生成的问题中的“4G福卡”可根据文档的总标题或文档相关的词条、关键词获得。

[0126] 本发明实施例针对不同文本类型的文档,分别选择与所述文本类型对应的不同的生成模型,使得生成的问题更有针对性,准确率更高,更加符合定制化问答服务的需求。利用生成的问题及其回答可支持客服机器人实现人机交互。

[0127] 在上述技术方案中，还可以对问题创建索引以节省存储空间。例如，创建索引的方式可包括倒排索引或key value(键值)索引。其中，倒排索引源于实际应用中需要根据属性的值来查找记录的情况。这种索引表中的每一项都包括一个属性值和具有该属性值的各记录的地址。由于不是由记录来确定属性值，而是由属性值来确定记录的位置，因而称为倒排索引。基于问题索引的查询答案的召回准确率更高。进一步地，还可以将生成的问答对，建立语义索引，用以支持客服机器人实现人机交互。

[0128] 在一个示例中，待处理文档可能是整篇文档都属于上述三种文本类型中的一种，则可选择其文本类型对应的生成模型针对整篇文档生成问题。在另一个示例中，待处理文档可以分为几个部分，而各部分文本的文本类型分别属于上述三种文本类型中的一种。在这种情况下可分别选择各部分文本的文本类型对应的生成模型，针对各部分文本分别生成问题。

[0129] 在一种实施方式中，根据文本结构识别待处理文档的文本类型，包括：根据文本结构识别待处理文档中的各部分文本的文本类型；

[0130] 选择与所述文本类型对应的生成模型，包括：选择与所述各部分文本的文本类型对应的生成模型；

[0131] 利用选择的所述生成模型针对所述待处理文档生成问题，包括：利用选择的所述生成模型针对所述各部分文本生成问题。

[0132] 首先根据文本结构识别待处理文档或待处理文档中的各部分属于哪种文本类型：

[0133] (1)识别待处理文档的文本结构中是否有问答结构，即问题和问题的回答。若有，则使用显式问题生成模型，针对待处理文档中的有问答结构的部分文本生成问题。

[0134] (2)识别待处理文档的文本结构中是否带有标题或表格，若有，则使用结构化和半结构化问题生成模型，针对待处理文档中的带有标题或表格的部分文本生成问题。

[0135] (3)若待处理文档中的部分文本中没有问答结构，也没有标题或表格，例如待处理文档中的部分文本是纯文本文档的格式，可将该部分文本称为纯文本部分。这种情况则使用自然语言问题生成模型，针对该部分文本生成问题。

[0136] 上述技术方案具有如下优点或有益效果：针对不同的文本类型的特点，对于整篇文档而言，或者对整篇文档的各部分文本而言，都选择最适用的生成模型，提高了生成问题的准确率。

[0137] 图4为本发明实施例提供的问题生成方法的利用显式问题生成模型生成问题的流程图。如图4所示，在一种实施方式中，图1中的步骤S110，根据文本结构识别待处理文档的文本类型，具体可包括步骤S210：识别所述待处理文档的文本结构中是否有问答结构。

[0138] 图1中的步骤S120，选择与所述文本类型对应的生成模型，具体可包括步骤S220：若所述待处理文档的文本结构中有问答结构，则将所述显式问题生成模型作为与所述文本类型对应的生成模型。

[0139] 图1中的步骤S130，利用选择的所述生成模型针对所述待处理文档生成问题，具体可包括步骤S230：利用所述显式问题生成模型针对所述待处理文档生成问题。

[0140] 参见图3表格中第一行所示的示例，若待处理文档的文本结构中有问答结构，则这部分文本的文本类型属于显式FAQ文本类型。对于显式FAQ文本类型，选择与之对应的显式问题生成模型针对该部分文本生成问题。

[0141] 图5为本发明实施例提供的问题生成方法的利用显式问题生成模型生成问题的流程图。如图5所示，在一种实施方式中，图4中的步骤S230，利用所述显式问题生成模型针对所述待处理文档生成问题，具体可包括：

[0142] 步骤S310，判断所述问答结构中的问题部分和对应的回答部分是否匹配，将匹配成功的所述问答结构对应的部分文本作为候选文本筛选出来；

[0143] 步骤S320，利用第一循环神经网络模型，对筛选出的所述候选文本进行分类，以从所述候选文本中识别出显式问题；

[0144] 步骤S330，将所述显式问题作为针对所述待处理文档生成的问题。

[0145] 在这种实施方式中，可通过文档结构解析和问题识别来生成问题。其中，文档结构解析的过程可包括用文本结构初筛出可能的显式问题。问题识别的过程可包括：将RNN(Recurrent Neural Network，循环神经网络)作为问题分类模型，以词为特征，使用人工标注训练数据，训练上述RNN模型，作为显式问题生成模型。

[0146] 具体地，显式问题生成模型的处理过程可分为以下步骤：

[0147] (1)文档结构解析：找出文本中的问题部分和对应的回答部分，并判断问题部分和回答部分是否匹配。若问题部分和回答部分匹配，则将问题部分和回答部分作为“可能的显式问题”筛选出来。

[0148] 例如：“找出文本中的问题部分和对应的回答部分”的方法可包括：找出问号，将问号之前的一句话确定为问题部分，将问号之后的一段话确定为回答部分。可通过语义理解判断问题和回答是否匹配。

[0149] (2)问题识别：利用RNN模型，即上述第一循环神经网络模型，对筛选出的“可能的显式问题”进行分类。第一循环神经网络模型输出的结果可分为两类，一类是最终确定其是显式问题，另一类是最终确定其不是显式问题。

[0150] 上述RNN模型使用的特征可包括：词性和词的依存关系。其中，可将文本进行分句、切词，根据需求提取不同的特征，作为训练模型的特征。例如，利用标点进行分句；利用NLP(Natural Language Processing，自然语言处理)进行切词。可以分别尝试将一个句子可以分成例如5个词、3个词或1个词的模式，选取效果最好的模式进行分词。

[0151] 图6为本发明实施例提供的问题生成方法的利用结构化和半结构化问题生成模型生成问题的流程图。如图6所示，在一种实施方式中，图1中的步骤S110，根据文本结构识别待处理文档的文本类型，具体可包括步骤S410：识别所述待处理文档的文本结构中是否有标题结构，所述标题结构包括标题或表格。

[0152] 图1中的步骤S120，选择与所述文本类型对应的生成模型，具体可包括步骤S420：若所述待处理文档的文本结构中有标题结构，则将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型。

[0153] 图1中的步骤S130，利用选择的所述生成模型针对所述待处理文档生成问题，具体可包括步骤S430：利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题。

[0154] 参见图3表格中第二行所示的示例，若待处理文档的文本结构中有标题结构，其中标题结构包括标题或表格，则这部分文本的文本类型属于结构化和半结构化文本类型。对于结构化和半结构化文本类型，选择与之对应的结构化和半结构化问题生成模型针对该部

分文本生成问题。

[0155] 图7为本发明实施例提供的问题生成方法的利用结构化和半结构化问题生成模型生成问题的流程图。如图7所示，在一种实施方式中，图6中的步骤S430，利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题，具体可包括：

[0156] 步骤S510，在所述待处理文档的文本结构中有标题的情况下，获取与所述标题相关的属性复述；

[0157] 步骤S520，根据所述属性复述生成问题。

[0158] 在这种实施方式中，可获取FAQ的检索系统中的搜索点击展现日志，依赖搜索点击展现日志挖掘属性的复述。例如：*计费方式->*如何计费，这两个属性可以互为复述。基于属性复述的生成，作为半结构化和结构化的问题生成模型。

[0159] 图8为本发明实施例提供的问题生成方法的利用结构化和半结构化问题生成模型生成问题的流程图。如图8所示，在一种实施方式中，图7中的步骤S510，获取与所述标题相关的属性复述，具体可包括：

[0160] 步骤S610，获取与所述标题相关的搜索点击展现日志；

[0161] 步骤S620，对所述搜索点击展现日志进行数据挖掘，得到与所述标题相关的属性复述；

[0162] 步骤S630，将所述属性复述存入属性复述表中。

[0163] 在一种实施方式中，图7中的步骤S520，根据所述属性复述生成问题，具体可包括：

[0164] 根据所述属性复述，利用第一编码器-解码器模型生成问题；或者，

[0165] 从所述属性复述表中查询与所述标题相关的属性复述，并根据查询到的所述属性复述生成问题。

[0166] 具体地，结构和半结构化生成模型可包括以下处理步骤：

[0167] (1)依赖搜索的点击展现日志，利用数据挖掘的方法挖掘属性的复述。将挖掘的属性的复述存入属性复述表中。

[0168] 例如两个用户A和B使用了不同的关键词搜索后，点击了同一个URL(Uniform Resource Locator，统一资源定位符)。则这两个不同的关键词所表达的意思可能是相同的，可以互为复述。

[0169] 在一个示例中，用户搜索“苦瓜的功效”。其中“苦瓜”是实体，“功效”是实体的属性。“功效”这个属性的复述还包括“作用”、“药效”等。也就是说“苦瓜的功效”、“苦瓜的作用”和“苦瓜的药效”所表达的意思是相同的。

[0170] (2)基于属性复述的生成，作为半结构化和结构化的问题生成模型。可包括两种实施方式：

[0171] 方式一：使用Seq2Seq(Sequence to Sequence，序列到序列)模型，即上述第一编码器-解码器模型，生成问题。上述模型使用的特征包括词法和句法特征，序列标注模型预测的答案起止位置，以及词特征。上述模型的输入信息是待处理文档的段落，输出信息是针对输入信息生成的问题。在一个示例中，待处理文档中的文本内容为：“北京是中国的首都。”则序列标注模型可标注出“北京”。然后再把“北京”和“北京是中国的首都”作为输入信息，输入给seq2seq模型以生成问题“中国的首都是哪儿”。

[0172] Seq2Seq模型，也可以称之为Encoder-Decoder模型(编码器-解码器模型)，它是

RNN模型的一个重要的变种。Encoder-Decoder结构不限制输入和输出的序列长度，因此应用的范围非常广泛，比如：机器翻译、文本摘要、阅读理解、语音识别等。

[0173] 由于seq2seq模型可以是输入和输出序列不等长，即Sequence to Sequence，因此它实现了从一个序列到另外一个序列的转换，比如它可以实现聊天机器人对话模型。经典的RNN模型固定了输入序列和输出序列的大小，而seq2seq模型则突破了该限制。

[0174] 编码器(encoder)和解码器(decoder)分别对应着输入序列和输出序列的两个RNN。常见的encoder-decoder结构，其基本思想就是利用两个RNN，一个RNN作为encoder，另一个RNN作为decoder。encoder负责将输入序列压缩成指定长度的向量，这个向量可以看成是这个序列的语义，这个过程称为编码。而decoder则负责根据语义向量生成指定的序列，这个过程也称为解码。

[0175] 方式二：查询属性复述表，选择一个相关的复述用以生成问题。

[0176] 例如，对于查询到的复述："苦瓜的功效"、"苦瓜的作用"和"苦瓜的药效"，可以从中选择一个最适合的复述用以生成问题。可利用语义理解、关键词匹配等方法，将查询到的各个复述与待处理文档中的表述做比对、分析，以确定用以生成问题的最适合的复述。

[0177] 另一种情况，对于所述待处理文档的文本结构中有表格的情形，可利用语义理解、关键词匹配等方法，识别表格中表头、各行记录、各列字段的内容，进而确定生成的问题及其对应的答案。例如，表格中的其中一列数据的内容可能是生成的问题，而另外一列数据的内容可能是其对应的答案。

[0178] 图9为本发明实施例提供的问题生成方法的利用自然语言问题生成模型生成问题的流程图。如图9所示，在一种实施方式中，图1中的步骤S110，根据文本结构识别待处理文档的文本类型，具体可包括步骤S710：识别所述待处理文档的文本结构中是否有问答结构和标题结构，所述标题结构包括标题或表格；

[0179] 图1中的步骤S120，选择与所述文本类型对应的生成模型，具体可包括步骤S720：若所述待处理文档的文本结构中没有问答结构且没有标题结构，则将所述自然语言问题生成模型作为与所述文本类型对应的生成模型；

[0180] 图1中的步骤S130，利用选择的所述生成模型针对所述待处理文档生成问题，具体可包括步骤S730：利用所述自然语言问题生成模型针对所述待处理文档生成问题。

[0181] 参见图3表格中第三行所示的示例，若待处理文档的文本结构中没有问答结构且没有标题结构，则这部分文本的文本类型属于自然语言文本类型。对于自然语言文本类型，选择与之对应的自然语言问题生成模型针对该部分文本生成问题。

[0182] 图10为发明实施例提供的问题生成方法的利用自然语言问题生成模型生成问题的流程图。如图10所示，在一种实施方式中，图9中的步骤S730，利用所述自然语言问题生成模型针对所述待处理文档生成问题，具体可包括：

[0183] 步骤S810，利用第二循环神经网络模型，从所述待处理文档中筛选出目标句子，所述目标句子包括语义完整的句子；

[0184] 步骤S820，利用第三循环神经网络模型，从所述目标句子中选择候选答案片段；

[0185] 步骤S830，根据所述候选答案片段，利用第二编码器-解码器模型生成问题。

[0186] 在这种实施方式中，可首先利用RNN模型分类筛选目标句子，针对筛选出的目标句子使用RNN序列标注选择候选答案片段，然后再用seq2seq模型生成问题。

[0187] 具体地，自然语言生成模型可包括以下处理步骤：

[0188] (1)首先利用第二个循环神经网络模型分类筛选目标句子，筛选出语义完整的句子。

[0189] (2)对筛选的目标句子，利用第三个循环神经网络模型选择候选答案片段，也就是选择出问题和对应的可能是答案的片段。其中，可利用序列标注训练第三循环神经网络模型。序列标注可包括句子标注，也就是标注出可以提问的问题。

[0190] (3)再利用seq2seq模型，即上述第二编码器-解码器模型，问题生成模型生成问题。

[0191] 在一个示例中，待处理文档中的文本内容为：“北京是中国的首都。”则序列标注可标注出“北京”。然后再把“北京”和“北京是中国的首都”作为输入信息，输入给seq2seq模型以生成问题“中国的首都是哪儿”。

[0192] 图11为本发明实施例提供的问题生成方法的流程图。如图11所示，在一种实施方式中，所述方法还包括步骤S140：针对生成的所述问题进行答案边界定位。

[0193] 上述技术方案具有如下优点或有益效果：通过问答技术，例如通过答案边界定位，能够得到问题对应答案的准确边界，进一步提高了生成的FAQ文档的准确性。

[0194] 在一个示例中，本发明实施例的问题生成方法包括两部分：在线部分的问题生成方法和离线部分的问题生成方法。图12为本发明实施例提供的问题生成方法的在线部分的示意图。图12中的“目标文档”也就是待处理文档。图12中的“通用文档解析”包括识别文档结构。具体地，“通用文档解析”可包括识别文档的主题(标题)、子标题、子标题下的段落和文本，可利用树形结构描述识别出的文档结构。“通用文档解析”还包括识别待处理文档中的各部分属于上述三种文本类型(显式FAQ文本类型、结构化和半结构化文本类型、自然语言文本类型)中的哪一种。识别出待处理文档中的各部分的文本类型之后，在下一步骤中选择对应的问题生成模型生成问题并进行答案边界定位。

[0195] 图13为本发明实施例提供的问题生成方法的答案边界定位的流程图。如图13所示，在一种实施方式中，图11中的步骤S140，针对生成的所述问题进行答案边界定位，具体可包括：

[0196] 步骤S910，利用双向注意流网络预测所述问题对应的答案片段的起止位置；

[0197] 步骤S920，利用学习排序模型将所述答案片段排序，根据排序结果对所述问题进行答案边界定位，其中，所述学习排序模型的特征包括所述答案片段的起止位置。

[0198] 具体地，答案边界定位可包括以下处理步骤：

[0199] (1)采用Bi-DAF(Bi-Directional Attention Flow network，双向注意流网络)作为阅读理解模型，能够准确的预测答案的起止位置。

[0200] 双向注意流网络是一种层次化的多阶段的结构，可在不同粒度等级上对上下文进行建模。双向注意流网络主要包括在字符粒度等级(Character Level)和词粒度水平(Word Level)上对上下文进行建模，并且使用双向注意流来获取问题-察觉的上下文的表示方法。

[0201] 一个示例性的双向注意流网络可包括以下层次：

[0202] 1.字符嵌入层(Character embedding layer)

[0203] 该层的主要作用是将词映射到一个固定大小的向量，该层可使用字符水平的卷积神经网络(Character level CNN)实现其功能。

[0204] 2.词嵌入层(Word embedding layer)

[0205] 可使用预先训练的词嵌入模型,将每一个词映射到固定大小的向量。

[0206] 3.上下文嵌入层(Contextual embedding layer)

[0207] 该层主要作用是给每一个词加一个上下文的线索(cue),前三层都是对问题和上下文进行应用。

[0208] 4.注意流层(Attention flow layer)

[0209] 组合问题和上下文的向量,生成一个问题-察觉的特征向量集合。

[0210] 5.模型层(Modeling layer)

[0211] 可使用循环神经网络对上下文进行扫描。

[0212] 6.输出层(Output layer)

[0213] 该层提供对问题的回答。

[0214] (2)采用LTR(Learning to rank,学习排序),进行答案片段排序。

[0215] 其中,将步骤(1)预测出的起止位置作为步骤(2)中LTR的一个特征。步骤(2)采用LTR从问题后面的大段文本中寻找问题对应的答案。其中,LTR模型根据问答特征对答案片段进行排序。

[0216] 其中,学习排序是一种监督学习的排序方法。利用LTR可通过构造相关度函数,按照相关度进行排序。对于传统的排序方法,很难融合多种信息,而且很可能会出现过拟合现象。学习排序很容易融合多种特征,而且有成熟深厚的理论基础,参数是通过迭代优化出来的,有一套成熟理论解决稀疏、过拟合等问题。

[0217] 在这种实施方式中,首先对目标文档分段,可识别自然段落或使用列表提取的方法进行分段。然后对段落提取特征并排序,如使用“领域特征”和/或“匹配特征”等相关工具提取特征。其中特征可包括以下几种:

[0218] 问题答案匹配特征:对齐匹配技术、DNN(Deep Neural Networks,深度神经网络)QP匹配技术、结合知识图谱的QP匹配技术,其中Q表示问题(Query),P表示分段的段落(Paragrap),对齐匹配包括Q和P对齐;

[0219] 领域特征:实体问答特征、how why问答特征、是非问答特征、描述类问答特征;

[0220] 结构特征:列表结构特征;

[0221] 文本特征:内容质量特征;

[0222] 交叉校验特征:文本聚合特征。

[0223] 图14为本发明实施例提供的问题生成方法的离线部分的示意图。如图14所示,离线部分的主要生成模型和数据,可包括两个部分和五个模型。其中,两个部分包括文档标题数据和问答标注数据。五个模型包括显式问题生成模型、结构化和半结构化问题生成模型、自然语言问题生成模型、Bi-DAF模型(阅读理解模型)和LTR模型(答案片段排序模型)。

[0224] 图15为本发明实施例提供的问题生成装置的结构框图。如图15所示,本发明实施例的问题生成装置包括:

[0225] 文本类型识别单元100,用于根据文本结构识别待处理文档的文本类型;

[0226] 生成模型选择单元200,用于选择与所述文本类型对应的生成模型,所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种;

[0227] 问题生成单元300，用于利用选择的所述生成模型针对所述待处理文档生成问题。

[0228] 图16为本发明实施例提供的问题生成装置的结构框图。如图16所示，在一种实施方式中，所述文本类型识别单元100包括第一识别子单元110，所述第一识别子单元110用于：识别所述待处理文档的文本结构中是否有问答结构；

[0229] 所述生成模型选择单元200包括第一选择子单元210，所述第一选择子单元210用于：若所述待处理文档的文本结构中有问答结构，则将所述显式问题生成模型作为与所述文本类型对应的生成模型；

[0230] 所述问题生成单元300包括第一生成子单元310，所述第一生成子单元310用于：利用所述显式问题生成模型针对所述待处理文档生成问题。

[0231] 在一种实施方式中，所述第一生成子单元310还用于：

[0232] 判断所述问答结构中的问题部分和对应的回答部分是否匹配，将匹配成功的所述问答结构对应的部分文本作为候选文本筛选出来；

[0233] 利用第一循环神经网络模型，对筛选出的所述候选文本进行分类，以从所述候选文本中识别出显式问题；

[0234] 将所述显式问题作为针对所述待处理文档生成的问题。

[0235] 在一种实施方式中，所述文本类型识别单元100包括第二识别子单元120，所述第二识别子单元120用于：识别所述待处理文档的文本结构中是否有标题结构，所述标题结构包括标题或表格；

[0236] 所述生成模型选择单元200包括第二选择子单元220，所述第二选择子单元220用于：若所述待处理文档的文本结构中有标题结构，则将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型；

[0237] 所述问题生成单元300包括第二生成子单元320，所述第二生成子单元320用于：利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题。

[0238] 图17为本发明实施例提供的问题生成装置的第二生成子单元的结构框图。如图17所示，在一种实施方式中，所述第二生成子单元320包括：

[0239] 复述获取子单元321，用于在所述待处理文档的文本结构中有标题的情况下，获取与所述标题相关的属性复述；

[0240] 复述问题生成子单元322，用于根据所述属性复述生成问题。

[0241] 在一种实施方式中，所述复述获取子单元321还用于：

[0242] 获取与所述标题相关的搜索点击展现日志；

[0243] 对所述搜索点击展现日志进行数据挖掘，得到与所述标题相关的属性复述；

[0244] 将所述属性复述存入属性复述表中。

[0245] 在一种实施方式中，所述复述问题生成子单元322还用于：

[0246] 根据所述属性复述，利用第一编码器-解码器模型生成问题；或者，

[0247] 从所述属性复述表中查询与所述标题相关的属性复述，并根据查询到的所述属性复述生成问题。

[0248] 参见图16，在一种实施方式中，所述文本类型识别单元100包括第三识别子单元130，所述第三识别子单元130用于：识别所述待处理文档的文本结构中是否有问答结构和标题结构，所述标题结构包括标题或表格；

[0249] 所述生成模型选择单元200包括第三选择子单元230，所述第三选择子单元230用于：若所述待处理文档的文本结构中没有问答结构且没有标题结构，则将所述自然语言问题生成模型作为与所述文本类型对应的生成模型；

[0250] 所述问题生成单元300包括第三生成子单元330，所述第三生成子单元330用于：利用所述自然语言问题生成模型针对所述待处理文档生成问题。

[0251] 在一种实施方式中，所述第三生成子单元330还用于：

[0252] 利用第二循环神经网络模型，从所述待处理文档中筛选出目标句子，所述目标句子包括语义完整的句子；

[0253] 利用第三循环神经网络模型，从所述目标句子中选择候选答案片段；

[0254] 根据所述候选答案片段，利用第二编码器-解码器模型生成问题。

[0255] 图18为本发明实施例提供的问题生成装置的结构框图。如图18所示，在一种实施方式中，所述装置还包括答案边界定位单元400，用于针对生成的所述问题进行答案边界定位。

[0256] 在一种实施方式中，所述答案边界定位单元400还用于：

[0257] 利用双向注意流网络预测所述问题对应的答案片段的起止位置；

[0258] 利用学习排序模型将所述答案片段排序，根据排序结果对所述问题进行答案边界定位，其中，所述学习排序模型的特征包括所述答案片段的起止位置。

[0259] 本发明实施例的问题生成装置中各单元的功能可以参见上述方法的相关描述，在此不再赘述。

[0260] 在一个可能的设计中，问题生成装置的结构中包括处理器和存储器，所述存储器用于存储支持问题生成装置执行上述问题生成方法的程序，所述处理器被配置为用于执行所述存储器中存储的程序。所述问题生成装置还可以包括通信接口，问题生成装置与其他设备或通信网络通信。

[0261] 图19为本发明实施例提供的问题生成装置的结构框图。如图19所示，该装置包括：存储器101和处理器102，存储器101内存储有可在处理器102上运行的计算机程序。所述处理器102执行所述计算机程序时实现上述实施例中的问题生成方法。所述存储器101和处理器102的数量可以为一个或多个。

[0262] 该装置还包括：

[0263] 通信接口103，用于与外界设备进行通信，进行数据交互传输。

[0264] 存储器101可能包含高速RAM存储器，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。

[0265] 如果存储器101、处理器102和通信接口103独立实现，则存储器101、处理器102和通信接口103可以通过总线相互连接并完成相互间的通信。所述总线可以是工业标准体系结构(ISA，Industry Standard Architecture)总线、外部设备互连(PCI，Peripheral Component)总线或扩展工业标准体系结构(EISA，Extended Industry Standard Component)总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示，图19中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0266] 可选的，在具体实现上，如果存储器101、处理器102及通信接口103集成在一块芯片上，则存储器101、处理器102及通信接口103可以通过内部接口完成相互间的通信。

[0267] 又一方面，本发明实施例提供了一种计算机可读存储介质，其存储有计算机程序，该程序被处理器执行时实现上述问题生成方法中任一所述的方法。

[0268] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0269] 此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或隐含地包括至少一个该特征。在本发明的描述中，“多个”的含义是两个或两个以上，除非另有明确具体的限定。

[0270] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本发明的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本发明的实施例所属技术领域的技术人员所理解。

[0271] 在流程图中表示或在此以其他方式描述的逻辑和/或步骤，例如，可以被认为是用于实现逻辑功能的可执行指令的定序列表，可以具体实现在任何计算机可读介质中，以供指令执行系统、装置或设备(如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、装置或设备取指令并执行指令的系统)使用，或结合这些指令执行系统、装置或设备而使用。就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例(非穷尽性列表)包括以下：具有一个或多个布线的电连接部(电子装置)，便携式计算机盘盒(磁装置)，随机存取存储器(RAM)，只读存储器(ROM)，可擦除可编辑只读存储器(EPROM或闪速存储器)，光纤装置，以及便携式只读存储器(CDROM)。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为可以例如通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

[0272] 应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列(PGA)，现场可编程门阵列(FPGA)等。

[0273] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

[0274] 此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，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模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读存储介质中。所述存储介质可以是只读存储器，磁盘或光盘等。

[0275] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到其各种变化或替换，这些都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以所述权利要求的保护范围为准。

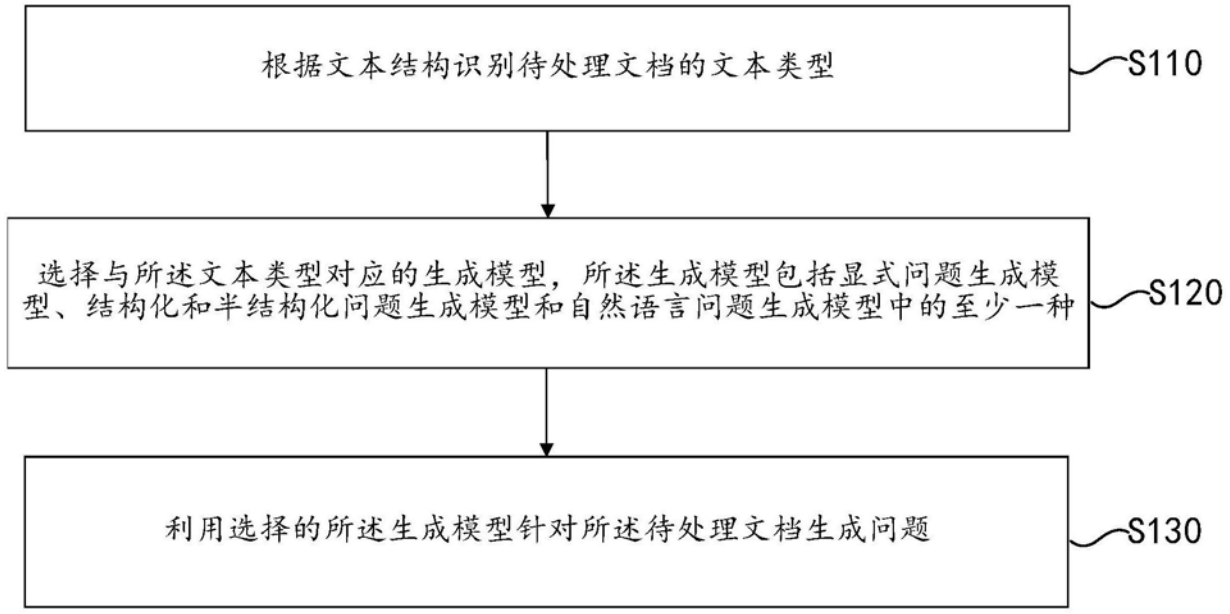

根据文本结构识别待处理文档的文本类型
S110
选择与所述文本类型对应的生成模型，所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种
S120
利用选择的所述生成模型针对所述待处理文档生成问题
S130

图1

图2

类型	样例
显式FAQ	1、98元不限量是否可以定购本地、国内流量包？ 答： 98和本地流量包互斥；与全国流量包不互斥。可以订购国际漫游包 Q：98元不限量可以定购本地、国内流量包吗？
结构化（表格）和半结构化（标题）	六、受理渠道 1、线下渠道：自有营业厅、合作营业厅 2、电话渠道：各分公司维系经理电话外呼、10010在线受理、10010正华外呼 3、线上渠道：网厅、手厅、短信、沃视窗、微信、沃管家、本地短厅。 Q：百度女神卡办理途径有哪些？
自然语言	（7）新用户入网首月计费规则： ➢ 入网当月套餐资费可选择全月套餐、套餐减半、或按量计费三种方式，次月按所选套餐计费； ➢ 全月套餐指申请套餐即时生效，用户入网当月即按照用户所选的套餐收取套餐月费，套餐所含内容不变。 ➢ 套餐减半指用户入网当月在其所选套餐的基础上，对套餐月费和套餐所含按量计费的业务内容均减半。 Q：4G福卡新用户入网如何计费？

图3

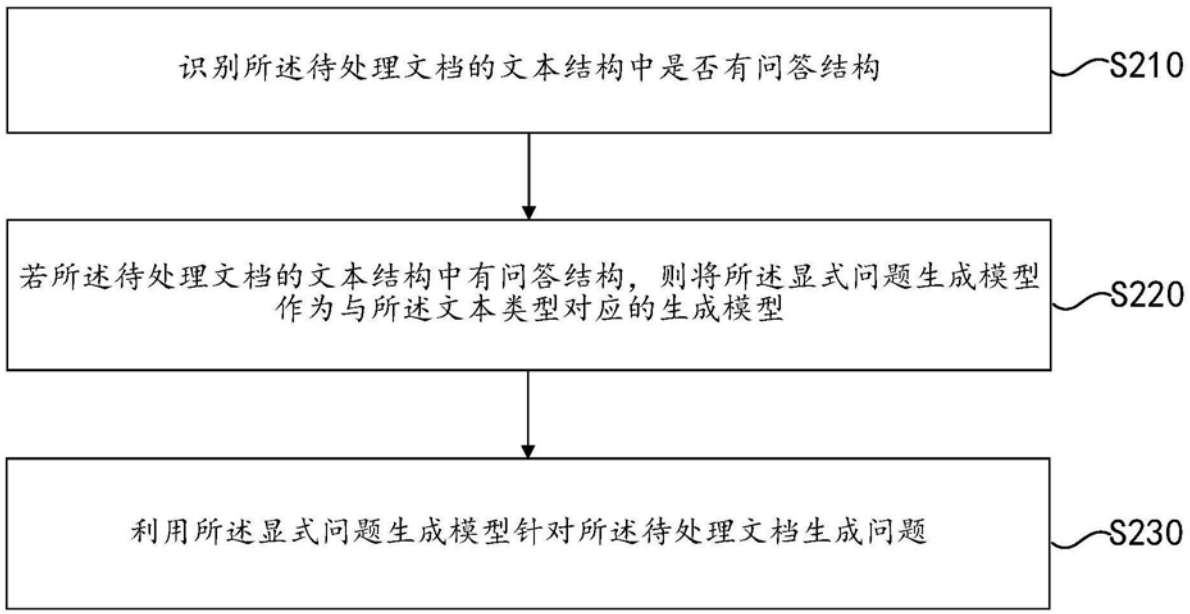

识别所述待处理文档的文本结构中是否有问答结构
S210
若所述待处理文档的文本结构中有问答结构，则将所述显式问题生成模型作为与所述文本类型对应的生成模型
S220
利用所述显式问题生成模型针对所述待处理文档生成问题
S230

图4

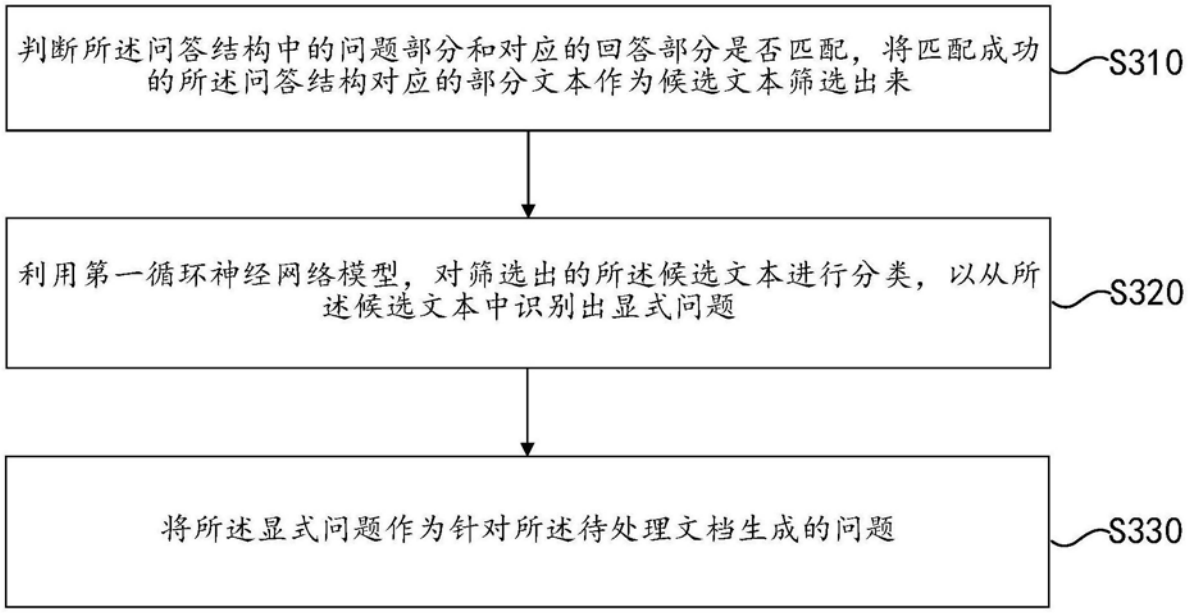

判断所述问答结构中的问题部分和对应的回答部分是否匹配，将匹配成功的所述问答结构对应的部分文本作为候选文本筛选出来
S310
利用第一循环神经网络模型，对筛选出的所述候选文本进行分类，以从所述候选文本中识别出显式问题
S320
将所述显式问题作为针对所述待处理文档生成的问题
S330

图5

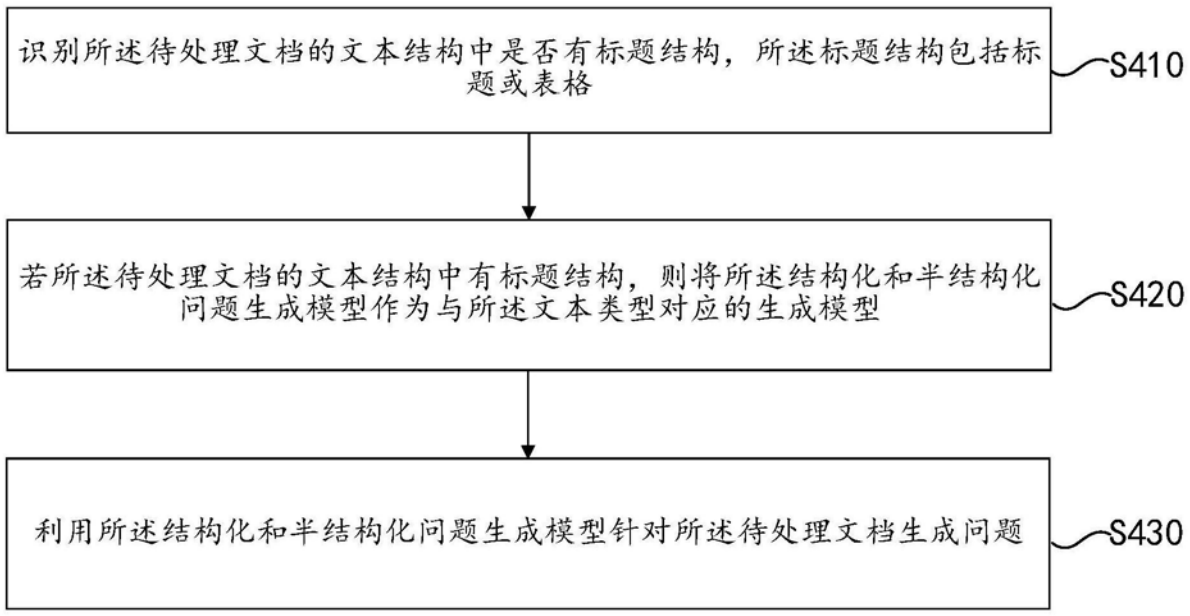

识别所述待处理文档的文本结构中是否有标题结构，所述标题结构包括标题或表格
S410
若所述待处理文档的文本结构中有标题结构，则将所述结构化和半结构化问题生成模型作为与所述文本类型对应的生成模型
S420
利用所述结构化和半结构化问题生成模型针对所述待处理文档生成问题
S430

图6

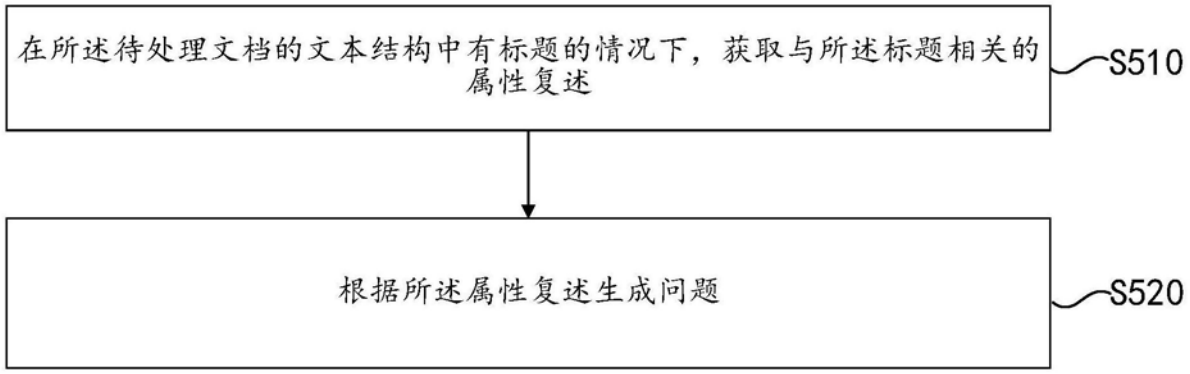

在所述待处理文档的文本结构中有标题的情况下，获取与所述标题相关的属性复述
S510
根据所述属性复述生成问题
S520

图7

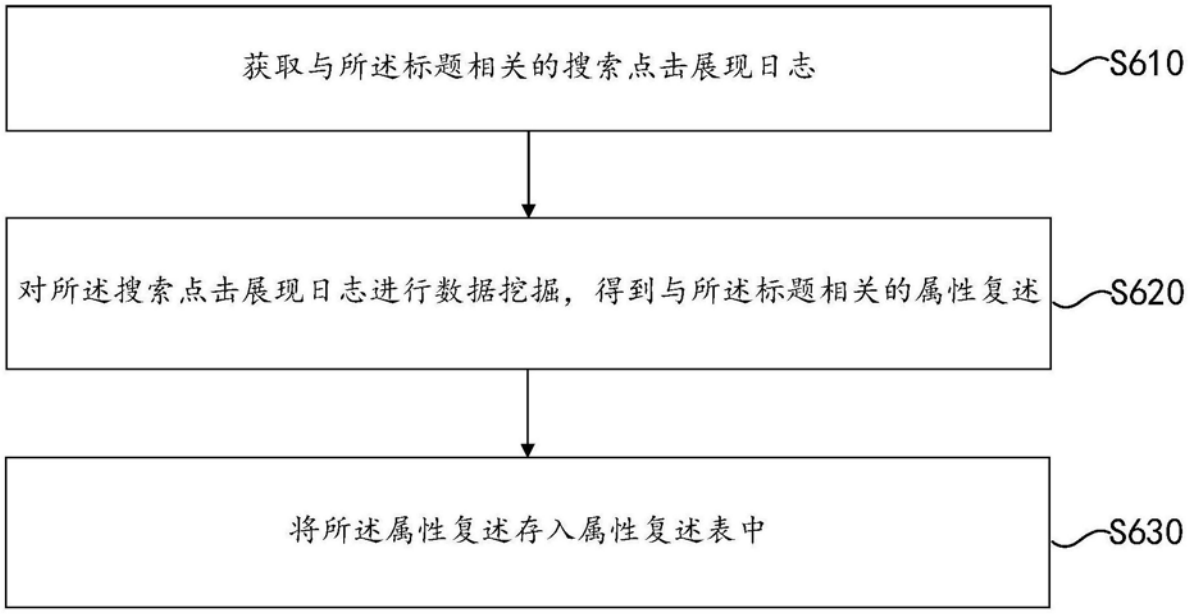

获取与所述标题相关的搜索点击展现日志
S610
对所述搜索点击展现日志进行数据挖掘，得到与所述标题相关的属性复述
S620
将所述属性复述存入属性复述表中
S630

图8

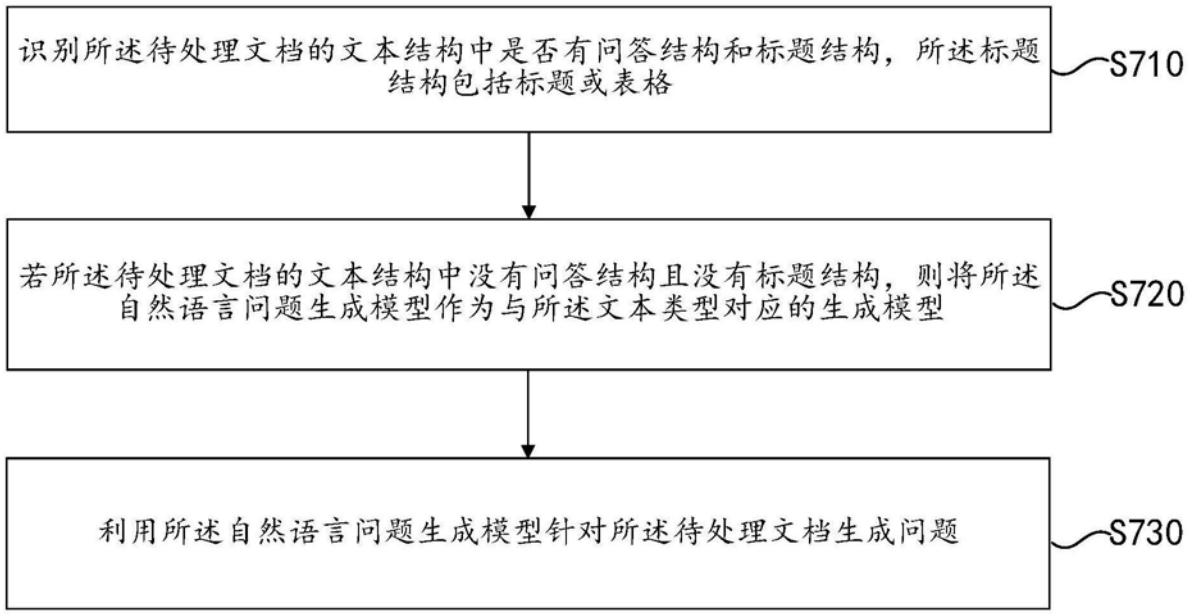

识别所述待处理文档的文本结构中是否有问答结构和标题结构，所述标题结构包括标题或表格
S710
若所述待处理文档的文本结构中没有问答结构且没有标题结构，则将所述自然语言问题生成模型作为与所述文本类型对应的生成模型
S720
利用所述自然语言问题生成模型针对所述待处理文档生成问题
S730

图9

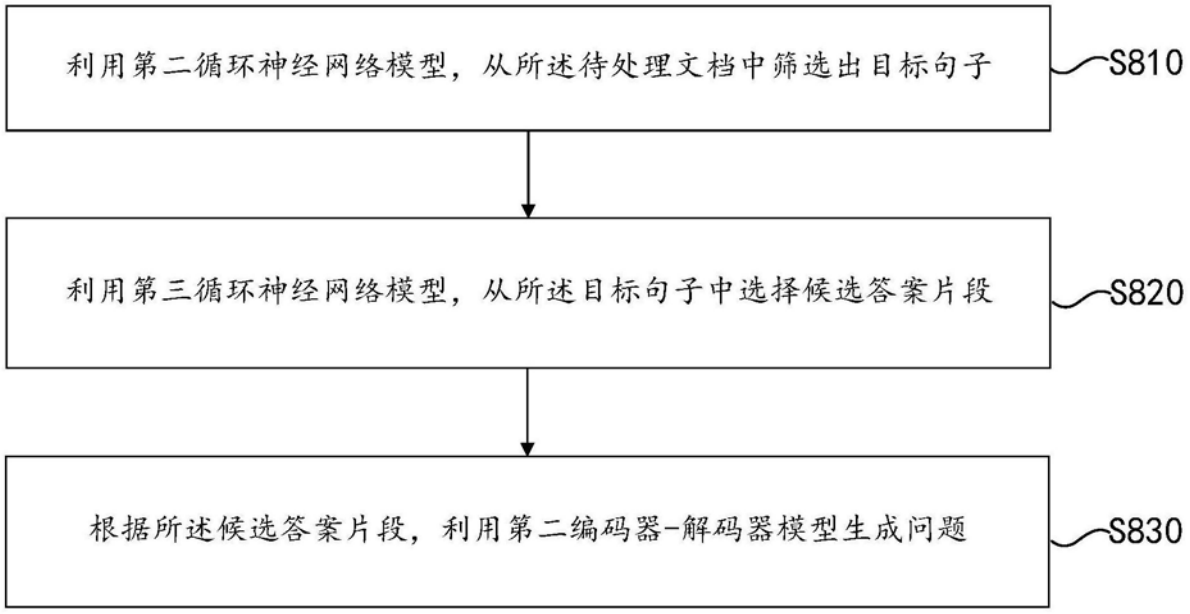

利用第二循环神经网络模型，从所述待处理文档中筛选出目标句子
S810
利用第三循环神经网络模型，从所述目标句子中选择候选答案片段
S820
根据所述候选答案片段，利用第二编码器-解码器模型生成问题
S830

图10

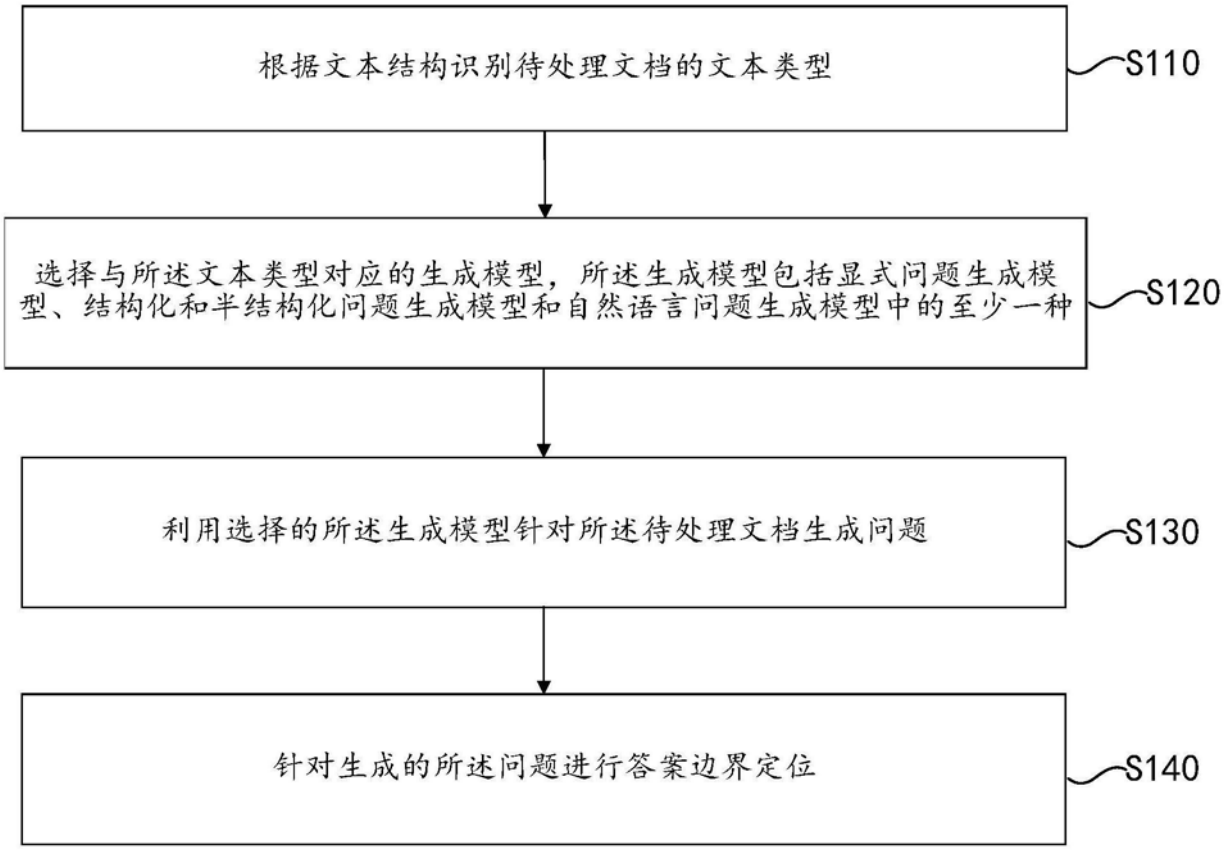

根据文本结构识别待处理文档的文本类型
S110
选择与所述文本类型对应的生成模型，所述生成模型包括显式问题生成模型、结构化和半结构化问题生成模型和自然语言问题生成模型中的至少一种
S120
利用选择的所述生成模型针对所述待处理文档生成问题
S130
针对生成的所述问题进行答案边界定位
S140

图11

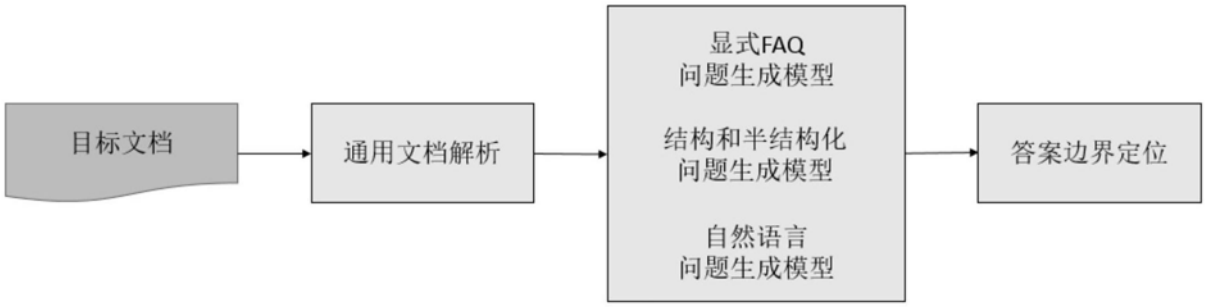

目标文档
通用文档解析
显式FAQ
问题生成模型
结构和半结构化
问题生成模型
自然语言
问题生成模型
答案边界定位

图12

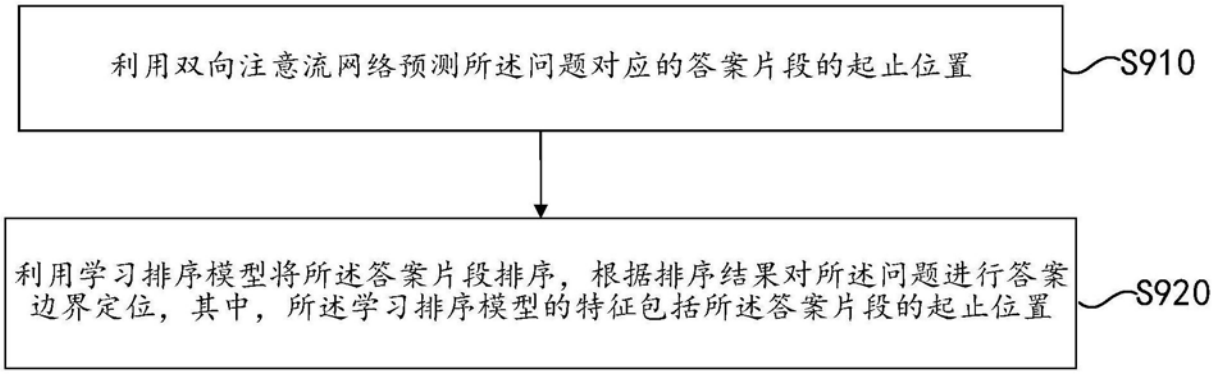

利用双向注意流网络预测所述问题对应的答案片段的起止位置
S910
利用学习排序模型将所述答案片段排序，根据排序结果对所述问题进行答案边界定位，其中，所述学习排序模型的特征包括所述答案片段的起止位置
S920

图13

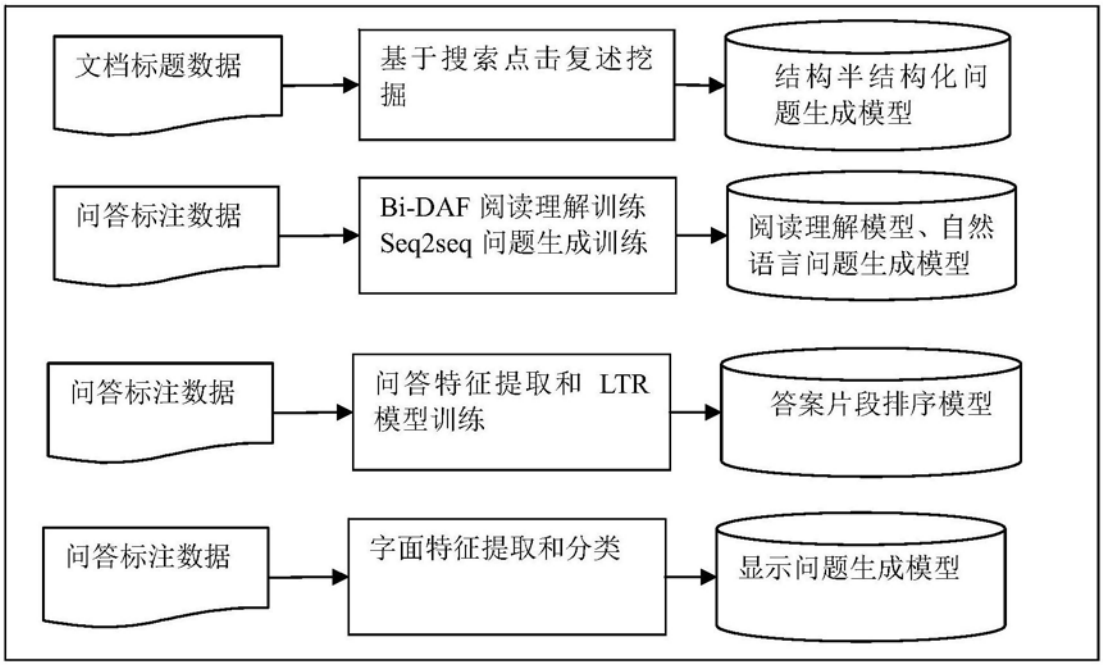

文档标题数据
基于搜索点击复述挖掘
结构半结构化问题生成模型
问答标注数据
Bi-DAF 阅读理解训练
Seq2seq 问题生成训练
阅读理解模型、自然语言问题生成模型
问答标注数据
问答特征提取和 LTR 模型训练
答案片段排序模型
问答标注数据
字面特征提取和分类
显示问题生成模型

图14

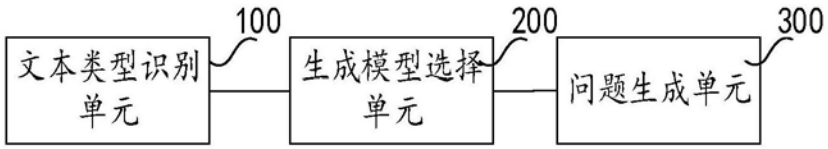


图15

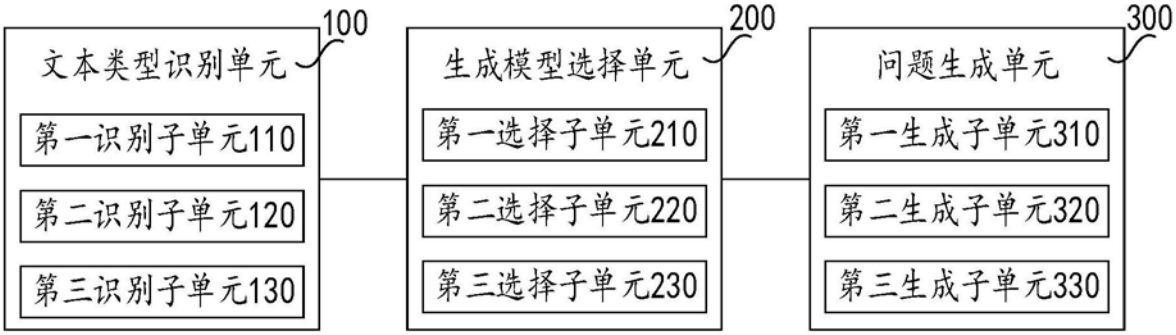


图16

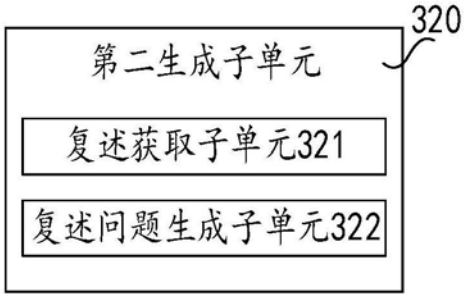


图17

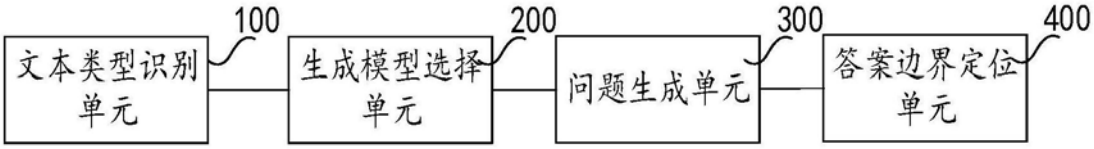


图18

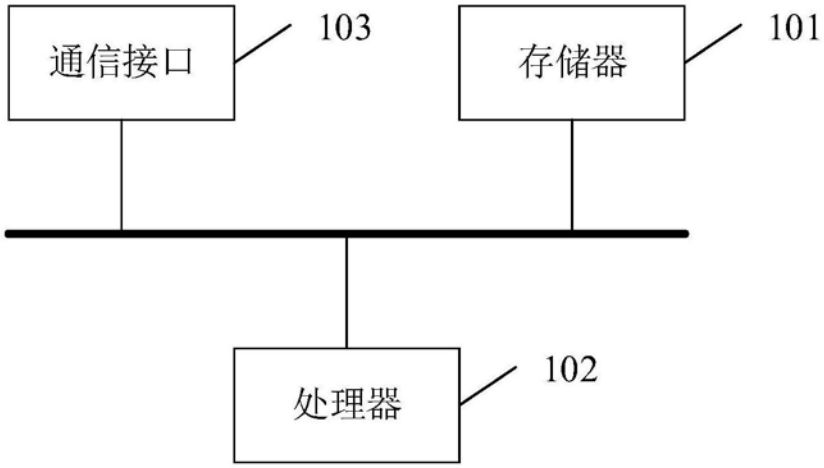


图19