(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111949791 B
(45)授权公告日 2024.01.30

(21)申请号 202010735443.0

(22)申请日 2020.07.28

(65)同一申请的已公布的文献号
申请公布号 CN 111949791 A

(43)申请公布日 2020.11.17

(73)专利权人 中国工商银行股份有限公司
地址 100140 北京市西城区复兴门内大街55号

(72)发明人 孔繁爽 李琦 梁莉娜 王小红

(74)专利代理机构 北京三友知识产权代理有限公司 11127
专利代理师 周达 刘飞

(51)Int.Cl.
G06F 16/35(2019.01)

(56)对比文件
CN 108959246 A,2018.12.07
CN 109670029 A,2019.04.23
CN 111241244 A,2020.06.05
CN 111382232 A,2020.07.07

审查员 刘晶

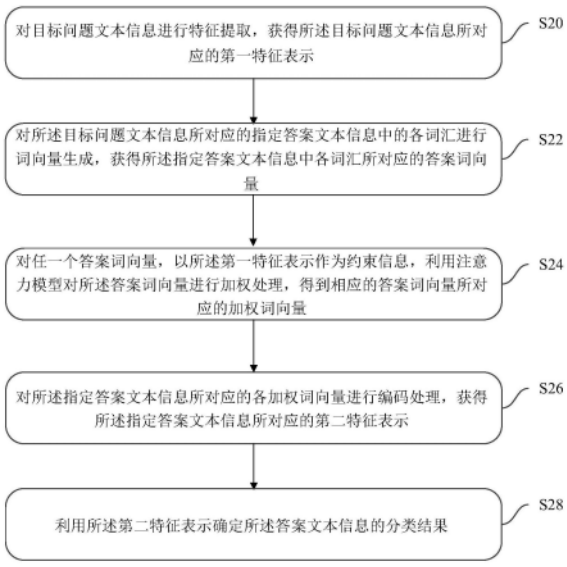
权利要求书2页 说明书10页 附图2页

(54)发明名称

一种文本分类方法、装置及设备

(57)摘要

本说明书实施例涉及人工智能数据处理技术领域,具有公开了一种文本分类方法、装置及设备,所述方法包括对目标问题文本信息进行特征提取,获得目标问题文本信息的第一特征表示;对目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成,获得指定答案文本信息中各词汇所对应的答案词向量;对任一个答案词向量,以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量;对指定答案文本信息所对应的各加权词向量进行编码处理,获得指定答案文本信息所对应的第二特征表示;利用第二特征表示确定所述答案文本信息的分类结果。从而可以进一步提高有效答案筛选的准确性。



CN 111949791 B

1.一种文本分类方法,其特征在于,应用于服务器,所述方法包括:

对目标问题文本信息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示;

对所述目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成,获得所述指定答案文本信息中各词汇所对应的答案词向量;

对任一个答案词向量,以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量;

对所述指定答案文本信息所对应的各加权词向量进行编码处理,获得所述指定答案文本信息所对应的第二特征表示;

利用所述第二特征表示确定所述答案文本信息的分类结果;

其中,所述利用所述第二特征表示确定所述指定答案文本信息的分类结果,包括:

将所述第二特征表示输入预先构建的分类模型中,获得所述指定答案文本信息的分类结果,所述分类模型采用分类算法构建;其中,所述分类结果包括有效答案和无效答案,或者所述分类结果包括有效答案的概率值和无效答案的概率值。

2.根据权利要求1所述的方法,其特征在于,所述对目标问题文本信息进行特征提取,包括:

对所述目标问题文本信息中各词汇进行词向量生成,获得所述目标问题文本信息中各词汇所对应的问题词向量;

对所述目标问题文本信息所对应的各问题词向量进行编码处理,获得所述目标问题文本信息所对应的第一特征表示。

3.根据权利要求1所述的方法,其特征在于,所述以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,包括:

以所述第一特征表示作为注意力模型的约束信息,以所述答案词向量作为注意力模型的值,输入注意力模型,得到所述答案词向量相对所述第一特征表示的相关系数;

计算答案词向量与相应的答案词向量所对应的相关系数的乘积,获得相应的答案词向量所对应的加权词向量。

4.根据权利要求1所述的方法,其特征在于,所述对所述指定答案文本信息所对应的各加权词向量进行编码处理,包括:

利用LSTM算法对对所述指定答案文本信息所对应的各加权词向量进行编码处理。

5.根据权利要求2所述的方法,其特征在于,所述对所述目标问题文本信息中各词汇进行词向量生成之前,还包括:

对所述目标问题文本信息以及所述目标问题文本信息所对应的指定答案文本信息进行分词处理,获得所述指定答案文本信息以及指定答案文本信息所对应的一个或者多个词汇。

6.一种文本分类装置,其特征在于,应用于服务器,所述装置包括:

特征提取模块,用于对目标问题文本信息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示;

词向量生成模块,用于对所述目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成,获得所述指定答案文本信息中各词汇所对应的答案词向量;

加权处理模块,用于对任一个答案词向量,以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量;

编码处理模块,用于对所述指定答案文本信息所对应的各加权词向量进行编码处理,获得所述指定答案文本信息所对应的第二特征表示;

分类模块,用于利用所述第二特征表示确定所述指定答案文本信息的分类结果;

其中,所述分类模块具体用于:将所述第二特征表示输入预先构建的分类模型中,获得所述指定答案文本信息的分类结果,所述分类模型采用分类算法构建;其中,所述分类结果包括有效答案和无效答案,或者所述分类结果包括有效答案的概率值和无效答案的概率值。

7.根据权利要求6所述的装置,其特征在于,所述特征提取模块包括:

词向量生成单元,用于对所述目标问题文本信息中各词汇进行词向量生成,获得所述目标问题文本信息中各词汇所对应的问题词向量;

编码处理单元,用于对所述目标问题文本信息所对应的各问题词向量进行编码处理,获得所述目标问题文本信息所对应的第一特征表示。

8.根据权利要求6所述的装置,其特征在于,所述加权处理模块用于以所述第一特征表示作为注意力模型的约束信息,以所述答案词向量作为注意力模型的值,输入注意力模型,得到所述答案词向量相对所述第一特征表示的相关系数;计算答案词向量与相应的答案词向量所对应的相关系数的乘积,获得相应的答案词向量所对应的加权词向量。

9.一种文本分类设备,其特征在于,应用于服务器,所述设备包括至少一个处理器及用于存储处理器可执行指令的存储器,所述指令被所述处理器执行时实现包括上述权利要求1-5任一项所述方法的步骤。

一种文本分类方法、装置及设备

技术领域

[0001] 本说明书涉及人工智能数据处理技术领域,特别地,涉及一种文本分类方法、装置及设备。

背景技术

[0002] 在软件发布测评或者业务问答等应用场景中,平台可以预先配置系列问题,相应的,用户可以针对问题进行回答。或者,用户也可以在平台发起提问,其他用户或者平台业务人员可以对该问题进行回答。平台可以通过分析不同问题所对应的答案,来获得用户对某项业务或者软件应用的反馈信息。对于某一个问题,可能对应有多个答案,而有些答案可能存在答非所问或者参考意义不大的情况。平台通常需要先对问题所对应的答案进行审核,筛选出较为有效的答案,以更为准确快速的了解用户的反馈。

[0003] 目前,通常采用直接分析答案或者将答案和问题拼接在一起的方式,来确定各答案的有效性。但实际应用中,答案通常是与问题相对应的,仅对答案分析,较难评估各答案的有效性。而将答案和问题拼接在一起进行答案有效性分析,虽然可以将答案和问题进行关联,但鉴于参考答案库的有限性以及用户回答表述形式的复杂多变性,实际处理时较易将实际有效但表述形式与参考答案表述形式差异性较大的答案排除,从而影响有效答案确定的准确性。而考虑上下文语义信息的深度学习算法,因其要求所涉及的上下文信息自身关联性较强,多应用于对话生成领域,较难直接迁移至问答应用场景下使用。因此,本技术领域亟需一种能够更为准确高效的问答类文本分类方法。

发明内容

[0004] 本说明书实施例的目的在于提供一种文本分类方法、装置及设备,可以进一步提高有效答案筛选的准确性。

[0005] 本说明书提供一种文本分类方法、装置及设备是包括如下方式实现的:

[0006] 一种文本分类方法,应用于服务器,所述方法包括:

[0007] 对目标问题文本信息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示;

[0008] 对所述目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成,获得所述指定答案文本信息中各词汇所对应的答案词向量;

[0009] 对任一个答案词向量,以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量;

[0010] 对所述指定答案文本信息所对应的各加权词向量进行编码处理,获得所述指定答案文本信息所对应的第二特征表示;

[0011] 利用所述第二特征表示确定所述答案文本信息的分类结果。

[0012] 本说明书提供的所述方法的另一些实施例中,所述对目标问题文本信息进行特征提取,包括:

[0013] 对所述目标问题文本信息中各词汇进行词向量生成,获得所述目标问题文本信息中各词汇所对应的问题词向量;

[0014] 对所述目标问题文本信息所对应的各问题词向量进行编码处理,获得所述目标问题文本信息所对应的第一特征表示。

[0015] 本说明书提供的所述方法的另一些实施例中,

[0016] 所述以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,包括:

[0017] 以所述第一特征表示作为注意力模型的约束信息,以所述答案词向量作为注意力模型的值,输入注意力模型,得到所述答案词向量相对所述第一特征表示的相关系数;

[0018] 计算答案词向量与相应的答案词向量所对应的相关系数的乘积,获得相应的答案词向量所对应的加权词向量。

[0019] 本说明书提供的所述方法的另一些实施例中,所述利用所述第二特征表示确定所述指定答案文本信息的分类结果,包括:

[0020] 将所述第二特征表示输入预先构建的分类模型中,获得所述指定答案文本信息的分类结果,所述分类模型采用分类算法构建。

[0021] 本说明书提供的所述方法的另一些实施例中,所述对所述指定答案文本信息所对应的各加权词向量进行编码处理,包括:

[0022] 利用LSTM算法对对所述指定答案文本信息所对应的各加权词向量进行编码处理。

[0023] 本说明书提供的所述方法的另一些实施例中,所述对所述目标问题文本信息中各词汇进行词向量生成之前,还包括:

[0024] 对所述目标问题文本信息以及所述目标问题文本信息所对应的指定答案文本信息进行分词处理,获得所述指定答案文本信息以及指定答案文本信息所对应的一个或者多个词汇。

[0025] 另一方面,本说明书实施例还提供一种文本分类装置,应用于服务器,所述装置包括:

[0026] 特征提取模块,用于对目标问题文本信息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示;

[0027] 词向量生成模块,用于对所述目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成,获得所述指定答案文本信息中各词汇所对应的答案词向量;

[0028] 相关性分析模块,用于对任一个答案词向量,以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量;

[0029] 编码处理模块,用于对所述指定答案文本信息所对应的各加权词向量进行编码处理,获得所述指定答案文本信息所对应的第二特征表示;

[0030] 分类模块,用于利用所述第二特征表示确定所述指定答案文本信息的分类结果。

[0031] 本说明书提供的所述装置的另一些实施例中,所述特征提取模块包括:

[0032] 词向量生成单元,用于对所述目标问题文本信息中各词汇进行词向量生成,获得所述目标问题文本信息中各词汇所对应的问题词向量;

[0033] 编码处理单元,用于对所述目标问题文本信息所对应的各问题词向量进行编码处

理,获得所述目标问题文本信息所对应的第一特征表示。

[0034] 本说明书提供的所述装置的另一些实施例中,所述加权处理模块用于以所述第一特征表示作为注意力模型的约束信息,以所述答案词向量作为注意力模型的值,输入注意力模型,得到所述答案词向量相对所述第一特征表示的相关系数;计算答案词向量与相应的答案词向量所对应的相关系数的乘积,获得相应的答案词向量所对应的加权词向量。

[0035] 另一方面,本说明书实施例还提供一种文本分类设备,应用于服务器,所述设备包括至少一个处理器及用于存储处理器可执行指令的存储器,所述指令被所述处理器执行时实现包括上述任意一个或者多个所述方法的步骤。

[0036] 本说明书一个或多个实施例提供的文本分类方法、装置及设备,可以获取表征问题文本信息的语义信息的第一特征表示,以及该问题文本信息所对应的任一答案文本信息中各词汇所对应的答案词向量。然后,可以利用注意力机制,以所述第一特征表示作为注意力模型的约束信息,分别以各答案词向量作为注意力模型的值,分析各答案词向量相对第一特征表示的相关性,并将该相关性作用于相应的答案词向量,获得各答案词向量所对应的加权词向量。相应的,该加权词向量融合了问题文本信息的语义信息。然后,可以对各加权词向量进行编码处理,获得所述答案文本信息所对应的第二特征表示,以利用该第二特征表示进行答案的分类处理。从而,利用本说明书各个实施例,可以有效考虑问题与答案之间的逻辑关系以及问题对答案的语义影响,得到融合了问题语义信息的答案表示。之后,再利用该融合了问题语义信息的第二特征表示对所述答案文本信息进行分类,可以进一步提高有效答案筛选的准确性。

附图说明

[0037] 为了更清楚地说明本说明书实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本说明书中记载的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。在附图中:

[0038] 图1为本说明书提供的一种文本分类方法实施例的流程示意图;

[0039] 图2为本说明书提供的一个实施例中的文本分类流程示意图;

[0040] 图3为本说明书提供的另一种文本分类装置的模块结构示意图。

具体实施方式

[0041] 为了使本技术领域的人员更好地理解本说明书中的技术方案,下面将结合本说明书一个或多个实施例中的附图,对本说明书一个或多个实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅是说明书一部分实施例,而不是全部的实施例。基于说明书一个或多个实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都应当属于本说明书实施例方案保护的范围。

[0042] 本说明书实施例提供的一个场景示例中,所述文本分类方法可以应用于软件发布测评或者业务问答等应用场景下执行问答类文本分类的服务器。所述服务器可以是指一个服务器或者多个服务器组成的服务器集群。

[0043] 对于某一目标问题及其对应的一个或者多个答案,服务器可以对目标问题文本信

息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示。然后,可以对所述目标问题文本信息所对应的指定答案文本信息进行词向量生成,获得所述指定答案文本信息中各词汇所对应的词向量。可以获取表征问题文本信息的语义信息的第一特征表示,以及该问题文本信息所对应的任一答案文本信息中各词汇所对应的答案词向量。然后,可以利用注意力机制,以所述第一特征表示作为注意力模型的约束信息,分别以各答案词向量作为注意力模型的值,分析各答案词向量相对第一特征表示的相关性,并将该相关性作用于相应的答案词向量,获得各答案词向量所对应的加权词向量。相应的,该加权词向量融合了问题文本信息的语义信息。然后,可以对各加权词向量进行编码处理,获得所述答案文本信息所对应的第二特征表示,以利用该第二特征表示进行答案的分类处理。

[0044] 从而,利用本说明书各个实施例,可以有效考虑问题与答案之间的逻辑关系以及问题对答案的语义影响,得到融合了问题语义信息的答案表示。之后,再利用该融合了问题语义信息的第二特征表示对所述答案文本信息进行分类,可以进一步提高有效答案筛选的准确性。

[0045] 图1是本说明书提供的所述文本分类方法实施例流程示意图。虽然本说明书提供了如下述实施例或附图所示的方法操作步骤或装置结构,但基于常规或者无需创造性的劳动在所述方法或装置中可以包括更多或者部分合并后更少的操作步骤或模块单元。在逻辑性上不存在必要因果关系的步骤或结构中,这些步骤的执行顺序或装置的模块结构不限于本说明书实施例或附图所示的执行顺序或模块结构。所述的方法或模块结构的在实际中的装置、服务器或终端产品应用时,可以按照实施例或者附图所示的方法或模块结构进行顺序执行或者并行执行(例如并行处理器或者多线程处理的环境、甚至包括分布式处理、服务器集群的实施环境)。具体的一个实施例如图1所示,本说明书提供的文本分类方法的一个实施例中,所述方法可以应用于所述服务器,所述方法可以包括如下步骤:

[0046] S20:对目标问题文本信息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示。

[0047] 对于某一目标问题及其对应的一个或者多个答案,服务器可以获取该目标问题所对应的目标问题文本信息。例如,所述目标问题文本信息可以为“X信用卡XXX出现问题,请问应该怎样处理”。

[0048] 然后,服务器可以对目标问题文本信息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示。服务器可以将目标问题文本信息映射到一个数值化的语义空间中,获得目标问题文本信息在该数值化的语义空间中的表示信息,以便于对目标问题文本信息进行计算机处理。相应的,所述目标问题文本信息所对应的第一特征表示为携带有目标问题文本信息的语义特征的数值化表示形式。

[0049] 一些实施例中,可以对所述目标问题文本信息中各词汇进行词向量生成,获得所述目标问题文本信息中各词汇所对应的问题词向量。然后,可以将所述多个问题词向量进行编码处理,获得所述目标问题文本信息所对应的第一特征表示。

[0050] 一些实施例中,服务器可以先对目标问题文本信息进行分词处理,获得目标问题文本信息所对应的一个或者多个词汇。如可以利用jieba分词、SnowNLP、THULAC、NLPIR等进行分词处理。另一些实施方式中,还可以同时进行去除停用词等处理,降低干扰信息。

[0051] 对于划分得到的各个词汇,服务器可以进一步生成各词汇所对应的词向量。所述

词向量为各个词汇所对应的数值化表示。即将各词汇映射到数值化的语义空间中,以便于进行计算机处理。可以利用统计的方法或者语言模型的方法生成各词汇所对应的词向量。如可以利用Skip-gram、CBOW、LBL、NNLM、C&W、GloVe等生成目标问题文本信息中的各词汇所对应的词向量。例如,可以利用300维的Glove生成各词汇所对应的词向量。为了便于表述,可以将目标问题文本信息所对应的词向量描述为问题词向量,将目标问题文本信息所对应答案文本信息的词向量描述为答案词向量。

[0052] 然后,服务器可以将所述多个问题词向量进行编码处理,获得所述目标问题文本信息所对应的第一特征表示。例如,可以将提取的词向量输入到序列编码器LSTM(Long Short-Term Memory,长短期记忆网络)中,进行语义压缩处理,取LSTM最后一层隐层的输出,作为目标问题文本信息的所对应的第一特征表示。利用LSTM对词向量进行编码时,在每一时刻,其输出的编码向量不仅依赖于当前时刻的输入,还考虑了上一时刻模型的状态,通过该历史依赖关系,可以使得编码处理后获得的第一特征表示能够更为有效的表征目标问题文本信息各词汇的上下文依存信息,从而更为有效的表征目标问题文本信息所表达的语义信息。当然,实际应用中也可以采用其他的算法进行编码处理,如还可以采用RNN(Recurrent Neural Network,循环神经网络)等。

[0053] 上述实施例中,通过先进行词汇划分,然后,基于各词汇所对应的词向量确定目标问题文本信息所对应的特征表示,可以更为简单方便的确定答案文本信息中各词汇相对目标问题文本信息的重要程度。

[0054] S22:对所述目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成,获得所述指定答案文本信息中各词汇所对应的答案词向量。

[0055] 对于某一目标问题对应的一个或者多个答案中的任意一个答案,服务器可以获取该答案的答案文本信息。所述答案文本信息如可以为“学习了”、“是XX这个意思吗”以及“应该对X信用卡进行XXX处理”等。其中,“学习了”、“是XX这个意思吗”属于没有解决问题的答案以及答非所问的答案。而“应该对X信用卡进行XXX处理”属于对该问题的有效答案。相应的,所述指定答案文本信息可以为目标问题对应的一个或者多个答案中的任意一个待确定类别的答案的文本信息。

[0056] 然后,服务器可以生成指定答案文本信息中的各词汇所对应的词向量,获得指定答案文本信息各词汇所对应的问题词向量。指定答案文本信息中各词汇所对应的词向量的生成方法可以参考目标问题文本信息中词汇的词向量生成方法实施。一些实施方式中,可以设置指定答案文本信息的词向量表示空间与目标问题文本信息中的词向量表示空间一致,也采用300维的Glove词向量进行提取。通过设置二者表示空间维度相同,可以更便于数据处理。

[0057] S24:对任一个答案词向量,以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量。

[0058] 对任一个答案词向量,服务器可以以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量。所述注意力模型可以是指利用注意力(Attention)机制构建的模型。

[0059] 一些实施例中,对任一个答案词向量,服务器可以以所述第一特征表示作为注意力模型的约束信息,以所述答案词向量作为注意力模型的值,输入注意力模型,得到所述答

案词向量相对所述第一特征表示的相关系数。然后,可以计算答案词向量与相应的答案词向量所对应的相关系数的乘积,获得相应的答案词向量所对应的加权词向量。或者,也可以进一步在计算二者乘积之后,进行对乘积值进行归一化等处理,获得相应答案词向量所对应的加权词向量。注意力模型中的相关系数计算过程可以参考注意力机制处理方法进行,这里不做赘述。

[0060] 所述相关系数表征了答案文本信息中各词汇相对目标问题文本信息的相关性。相关系数越大,则答案中的词汇与目标问题的语义信息关联性越强,而相关系数越小,则答案中的词汇与目标问题的语义信息关联性越弱。通过计算二者的相关性,在利用答案的各词向量进行语义编码处理时,以相关系数作为相应词向量的权重,可以进一步突出与目标问题的语义信息关联性较强的词汇的特征表示,弱化与目标问题的语义信息关联性较弱的词汇的特征表示,使得答案的特征表示有效融合目标问题的语义信息,即,使得答案的特征表示可以有效考虑目标问题与指定答案之间的逻辑关系以及目标问题对指定答案的语义影响。

[0061] 之后,再利用融合了目标问题的语义信息的答案特征表示进行答案的有效性分类,可以实现同样的答案在不同问题下分类结果不同,准确剔除无效答案的同时,尽可能准确的保留与问题关联性较强的有效答案。

[0062] S26:对所述答案文本信息所对应的各加权词向量进行编码处理,获得所述指定答案文本信息所对应的第二特征表示。

[0063] 服务器可以将转换后的加权词向量输入进行编码处理,得到所述指定答案文本信息所对应的第二特征表示。例如,可以参考目标问题文本信息,利用LSTM或者RNN等对加权词向量进行编码处理。

[0064] S28:利用所述第二特征表示确定所述指定答案文本信息的分类结果。

[0065] 服务器可以利用所述第二特征表示确定所述指定答案文本信息的分类结果。所述分类结果可以包括有效答案以及无效答案。或者也可以为有效答案、无效答案的概率值等。所述有效答案表示需要筛选出的答案文本信息所对应的类别,所述无效答案表示不需要筛选出的答案文本信息所对应的类别。当然,实际应用场景中,也可以存在其他的分类,这里不做限定。

[0066] 例如,服务器可以将该指定答案文本信息所对应的第二特征表示与参考答案库中给的参考答案的特征表示进行比对,确定该指定答案文本信息属于有效答案的概率。其中,所述参考答案库中的参考答案为预先配置的各问题的有效答案。例如,可以预先根据各问题的实际应用场景,由业务人员预先配置有效答案作为参考答案。之后,还可以在实际应用中,将通过上述实施例提供的方案确定的有效答案动态更新值参考答案库中,以更新优化参考答案库,丰富参考答案库,提高答案分类的准确性。

[0067] 一些实施例中,服务器还可以将所述答案文本信息所对应的第二特征表示输入预先构建的分类模型中,进行分类概率计算,获得所述答案文本信息属于某类别的概率值。所述分类模型如可以采用如多层感知机等分类算法进行构建。通过构建分类模型的方式进行答案分类处理,可以进一步提高分类处理的效率。

[0068] 图2表示文本分类处理方法流程图。如图2所示,本说明书的一个实施场景示例中,可以利用下述步骤构建问答文本分类处理模型,以进行问答文本分类处理。可以先对样本

数据进行预处理。如可以先将样本数据集分别分割成训练集与测试集。对训练集、测试集中的问题及相应的回答句子,可以使用jieba分词进行分词操作,同时去除停用词。

[0069] 假设,对于样本数据中任意问题文本信息及相应的答案文本信息,通过分词处理后,得到该问题文本信息所对应的问题词汇$w_1,w_2,w_2...w_n$以及该问题文本信息所对应的答案词汇$p_1,p_2,p_3...p_m$。

[0070] 然后,可以生成问题词汇$w_1,w_2,w_2...w_n$对应的问题词向量,即图2中左侧Embeding的输出,其中,词向量可以采用300维Glove词向量。可以将问题词向量输入到序列编码器LSTM中,进行问题文本信息的语义压缩,取LSTM最后一层隐层的输出,得到问题文本信息的第一特征表示向量q。

[0071] 然后,可以生成上述问题文本信息对应的答案文本信息中各词汇$p_1,p_2,p_3...p_m$所对应的答案词向量,即图2中右侧Embeding的输出。该答案词向量表示空间与问题词向量表示空间一致,同为300维Glove词向量。

[0072] 以第一特征表示向量q作为注意力机制的信号,答案文本信息对应的答案词向量分别作为注意力机制的值,首先计算答案词向量相对第一特征表示向量q的相关系数。其中,Softmax的输出即为各答案词向量的相关系数。然后,计算相关系数与其对应的答案词向量的乘积,获得该答案词向量所对应的加权词向量。各答案词向量所对应的加权词向量即图2中Att.(Attention)的输出。

[0073] 将各加权词向量输入到解码器LSTM中,进行语义编码,得到答案文本信息所对应的第二特征表示向量e。

[0074] 可以以第二特征表示向量e作为分类算法的输入,以其所对应的答案文本信息的分类标签作为输出,进行分类算法的训练。例如,可以采用交叉熵损失计算方式对算法分类结果与真实标签比较计算损失。以及采用minibatch训练方法,得到损失后,使用SGD优化器进行模型梯度更新,重复上述步骤,直至连续10个epoch训练损失不再下降,得到最终的模型和参数,从而获得训练好的问答文本分类处理模型。

[0075] 对于某目标问题,可以将目标问题文本信息以及对应的待分类答案文本信息输入上述训练好的问答文本分类处理模型中,在目标问题文本信息所对应的第一特征表示的约束下,进行第二特征表示的构建以及基于构建的第二特征表示进行答案分类结果的确定,获得该待分类答案文本信息的分类结果。

[0076] 上述一个或者多个实施例提供的方案,将问题与答案分别作为注意力机制的信号和值,通过注意力机制的映射,将问题与答案的语义空间统一,得到融合问题语义信息的答案句子表征,有效将问题信息融合到答案的新表征空间,为问题答案参与其他任务训练提供了初始语义表示,从而充分考虑了问题、答案的上下文逻辑关系,能够实现同样回答在不同问题下分类结果不同的效果,进而提高了答案分类的准确性。相比于传统文本二分类方法,上述实施例提供的方案分类特征更丰富,适应性更好。同时,使用注意力机制将问题信息融合到答案表示的框架,可在各编码阶段根据语料特征使用任意深度学习表示方法,如CNN,RNN,GRU,LSTM等,使用更为灵活。

[0077] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。具体的可以参照前述相关处理相关实施例的描述,在此不做一一赘述。

[0078] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下,在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是可以的或者可能是有利的。

[0079] 本说明书一个或多个实施例提供的文本分类方法,可以获取表征问题文本信息的语义信息的第一特征表示,以及该问题文本信息所对应的任一答案文本信息中各词汇所对应的答案词向量。然后,可以利用注意力机制,以所述第一特征表示作为注意力模型的约束信息,分别以各答案词向量作为注意力模型的值,分析各答案词向量相对第一特征表示的相关性,并将该相关性作用于相应的答案词向量,获得各答案词向量所对应的加权词向量。相应的,该加权词向量融合了问题文本信息的语义信息。然后,可以对各加权词向量进行编码处理,获得所述答案文本信息所对应的第二特征表示,以利用该第二特征表示进行答案的分类处理。从而,利用本说明书各个实施例,可以有效考虑问题与答案之间的逻辑关系以及问题对答案的语义影响,得到融合了问题语义信息的答案表示。之后,再利用该融合了问题语义信息的第二特征表示对所述答案文本信息进行分类,可以进一步提高有效答案筛选的准确性。

[0080] 基于上述所述的文本分类方法,本说明书一个或多个实施例还提供一种文本分类装置。所述的装置可以包括使用了本说明书实施例所述方法的系统、软件(应用)、模块、组件、服务器等并结合必要的实施硬件的装置。基于同一创新构思,本说明书实施例提供的一个或多个实施例中的装置如下面的实施例所述。由于装置解决问题的实现方案与方法相似,因此本说明书实施例具体的装置的实施可以参见前述方法的实施,重复之处不再赘述。以下所使用的,术语“单元”或者“模块”可以实现预定功能的软件和/或硬件的组合。尽管以下实施例所描述的装置较佳地以软件来实现,但是硬件,或者软件和硬件的组合的实现也是可能并被构想的。具体的,图3表示说明书提供的一种文本分类装置实施例的模块结构示意图,如图3所示,应用于边端服务器,所述装置可以包括:

[0081] 特征提取模块102,可以用于对目标问题文本信息进行特征提取,获得所述目标问题文本信息所对应的第一特征表示。

[0082] 词向量生成模块104,可以用于对所述目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成,获得所述指定答案文本信息中各词汇所对应的答案词向量。

[0083] 加权处理模块106,可以用于对任一个答案词向量,以所述第一特征表示作为约束信息,利用注意力模型对所述答案词向量进行加权处理,得到相应的答案词向量所对应的加权词向量。

[0084] 编码处理模块108,可以用于对所述指定答案文本信息所对应的各加权词向量进行编码处理,获得所述指定答案文本信息所对应的第二特征表示。

[0085] 分类模块110,可以用于利用所述第二特征表示确定所述指定答案文本信息的分类结果。

[0086] 另一些实施例中,所述特征提取模块102可以包括:

[0087] 词向量生成单元,可以用于对所述目标问题文本信息中各词汇进行词向量生成,

获得所述目标问题文本信息中各词汇所对应的问题词向量。

[0088] 编码处理单元,可以用于对所述目标问题文本信息所对应的各问题词向量进行编码处理,获得所述目标问题文本信息所对应的第一特征表示。

[0089] 另一些实施例中,所述加权处理模块106可以用于以所述第一特征表示作为注意力模型的约束信息,以所述答案词向量作为注意力模型的值,输入注意力模型,得到所述答案词向量相对所述第一特征表示的相关系数,然后,可以计算答案词向量与相应的答案词向量所对应的相关系数的乘积,获得相应的答案词向量所对应的加权词向量。

[0090] 需要说明的,上述所述的装置根据方法实施例的描述还可以包括其他的实施方式。具体的实现方式可以参照相关方法实施例的描述,在此不作一一赘述。

[0091] 本说明书一个或多个实施例提供的文本分类装置,可以获取表征问题文本信息的语义信息的第一特征表示,以及该问题文本信息所对应的任一答案文本信息中各词汇所对应的答案词向量。然后,可以利用注意力机制,以所述第一特征表示作为注意力模型的约束信息,分别以各答案词向量作为注意力模型的值,分析各答案词向量相对第一特征表示的相关性,并将该相关性作用于相应的答案词向量,获得各答案词向量所对应的加权词向量。相应的,该加权词向量融合了问题文本信息的语义信息。然后,可以对各加权词向量进行编码处理,获得所述答案文本信息所对应的第二特征表示,以利用该第二特征表示进行答案的分类处理。从而,利用本说明书各个实施例,可以有效考虑问题与答案之间的逻辑关系以及问题对答案的语义影响,得到融合了问题语义信息的答案表示。之后,再利用该融合了问题语义信息的第二特征表示对所述答案文本信息进行分类,可以进一步提高有效答案筛选的准确性。

[0092] 本说明书还提供一种文本分类设备,所述设备可以应用于单独的文本分类系统中,也可以应用在多种计算机数据处理系统中。所述的系统可以为单独的服务器,也可以包括使用了本说明书的一个或多个所述方法或一个或多个实施例装置的服务器集群、系统(包括分布式系统)、软件(应用)、实际操作装置、逻辑门电路装置、量子计算机等并结合必要的实施硬件的终端装置。一些实施例中,设备可以包括至少一个处理器及用于存储处理器可执行指令的存储器,所述指令被所述处理器执行时实现包括上述任意一个或者多个实施例所述方法的步骤。

[0093] 所述存储器可以包括用于存储信息的物理装置,通常是将信息数字化后再以利用电、磁或者光学等方式的媒体加以存储。所述存储介质有可以包括:利用电能方式存储信息的装置如,各式存储器,如RAM、ROM等;利用磁能方式存储信息的装置如,硬盘、软盘、磁带、磁芯存储器、磁泡存储器、U盘;利用光学方式存储信息的装置如,CD或DVD。当然,还有其他方式的可读存储介质,例如量子存储器、石墨烯存储器等等。

[0094] 需要说明的,上述所述的设备根据方法或者装置实施例的描述还可以包括其他的实施方式,具体的实现方式可以参照相关方法实施例的描述,在此不作一一赘述。

[0095] 上述实施例所述的文本分类设备,可以获取表征问题文本信息的语义信息的第一特征表示,以及该问题文本信息所对应的任一答案文本信息中各词汇所对应的答案词向量。然后,可以利用注意力机制,以所述第一特征表示作为注意力模型的约束信息,分别以各答案词向量作为注意力模型的值,分析各答案词向量相对第一特征表示的相关性,并将该相关性作用于相应的答案词向量,获得各答案词向量所对应的加权词向量。相应的,该加

权词向量融合了问题文本信息的语义信息。然后,可以对各加权词向量进行编码处理,获得所述答案文本信息所对应的第二特征表示,以利用该第二特征表示进行答案的分类处理。从而,利用本说明书各个实施例,可以有效考虑问题与答案之间的逻辑关系以及问题对答案的语义影响,得到融合了问题语义信息的答案表示。之后,再利用该融合了问题语义信息的第二特征表示对所述答案文本信息进行分类,可以进一步提高有效答案筛选的准确性。

[0096] 需要说明的是,本说明书实施例并不局限于必须是符合标准数据模型/模板或本说明书实施例所描述的情况。某些行业标准或者使用自定义方式或实施例描述的实施基础上略加修改后的实施方案也可以实现上述实施例相同、等同或相近、或变形后可预料的实施效果。应用这些修改或变形后的数据获取、存储、判断、处理方式等获取的实施例,仍然可以属于本说明书的可选实施方案范围之内。

[0097] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于系统实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。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说明书的至少一个实施例或示例中。在本说明书中,对上述术语的示意性表述并不必须针对的是相同的实施例或示例。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。

[0098] 以上所述仅为本说明书的实施例而已,并不用于限制本说明书。对于本领域技术人员来说,本说明书可以有各种更改和变化。凡在本说明书的精神和原理之内所作的任何修改、等同替换、改进等,均应包含在本说明书的权利要求范围之内。

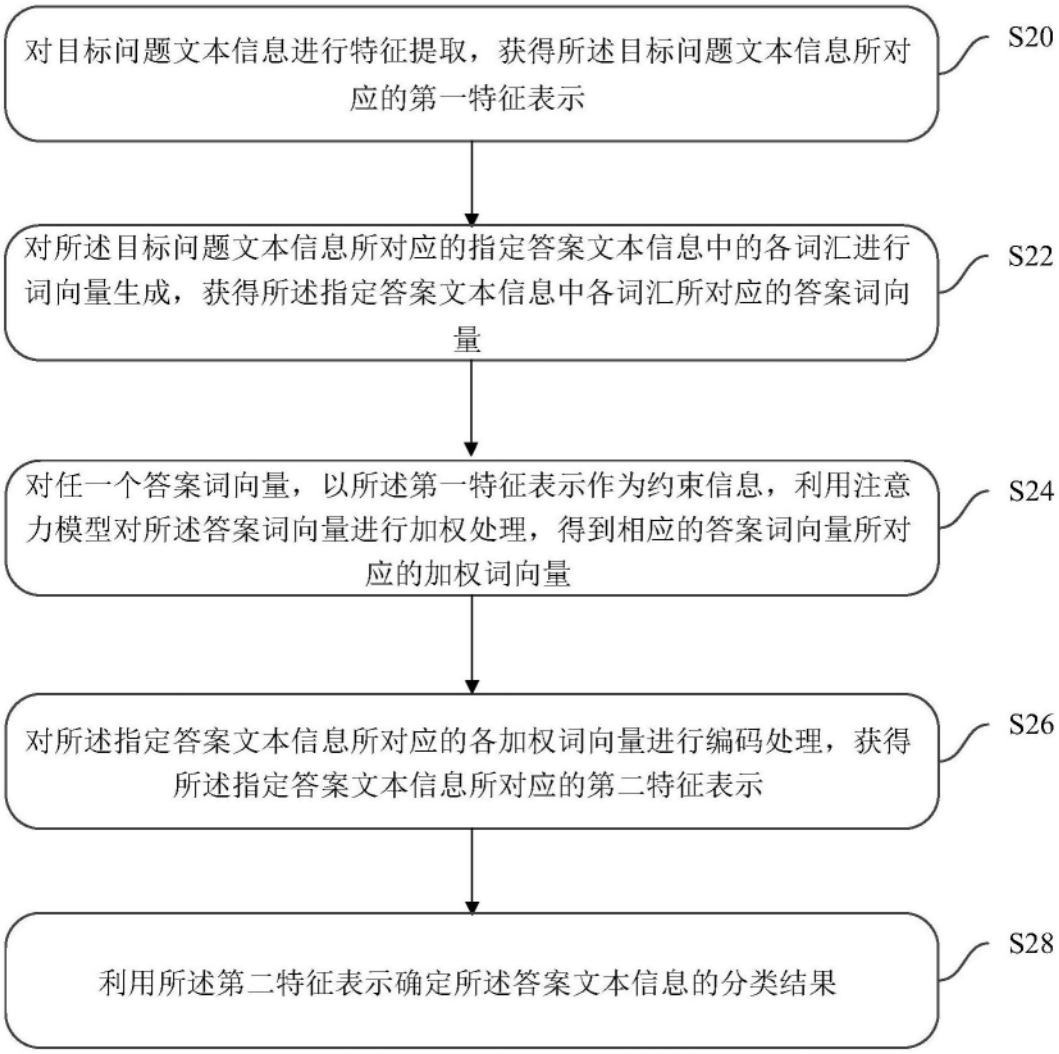

对目标问题文本信息进行特征提取，获得所述目标问题文本信息所对应的第一特征表示
S20
对所述目标问题文本信息所对应的指定答案文本信息中的各词汇进行词向量生成，获得所述指定答案文本信息中各词汇所对应的答案词向量
S22
对任一个答案词向量，以所述第一特征表示作为约束信息，利用注意力模型对所述答案词向量进行加权处理，得到相应的答案词向量所对应的加权词向量
S24
对所述指定答案文本信息所对应的各加权词向量进行编码处理，获得所述指定答案文本信息所对应的第二特征表示
S26
利用所述第二特征表示确定所述答案文本信息的分类结果
S28

图1

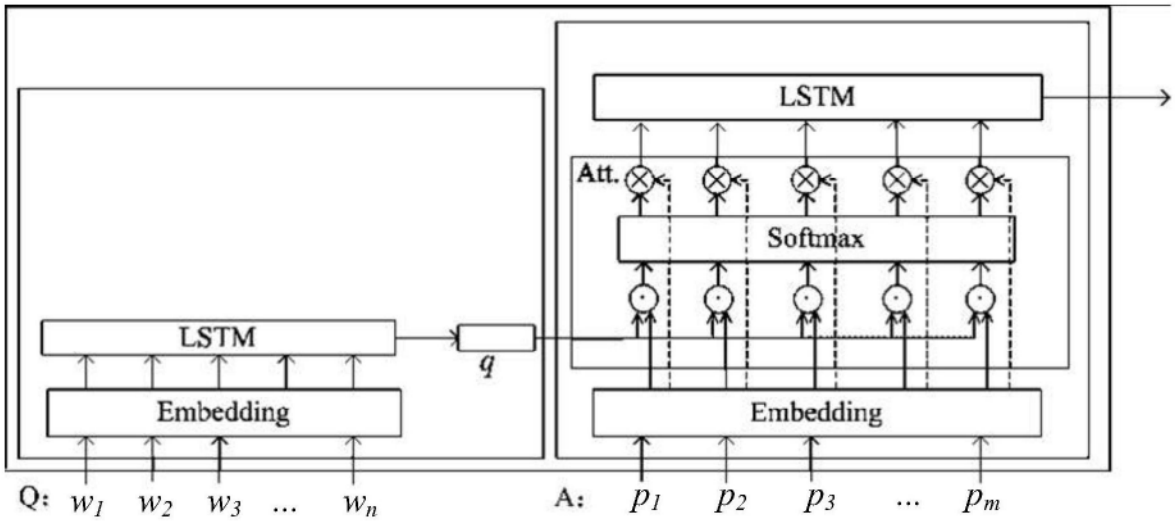

LSTM
Att.
Softmax
LSTM
q
Embedding
Embedding
Q: w_1 w_2 w_3 ... w_n
A: p_1 p_2 p_3 ... p_m

图2

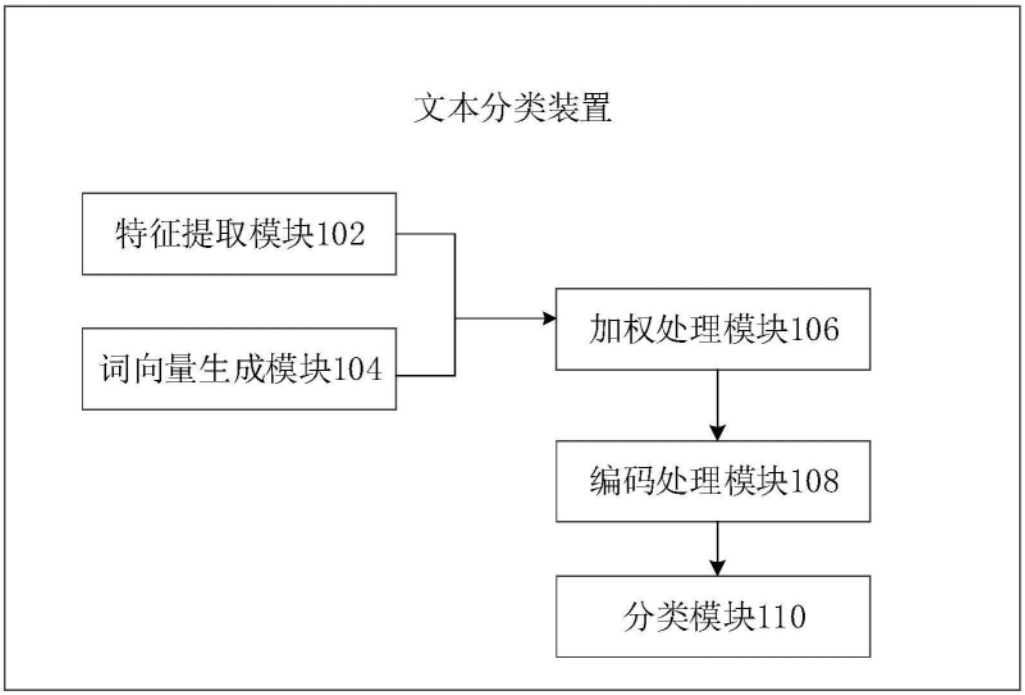

文本分类装置
特征提取模块102
词向量生成模块104
加权处理模块106
编码处理模块108
分类模块110

图3