(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 112329430 B
(45)授权公告日 2021.03.16

(21)申请号 202110000674.1
(22)申请日 2021.01.04
(65)同一申请的已公布的文献号
申请公布号 CN 112329430 A
(43)申请公布日 2021.02.05
(73)专利权人 恒生电子股份有限公司
地址 310053 浙江省杭州市滨江区江南大道3588号恒生大厦11楼
(72)发明人 王炯亮 娄东方 林金曙 高峰 陈哲 许浩
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 张柳
(51)Int.Cl.
G06F 40/194(2020.01)
G06K 9/62(2006.01)
(56)对比文件
CN 107729300 A,2018.02.23
审查员 易铭

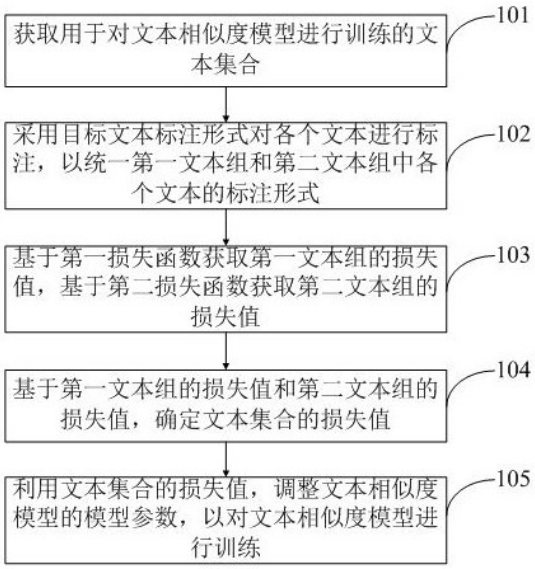
权利要求书6页 说明书14页 附图5页

(54)发明名称

一种模型训练方法、文本相似度确定方法及装置

(57)摘要

本申请提供一种模型训练方法、文本相似度确定方法及装置，获取用于对文本相似度模型进行训练的文本集合，文本集合包括第一文本组和第二文本组，第一文本组中的各个文本以第一文本标注形式标注，第二文本组中的各个文本以第二文本标注形式标注；采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式；基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值；基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值；利用文本集合的损失值，调整文本相似度模型的模型参数，以训练文本相似度模型，使得文本相似度模型的训练方法具备更强的兼容性，并提高模型准确度。



CN 112329430 B

1.一种模型训练方法,其特征在于,所述方法包括:

获取用于对文本相似度模型进行训练的文本集合,所述文本集合包括第一文本组和第二文本组,所述第一文本组中的各个文本以第一文本标注形式标注,所述第二文本组中的各个文本以第二文本标注形式标注,其中,所述第一文本标注形式在标注过程中的持续一致性强于所述第二文本标注形式在标注过程中的持续一致性,但所述第一文本标注形式的监督性能差于所述第二文本标注形式的监督性能;

采用目标文本标注形式对各个文本进行标注,以统一所述第一文本组和所述第二文本组中各个文本的标注形式,其中,所述目标文本标注形式为所述第一文本标注形式;

基于第一损失函数获取所述第一文本组的损失值,基于第二损失函数获取所述第二文本组的损失值;

基于所述第一文本组的损失值和所述第二文本组的损失值,确定所述文本集合的损失值;

利用所述文本集合的损失值,调整所述文本相似度模型的模型参数,以对所述文本相似度模型进行训练。

2.根据权利要求1所述的方法,其特征在于,所述基于第一损失函数获取所述第一文本组的损失值,基于第二损失函数获取所述第二文本组的损失值包括:

获取以目标文本标注形式标注的第一文本组中各个文本的第一文本向量,获取以目标文本标注形式标注的第二文本组中各个文本的第二文本向量;

基于所述第一文本向量,确定所述第一文本组中各个文本之间的相似度,基于所述第二文本向量,确定所述第二文本组中各个文本之间的相似度;

基于所述第一损失函数和所述第一文本向量,获取所述第一文本组的损失值;

基于所述第二损失函数和所述第二文本向量,获取所述第二文本组的损失值。

3.根据权利要求2所述的方法,其特征在于,所述基于所述第一文本向量,确定所述第一文本组中各个文本之间的相似度包括:

对所述第一文本组中的任意两个文本,利用公式 $sim(v_1, v_2)=e^{-\frac{\|v_1-v_2\|_2^2}{\sigma}}$ 得到任意两个文本之间的相似度,$\|\cdot\|_2$ 表示2-范数,$\sigma>0$ 表示2-范数标准化因子,v_1 和 v_2 表示任意两个文本各自的第一文本向量;

所述基于所述第二文本向量,确定所述第二文本组中各个文本之间的相似度包括:

对所述第二文本组中的任意两个文本,利用公式 $sim(y_1, y_2)=e^{-\frac{\|y_1-y_2\|_2^2}{\sigma}}$ 得到任意两个文本之间的相似度,$\|\cdot\|_2$ 表示2-范数,$\sigma>0$ 表示2-范数标准化因子,y_1 和 y_2 表示任意两个文本各自的第二文本向量。

4.根据权利要求1所述的方法,其特征在于,所述第一文本标注形式为三元组标注形

式,所述第二文本标注形式为0-1标注形式;

所述采用目标文本标注形式对各个文本进行标注,以统一所述第一文本组和所述第二文本组中各个文本的标注形式包括:

如果所述第二文本组中两个文本对应0标签,则将所述第二文本组以(A,B,A)形式进行标注,A和B为所述第二文本组中的两个文本;

如果所述第二文本组中两个文本对应1标签,则将所述第二文本组以(A,A,B)形式进行标注,A和B为所述第二文本组中的两个文本。

5.根据权利要求4所述的方法,其特征在于,所述基于第一损失函数获取所述第一文本组的损失值,基于第二损失函数获取所述第二文本组的损失值包括:

对以三元组标注形式标注的第一文本组(A,P,N),利用

$$loss_{triplet}=\frac{1}{n}\sum_{(A,P,N)}max(sim(v_A-v_N)-sim(v_A-v_P)+\delta,0)$$ 获取所述第一文本组的损失值,A,P和N为第一文本组中的三个文本,n为第一文本组的总数,v_A 为第一文本组中文本A的文本向量,v_P 为第一文本组中文本P的文本向量,v_N 第一文本组中文本N的文本向量,$\delta>0$ 表示第一文本组中相似度高的文本对与相似度低的文本对之间的相似度差异阈值;

对具有0标签且以三元组标注形式标注的第二文本组(A,B,A),利用 $$loss_0=\frac{1}{n}\sum_{(A,B,A)}sim(v_A,v_B)$$ 获取所述具有0标签且以三元组标注形式标注的第二文本组的损失值,A,B和A为具有0标签且以三元组标注形式标注的第二文本组中的三个文本,n为具有0标签且以三元组标注形式标注的第二文本组的总数,v_A 为具有0标签且以三元组标注形式标注的第二文本组中文本A的文本向量,v_B 为具有0标签且以三元组标注形式标注的第二文本组中文本B的文本向量;

对具有1标签且以三元组标注形式标注的第二文本组(A,A,B),利用 $$loss_1=\frac{1}{n}\sum_{(A,A,B)}(1-sim(v_A,v_B))$$ 获取所述具有1标签且以三元组标注形式标注的第二文本组的损失值,A,A和B为具有1标签且以三元组标注形式标注的第二文本组中的三个文本,n为具有1标签且以三元组标注形式标注的第二文本组的总数,v_A 为具有1标签且以三元组标注形式标注的第二文本组中文本A的文本向量,v_B 为具有1标签且以三元组标注形式标注的第二文本组中文本B的文本向量。

6.根据权利要求5所述的方法,其特征在于,所述基于所述第一文本组的损失值和所述第二文本组的损失值,确定所述文本集合的损失值包括:

利用 $loss = loss_{triplet} + \lambda_0 \cdot loss_0 + \lambda_1 \cdot loss_1$ 得到文本集合的损失值$loss$，λ_0 为 $loss_0$ 的权重，λ_1 为 $loss_1$ 的权重。

7.根据权利要求1所述的方法，其特征在于，还包括：

获取测试文本组；

利用训练得到的文本相似度模型对所述测试文本组中的文本进行编码，得到所述测试文本组中各文本的文本向量；

基于所述测试文本组中各文本的文本向量，确定所述测试文本组中各文本的相似结果；

如果所述相似结果与所述测试文本组的已知测试结果不一致，则以所述目标文本标注形式对所述测试文本组进行标注；

利用标注后的测试文本组中的各文本，对训练得到的文本相似度模型的模型参数进行调整。

8.一种文本相似度确定方法，其特征在于，所述方法包括：

获取第一待处理文本和第二待处理文本；

调用文本相似度模型，所述文本相似度模型是通过文本集合的损失值调整模型参数得到，所述文本集合的损失值基于文本集合中的第一文本组的损失值和第二文本组的损失值得到，所述第一文本组中的各个文本以第一文本标注形式标注，所述第二文本组中的各个文本以第二文本标注形式标注，在得到所述第一文本组的损失值和所述第二文本组的损失值之前，采用目标文本标注形式对各个文本进行标注，以统一所述第一文本组和所述第二文本组中各个文本的标注形式，其中，所述第一文本标注形式在标注过程中的持续一致性强于所述第二文本标注形式在标注过程中的持续一致性，但所述第一文本标注形式的监督性能差于所述第二文本标注形式的监督性能，所述目标文本标注形式为所述第一文本标注形式；

获得所述文本相似度模型输出的指示所述第一待处理文本和所述第二待处理文本是否相似的处理结果。

9.一种模型训练装置，其特征在于，所述装置包括：

获取单元，用于获取用于对文本相似度模型进行训练的文本集合，所述文本集合包括第一文本组和第二文本组，所述第一文本组中的各个文本以第一文本标注形式标注，所述第二文本组中的各个文本以第二文本标注形式标注，其中，所述第一文本标注形式在标注过程中的持续一致性强于所述第二文本标注形式在标注过程中的持续一致性，但所述第一文本标注形式的监督性能差于所述第二文本标注形式的监督性能；

标注单元，用于采用目标文本标注形式对各个文本进行标注，以统一所述第一文本组和所述第二文本组中各个文本的标注形式，其中，所述目标文本标注形式为所述第一文本标注形式；

损失值确定单元，用于基于第一损失函数获取所述第一文本组的损失值，基于第二损失函数获取所述第二文本组的损失值；基于所述第一文本组的损失值和所述第二文本组的损失值，确定所述文本集合的损失值；

调整单元，用于利用所述文本集合的损失值，调整所述文本相似度模型的模型参数，以对所述文本相似度模型进行训练。

10.根据权利要求9所述的装置，其特征在于，所述损失值确定单元，用于获取以目标文本标注形式标注的第一文本组中各个文本的第一文本向量，获取以目标文本标注形式标注的第二文本组中各个文本的第二文本向量；基于所述第一文本向量，确定所述第一文本组中各个文本之间的相似度，基于所述第二文本向量，确定所述第二文本组中各个文本之间的相似度；基于所述第一损失函数和所述第一文本向量，获取所述第一文本组的损失值；基于所述第二损失函数和所述第二文本向量，获取所述第二文本组的损失值。

11.根据权利要求10所述的装置，其特征在于，所述损失值确定单元确定所述第一文本组中各个文本之间的相似度包括：对所述第一文本组中的任意两个文本，利用公式 $sim(v_1, v_2) = e^{-\frac{\|v_1 - v_2\|_2^2}{\sigma}}$ 得到任意两个文本之间的相似度，$\|\cdot\|_2$ 表示2-范数，$\sigma>0$ 表示2-范数标准化因子，v_1 和 v_2 表示任意两个文本各自的第一文本向量；

所述损失值确定单元确定所述第二文本组中各个文本之间的相似度包括：对所述第二文本组中的任意两个文本，利用公式 $sim(y_1, y_2) = e^{-\frac{\|y_1 - y_2\|_2^2}{\sigma}}$ 得到任意两个文本之间的相似度，$\|\cdot\|_2$ 表示2-范数，$\sigma>0$ 表示2-范数标准化因子，y_1 和 y_2 表示任意两个文本各自的第二文本向量。

12.根据权利要求9所述的装置，其特征在于，所述第一文本标注形式为三元组标注形式，所述第二文本标注形式为0-1标注形式；

所述标注单元，用于如果所述第二文本组中两个文本对应0标签，则将所述第二文本组以(A,B,A)形式进行标注，A和B为所述第二文本组中的两个文本；以及用于如果所述第二文本组中两个文本对应1标签，则将所述第二文本组以(A,A,B)形式进行标注，A和B为所述第二文本组中的两个文本。

13.根据权利要求12所述的装置，其特征在于，所述损失值确定单元，用于：

对以三元组标注形式标注的第一文本组(A,P,N)，利用

$$loss_{triplet} = \frac{1}{n}\sum_{(A,P,N)} max(sim(v_A - v_N) - sim(v_A - v_P) + \delta, 0)$$

获取所述第一文本组的损失值，A,P和N为第一文本组中的三个文本，n为第一文本组的总数，v_A 为第一文本组中文本A的文本向量，v_P 为第一文本组中文本P的文本向量，v_N 第一文本组中文本N的文本向量，$\delta>0$ 表示第一文本组中相似度高的文本对与相似度低的文本对之间的相似度差异阈值；

对具有0标签且以三元组标注形式标注的第二文本组（A，B，A），利用 $loss_0=\frac{1}{n_{(A,B,A)}}\sum sim(v_A,v_B)$ 获取所述具有0标签且以三元组标注形式标注的第二文本组的损失值，A，B和A为具有0标签且以三元组标注形式标注的第二文本组中的三个文本，n为具有0标签且以三元组标注形式标注的第二文本组的总数，v_A 为具有0标签且以三元组标注形式标注的第二文本组中文本A的文本向量，v_B 为具有0标签且以三元组标注形式标注的第二文本组中文本B的文本向量；

对具有1标签且以三元组标注形式标注的第二文本组（A，A，B），利用 $loss_1=\frac{1}{n_{(A,A,B)}}\sum (1-sim(v_A,v_B))$ 获取所述具有1标签且以三元组标注形式标注的第二文本组的损失值，A，A和B为具有1标签且以三元组标注形式标注的第二文本组中的三个文本，n为具有1标签且以三元组标注形式标注的第二文本组的总数，v_A 为具有1标签且以三元组标注形式标注的第二文本组中文本A的文本向量，v_B 为具有1标签且以三元组标注形式标注的第二文本组中文本B的文本向量。

14.根据权利要求13所述的装置，其特征在于，所述损失值确定单元，用于利用 $loss=loss_{triplet}+\lambda_0\cdot loss_0+\lambda_1\cdot loss_1$ 得到文本集合的损失值$loss$，λ_0 为 $loss_0$ 的权重，λ_1 为 $loss_1$ 的权重。

15.根据权利要求9所述的装置，其特征在于，还包括：编码单元和确定单元；

所述获取单元，还用于获取测试文本组；

所述编码单元，用于利用训练得到的文本相似度模型对所述测试文本组中的文本进行编码，得到所述测试文本组中各文本的文本向量；

所述确定单元，用于基于所述测试文本组中各文本的文本向量，确定所述测试文本组中各文本的相似结果；

所述标注单元，还用于如果所述相似结果与所述测试文本组的已知测试结果不一致，则以所述目标文本标注形式对所述测试文本组进行标注；

所述调整单元，还用于利用标注后的测试文本组中的各文本，对训练得到的文本相似度模型的模型参数进行调整。

16.一种文本相似度确定装置，其特征在于，所述装置包括：

文本获取单元，用于获取第一待处理文本和第二待处理文本；

调用单元，用于调用文本相似度模型，所述文本相似度模型是通过文本集合的损失值调整模型参数得到，所述文本集合的损失值基于文本集合中的第一文本组的损失值和第二文本组的损失值得到，所述第一文本组中的各个文本以第一文本标注形式标注，所述第二文本组中的各个文本以第二文本标注形式标注，在得到所述第一文本组的损失值和所述第二文本组的损失值之前，采用目标文本标注形式对各个文本进行标注，以统一所述第一文

本组和所述第二文本组中各个文本的标注形式，其中，所述第一文本标注形式在标注过程中的持续一致性强于所述第二文本标注形式在标注过程中的持续一致性，但所述第一文本标注形式的监督性能差于所述第二文本标注形式的监督性能，所述目标文本标注形式为所述第一文本标注形式；

结果获得单元，用于获得所述文本相似度模型输出的指示所述第一待处理文本和所述第二待处理文本是否相似的处理结果。

17.一种电子设备，其特征在于，包括：

处理器；

存储器，用于存储可执行指令；

其中，所述处理器配置为经由执行所述可执行指令来执行如权利要求1至7中任意一项所述的模型训练方法和/或权利要求8所述的文本相似度确定方法。

18.一种存储介质，其特征在于，所述存储介质中存储有计算机程序代码，所述计算机程序代码被运行时执行如权利要求1至7中任意一项所述的模型训练方法和/或权利要求8所述的文本相似度确定方法。

一种模型训练方法、文本相似度确定方法及装置

技术领域

[0001] 本申请属于人工智能技术领域，尤其涉及一种模型训练方法、文本相似度确定方法及装置。

背景技术

[0002] 随着互联网和人工智能技术的迅速发展，基于自然语言的相似搜索和问答成为各大网站、APP(Application，应用程序)、智能客服系统等的必备技能，在基于自然语言的相似搜索和问答过程中文本相似度模型是关键模型，文本相似度模型用于判别两个文本之间是否相似以及两个文本相似时的相似程度，但是目前文本多样化表述使得文本相似度模型的模型准确度降低。

发明内容

[0003] 有鉴于此，本申请的目的在于提供一种模型训练方法、文本相似度确定方法及装置，用于使得文本相似度模型的训练方法具备更强的兼容性，并提高模型准确度。技术方案如下：

[0004] 一方面，本申请提供一种模型训练方法，所述方法包括：

[0005] 获取用于对文本相似度模型进行训练的文本集合，所述文本集合包括第一文本组和第二文本组，所述第一文本组中的各个文本以第一文本标注形式标注，所述第二文本组中的各个文本以第二文本标注形式标注；

[0006] 采用目标文本标注形式对各个文本进行标注，以统一所述第一文本组和所述第二文本组中各个文本的标注形式；

[0007] 基于第一损失函数获取所述第一文本组的损失值，基于第二损失函数获取所述第二文本组的损失值；

[0008] 基于所述第一文本组的损失值和所述第二文本组的损失值，确定所述文本集合的损失值；

[0009] 利用所述文本集合的损失值，调整所述文本相似度模型的模型参数，以对所述文本相似度模型进行训练。

[0010] 另一方面，本申请提供一种文本相似度确定方法，所述方法包括：

[0011] 获取第一待处理文本和第二待处理文本；

[0012] 调用文本相似度模型，所述文本相似度模型是通过文本集合的损失值调整模型参数得到，所述文本集合的损失值基于文本集合中的第一文本组的损失值和第二文本组的损失值得到，所述第一文本组中的各个文本以第一文本标注形式标注，所述第二文本组中的各个文本以第二文本标注形式标注，在得到所述第一文本组的损失值和所述第二文本组的损失值之前，采用目标文本标注形式对各个文本进行标注，以统一所述第一文本组和所述第二文本组中各个文本的标注形式；

[0013] 获得所述文本相似度模型输出的指示所述第一待处理文本和所述第二待处理文

本是否相似的处理结果。

[0014] 再一方面，本申请提供一种模型训练装置，所述装置包括：

[0015] 获取单元，用于获取用于对文本相似度模型进行训练的文本集合，所述文本集合包括第一文本组和第二文本组，所述第一文本组中的各个文本以第一文本标注形式标注，所述第二文本组中的各个文本以第二文本标注形式标注；

[0016] 标注单元，用于采用目标文本标注形式对各个文本进行标注，以统一所述第一文本组和所述第二文本组中各个文本的标注形式；

[0017] 损失值确定单元，用于基于第一损失函数获取所述第一文本组的损失值，基于第二损失函数获取所述第二文本组的损失值；基于所述第一文本组的损失值和所述第二文本组的损失值，确定所述文本集合的损失值；

[0018] 调整单元，用于利用所述文本集合的损失值，调整所述文本相似度模型的模型参数，以对所述文本相似度模型进行训练。

[0019] 再一方面，本申请提供一种文本相似度确定装置，所述装置包括：

[0020] 文本获取单元，用于获取第一待处理文本和第二待处理文本；

[0021] 调用单元，用于调用文本相似度模型，所述文本相似度模型是通过文本集合的损失值调整模型参数得到，所述文本集合的损失值基于文本集合中的第一文本组的损失值和第二文本组的损失值得到，所述第一文本组中的各个文本以第一文本标注形式标注，所述第二文本组中的各个文本以第二文本标注形式标注，在得到所述第一文本组的损失值和所述第二文本组的损失值之前，采用目标文本标注形式对各个文本进行标注，以统一所述第一文本组和所述第二文本组中各个文本的标注形式；

[0022] 结果获得单元，用于获得所述文本相似度模型输出的指示所述第一待处理文本和所述第二待处理文本是否相似的处理结果。

[0023] 再一方面，本申请提供一种电子设备，包括：

[0024] 处理器；

[0025] 存储器，用于存储可执行指令；

[0026] 其中，所述处理器配置为经由执行所述可执行指令来执行上述模型训练方法和/或上述文本相似度确定方法。

[0027] 再一方面，本申请提供一种存储介质，所述存储介质中存储有计算机程序代码，所述计算机程序代码被运行时执行上述模型训练方法和/或上述文本相似度确定方法。

[0028] 上述模型训练方法、文本相似度确定方法及装置，获取用于对文本相似度模型进行训练的文本集合，文本集合包括第一文本组和第二文本组，第一文本组中的各个文本以第一文本标注形式标注，第二文本组中的各个文本以第二文本标注形式标注；采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式；基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值；基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值；利用文本集合的损失值，调整文本相似度模型的模型参数，以对文本相似度模型进行训练，实现基于多种标注形式的文本组对文本相似度模型进行训练，这样在进行模型训练过程中能够兼容不同标注形式的优点调整文本集合，使得文本相似度模型的训练方式灵活便捷，从而使得文本相似度模型的训练方法具备更强的兼容性。并且基于每种标注形式各自对应的损失函

数获取损失值，使得在调整模型参数过程中保留每种标注形式的优势，提高文本相似度模型的准确度。

附图说明

[0029] 为了更清楚地说明本申请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，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0030] 图1是本申请实施例提供的一种模型训练方法的流程图；

[0031] 图2是本申请实施例提供的一种获取第一文本组和第二文本组的损失值的流程图；

[0032] 图3是本申请实施例提供的另一种模型训练方法的流程图；

[0033] 图4是本申请实施例提供的一种文本相似度确定方法的流程图；

[0034] 图5是本申请实施例提供的一种模型训练装置的结构示意图；

[0035] 图6是本申请实施例提供的另一种模型训练装置的结构示意图；

[0036] 图7是本申请实施例提供的一种文本相似度确定装置的结构示意图。

具体实施方式

[0037] 申请人通过对目前文本相似度模型的研究发现：文本相似度模型的训练难度主要集中在两个方面，一方面是针对多种文本相似度标注的模型训练和模型参数调整，另一方面是相似度度量指标定义的鲁棒性问题。

[0038] 其中关于针对多种文本相似度数据标注的模型训练和模型参数调整，目前文本相似度标注包括两种文本标注方式，一种是三元组标注形式(文本A，文本P，文本N)，满足文本A与文本P的相似度大于文本A与文本N的相似度，一般通过构建triplet loss损失函数训练文本相似度模型。另一种是0-1标注形式(文本A，文本B，标签0/1)，0表示文本A与文本B不相似，1表示文本A与文本B相似，一般基于分类损失函数训练文本相似度模型。这两种文本标注方式各有优劣，三元组标注形式标注简单，标注具有可持续性且文本一致性强，细粒度，但监督性能较弱，需标注大量文本才能完成文本相似度模型训练；0-1标注形式的监督性能较强，但标注过程持续一致性较差，且分类损失函数与相似度问题不能很好地融合。目前针对文本相似度模型的训练采用上述两种文本标注方式中的任意一种对训练模型的文本进行标注，单一文本标注形式限制了模型训练及模型参数调整的灵活性，且被所采用标注方式本身的缺点所拖累。并且每个标注方式对应的损失函数也有一定限制，使得在基于单一损失函数得到的损失值进行模型训练时，会降低文本相似度模型的准确度。

[0039] 为此，本实施例提供一种模型训练方法、文本相似度确定方法及装置，基于多种标注形式的文本组对文本相似度模型进行训练，以在进行模型训练过程中能够兼容不同标注形式的优点调整文本集合，使得文本相似度模型的训练方式灵活便捷，从而使得文本相似度模型的训练方法具备更强的兼容性。并且基于每种标注形式各自对应的损失函数获取损失值，使得在调整模型参数过程中保留每种标注形式的优势，提高文本相似度模型的准确度。

[0040] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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，都属于本申请保护的范围。

[0041] 请参见图1，其示出了本申请实施例提供的一种模型训练方法的流程图，可以包括以下步骤：

[0042] 101：获取用于对文本相似度模型进行训练的文本集合，文本集合包括第一文本组和第二文本组，第一文本组中的各个文本以第一文本标注形式标注，第二文本组中的各个文本以第二文本标注形式标注，以通过多种不同标注形式的文本组对文本相似度模型进行训练。

[0043] 其中第一文本标注形式和第二文本标注形式的关系是但不限于是：第一文本标注形式在标注过程中的持续一致性强于第二文本标注形式在标注过程中的持续一致性，但第一文本标注形式的监督性能差于第二文本标注形式的监督性能。例如第一文本标注形式为三元组标注形式，第二文本标注形式为0-1标注形式。

[0044] 102：采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式。目标文本标注形式可以是集成第一文本标注形式和第二文本标注形式各自优点的一种文本标注形式，类似于上述三元组标注形式和0-1标注形式，目标文本标注形式规定一个文本组中文本的关系，因此可以直接利用目标文本标注形式对第一文本组和第二文本组中至少一个文本组中的各个文本进行标注。

[0045] 目标文本标注形式还可以是第一文本标注形式和第二文本标注形式中的任意一种，例如目标文本标注形式为第一文本标注形式，如目标文本标注形式为三元组标注形式。在目标文本标注形式为三元组标注形式的情况下，针对0-1标注形式标注的第二文本组来说，其形式转化方式如下：

[0046] 如果第二文本组中两个文本对应0标签，即第二文本组的标注为(A,B,0)，则将第二文本组以(A,B,A)形式进行标注，A和B为第二文本组中的两个文本；如果第二文本组中两个文本对应1标签，即第二文本组的标注为(A,B,1)，则将第二文本组以(A,A,B)形式进行标注，A和B为第二文本组中的两个文本，通过上述标注形式转化，第一文本组和第二文本组被一致地转化为利用三元组进行表示，如通过(A,P,N)、(A,B,A)、(A,A,B)分别表示第一文本组、具有0标签的第二文本组和具有1标签的第二文本组。

[0047] 103：基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值。第一损失函数与第一文本标注形式对应，第二损失函数与第二文本标注形式对应，使得每个文本组能够基于各自对应的损失函数得到损失值，以符合每个文本组的文本标注需求。

[0048] 104：基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值。文本集合的损失值能够将第一文本组和第二文本组的损失值进行融合，使得损失值能够兼顾具有不同标注形式的文本组，从而在利用文本集合的损失值进行训练过程中能够兼顾不同标注形式的文本组，提高文本相似度模型的准确度。

[0049] 在本实施例中，如果第一文本标注形式为三元组标注形式，第二文本标注形式为0-1标注形式，则对应的基于第一损失函数获取第一文本组的损失值，基于第二损失函数获

取第二文本组的损失值的过程如下：

[0050] 对以三元组标注形式标注的第一文本组（A，P，N），利用 $loss_{triplet}=\frac{1}{n}\sum_{(A,P,N)}max(sim(v_A-v_N)-sim(v_A-v_P)+\delta,0)$ 获取第一文本组的损失值，A，P和N为第一文本组中的三个文本，n为第一文本组的总数，v_A 为第一文本组中文本A的文本向量，v_P 为第一文本组中文本P的文本向量，v_N 第一文本组中文本N的文本向量，$\delta>0$ 表示第一文本组中相似度高的文本对与相似度低的文本对之间的相似度差异阈值。

[0051] 对具有0标签且以三元组标注形式标注的第二文本组(A,B,A)，第二文本组中文本之间的相似度越低越好，相对应的第二损失函数需要压缩第二文本组中文本之间的相似度，因此可以利用 $loss_0=\frac{1}{n}\sum_{(A,B,A)}sim(v_A,v_B)$ 获取具有0标签且以三元组标注形式标注的第二文本组的损失值，A，B和A为具有0标签且以三元组标注形式标注的第二文本组中的三个文本，n为具有0标签且以三元组标注形式标注的第二文本组的总数，v_A 为具有0标签且以三元组标注形式标注的第二文本组中文本A的文本向量，v_B 为具有0标签且以三元组标注形式标注的第二文本组中文本B的文本向量。

[0052] 对具有1标签且以三元组标注形式标注的第二文本组(A,A,B)，第二文本组中文本之间的相似度越高越好，相对应的第二损失函数需要增强第二文本组中文本之间的相似度，因此可以利用 $loss_1=\frac{1}{n}\sum_{(A,A,B)}(1-sim(v_A-v_B))$ 获取具有1标签且以三元组标注形式标注的第二文本组的损失值，A，A和B为具有1标签且以三元组标注形式标注的第二文本组中的三个文本，n为具有1标签且以三元组标注形式标注的第二文本组的总数，v_A 为具有1标签且以三元组标注形式标注的第二文本组中文本A的文本向量，v_B 为具有1标签且以三元组标注形式标注的第二文本组中文本B的文本向量。

[0053] 通过上述三种损失函数，得到所有第一文本组的损失值 $loss_{triplet}$ 、所有具有0标签且以三元组标注形式标注的第二文本组的损失值 $loss_0$ 和所有具有1标签且以三元组标注形式标注的第二文本组 $loss_1$ ，然后利用但不限于利用

$loss = loss_{triplet} + \lambda_0 \cdot loss_0 + \lambda_1 \cdot loss_1$ 得到文本集合的损失值*loss*，λ_0 为 $loss_0$ 的权重，λ_1 为 $loss_1$ 的权重。

[0054] 其中，$\lambda_0, \lambda_1 > 0$，模型参数将通过最小化loss进行调整。如果具有0标签且以三元组标注形式标注的第二文本组对应的相似度阈值为0.3，则损失值为0.3；如果具有1标签且以三元组标注形式标注的第二文本组对应的相似度阈值为0.8，则损失值为0.2；如果第一文本组对应的相似度阈值 δ 为0.4，则损失值接近0，最大不超过0.1。为了便于优化，应当保证三个文本组的损失值在同一个量级，故 λ_0, λ_1 的取值可以为0.3和0.5。

[0055] 105：利用文本集合的损失值，调整文本相似度模型的模型参数，以对文本相似度模型进行训练，文本相似度模型的模型参数调整过程与目前利用损失值进行模型参数调整相同，本实施例不再赘述。

[0056] 上述模型训练方法，获取用于对文本相似度模型进行训练的文本集合，文本集合包括第一文本组和第二文本组，第一文本组中的各个文本以第一文本标注形式标注，第二文本组中的各个文本以第二文本标注形式标注；采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式；基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值；基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值；利用文本集合的损失值，调整文本相似度模型的模型参数，以对文本相似度模型进行训练，实现基于多种标注形式的文本组对文本相似度模型进行训练，这样在进行模型训练过程中能够兼容不同标注形式的优点调整文本集合，使得文本相似度模型的训练方式灵活便捷，从而使得文本相似度模型的训练方法具备更强的兼容性。并且基于每种标注形式各自对应的损失函数获取损失值，使得在调整模型参数过程中保留每种标注形式的优势，提高文本相似度模型的准确度。

[0057] 在本实施例中，获取第一文本组和第二文本组的损失值的一种可行方式如图2所示，可以包括以下步骤：

[0058] 201：获取以目标文本标注形式标注的第一文本组中各个文本的第一文本向量，获取以目标文本标注形式标注的第二文本组中各个文本的第二文本向量。如利用但不限于利用深度学习模型获取各个文本的文本向量，如利用预训练模型ALBERT获取各个文本的文本向量。

[0059] 202：基于第一文本向量，确定第一文本组中各个文本之间的相似度，基于第二文本向量，确定第二文本组中各个文本之间的相似度。

[0060] 相似度度量指标定义的鲁棒性问题主要是通过一个合理有效地相似度算法度量文本之间的相似度。目前相似度算法包括：欧式距离、曼哈顿距离、余弦相似度和汉明距离等，最常用的相似度算法是余弦相似度和欧式距离，余弦相似度和欧式距离时针对文本的文本向量进行处理。余弦相似度通过计算两个文本向量之间的夹角余弦值来衡量文本之间是否相似以及相似程度，夹角余弦值越大表明文本越相似。欧式距离计算两个文本向量之

间的几何距离，几何距离的值越小表明文本越相似。对于高维文本向量来说(向量维度大于预设阈值)，由于在余弦相似度计算过程中，向量标准化操作会将文本向量中各分量的差异缩小，导致余弦相似度无法准确刻画两个文本的相似度，因为将文本向量中各分量的差异缩小，导致在测试过程中错误召回大量无关文本，说明余弦相似度的抗干扰能力弱，而欧式距离得到的距离值范围是 $[0, +\infty)$，取值范围太大，在工业应用场景中很难给出合理的用于确定文本相似的相似度阈值。

[0061] 本实施例在确定第一文本组中各个文本之间的相似度以及第二文本组中各个文本之间的相似度过程中可利用上述相似度算法，但是鉴于常用的余弦相似度和欧氏距离存在的问题，本实施例提供如下一种方式来确定相似度：

[0062] 一、基于第一文本向量，确定第一文本组中各个文本之间的相似度包括：

[0063] 对第一文本组中的任意两个文本，利用公式 $sim(v_1, v_2) = e^{-\frac{\|v_1 - v_2\|_2^2}{\sigma}}$ 得到任意两个文本之间的相似度，$\|\cdot\|_2$ 表示2-范数，$\sigma > 0$ 表示2-范数标准化因子，v_1 和 v_2 表示任意两个文本各自的第一文本向量。

[0064] 二、基于第二文本向量，确定第二文本组中各个文本之间的相似度包括：

[0065] 对第二文本组中的任意两个文本，利用公式 $sim(y_1, y_2) = e^{-\frac{\|y_1 - y_2\|_2^2}{\sigma}}$ 得到任意两个文本之间的相似度，$\|\cdot\|_2$ 表示2-范数，$\sigma > 0$ 表示2-范数标准化因子，y_1 和 y_2 表示任意两个文本各自的第二文本向量。

[0066] 上述相似度计算利用的公式称为负指数相似度，其包括两部分：距离负指数变换和2-范数，2-范数能够保留文本向量中各分量之间的差异，差异越大2-范数距离也会越大，相对应的相似度越低，而无关文本(即不相似文本)的无关性体现在文本向量中各分量之间的差异，因此通过上述公式具备抗无关文本干扰的能力，使得相似度的准确度提高。

[0067] 在任一文本组中任意两个文本变化少数几个字词后，对应的文本向量中各分量的差异变化也是微小的，因为2-范数能够保留这种差异，任一文本组中的任意两个文本即便变化少数几个字词，得到的相似度的变化不大，从而解决变化字词出现相似度的变化较大的问题。并且上述公式对应的相似度取值范围是 (0,1]，相对应的相似度阈值设置也可以在0至1之间选择，解决欧式距离因取值范围是 $[0, +\infty)$ 导致的相似度阈值设置困难的问题，提升相似度阈值的可干预性，经过多次试验本实施例对应的阈值可以为0.7或0.8。

[0068] 203：基于第一损失函数和第一文本向量，获取第一文本组的损失值。

[0069] 如利用但不限于利用

$$loss_{triplet}=\frac{1}{n}\sum_{(A,P,N)}max(sim(v_A-v_N)-sim(v_A-v_P)+\delta,0)$$ 获取第一文本组的损失值。

[0070] 204:基于第二损失函数和第二文本向量,获取第二文本组的损失值。

[0071] 如对具有0标签且以三元组标注形式标注的第二文本组(A,B,A),利用但不限于利用 $$loss_0=\frac{1}{n}\sum_{(A,B,A)}sim(v_A,v_B)$$ 获取具有0标签且以三元组标注形式标注的第二文本组的损失值;对具有1标签且以三元组标注形式标注的第二文本组(A,A,B),利用但不限于利用 $$loss_1=\frac{1}{n}\sum_{(A,A,B)}(1-sim(v_A,v_B))$$ 获取具有1标签且以三元组标注形式标注的第二文本组的损失值。

[0072] 请参阅图3,其示出了本申请实施例提供的另一种模型训练方法的流程图,在训练得到文本相似度模型后对文本相似度模型的模型参数进行调优,在图1基础上还可以包括以下步骤:

[0073] 106:获取测试文本组。在本实施例中测试文本组中各文本可以以第一文本标注形式和第二文本标注形式中的任意一种进行标注,并且可以同时获取多个测试文本组,通过多个测试文本组对训练得到的文本相似度模型进行测试,对训练得到的文本相似度模型进行测试则是对文本相似度模型的模型参数进行调优的过程。

[0074] 107:利用训练得到的文本相似度模型对测试文本组中的文本进行编码,得到测试文本组中各文本的文本向量。

[0075] 108:基于测试文本组中各文本的文本向量,确定测试文本组中各文本的相似结果。如通过但不限于通过 $$sim(v_1,v_2)=e^{-\frac{\|v_1-v_2\|_2^2}{\sigma}}$$ 获取测试文本组中各文本的相似度。

[0076] 109:如果相似结果与测试文本组的已知测试结果不一致,则以目标文本标注形式对测试文本组进行标注。

[0077] 如果相似结果与测试文本组的已知测试结果不一致,说明相似结果有误,因为相似结果基于文本相似度模型编码出的文本向量得到,说明文本相似度模型编码出的文本向量有误,确定文本相似度模型存在问题,进而需要对文本相似度模型的模型参数进行调优。

[0078] 在对文本相似度模型的模型参数进行调优,首先以目标文本标注形式对测试文本组进行标注,如以三元组标注形式对测试文本组进行标注。

[0079] 110:利用标注后的测试文本组中的各文本,对训练得到的文本相似度模型的模型参数进行调整。如将标注后的测试文本组加入到文本集合中,利用上述图1所示的方式得到

文本集合的损失值，然后利用文本集合的损失值重新调整文本相似度模型的模型参数。

[0080] 上述模型训练方法，在训练得到文本相似度模型后，通过测试文本组可继续对文本相似度模型的模型参数进行调整，以优化文本相似度模型的模型参数。

[0081] 下面以智能客服场景下的文本集合为例对本实施例提供的模型训练方法进行验证，其中文本集合包括以三元组标注形式标注的文本组有19625条，无关文本组有10000条。评价指标包括：三元组准确率(越大越好)、1标签与0标签相似度均值差异(越大越好)、完全无关句子相似度均值(越小越好)、完全无关句子相似度标准差(越小越好)，评价结果如表1所示。

[0082] 表1 评价结果

[0083]

试验设定	相似度	三元组准确率(%)	1标签与0标签相似度均值差异	完全无关句子相似度均值	完全无关句子相似度标准差
三元组标注	余弦相似度	96.7	0.095	0.493	0.172
三元组标注	负指数相似度	99	0.202	0.122	0.105
三元组标注和0-1标注	余弦相似度	95.3	0.123	0.478	0.198
三元组标注和0-1标注	负指数相似度	97.5	0.273	0.105	0.138

[0084] 从上述表1可知：(1) 在相同的数据条件下，采用负指数相似度较余弦相似度的测试三元组准确率更高、1标签与0标签相似度均值差异大(有利于设定相似度阈值)、完全无关句子相似度均值和标准差都更小(解决了无关句子被错误召回的问题)。(2) 在采用负指数相似度的条件下，三元组标注和0-1标注的准确率虽然有所降低，但1标签与0标签相似度均值差异得以放大；完全无关句子相似度均值更小，文本相似度模型更稳健。

[0085] 请参见图4，其示出了本申请实施例提供的一种文本相似度确定方法，以通过上述模型训练方法得到的文本相似度模型确定两个文本之间是否相似，可以包括以下步骤：

[0086] 301：获取第一待处理文本和第二待处理文本。其中第一待处理文本和第二待处理文本是用于确定是否相似的两个文本，对于第一待处理文本和第二待处理文本的来源和获取方式本实施例不进行限定。

[0087] 302：调用文本相似度模型，其中文本相似度模型是通过文本集合的损失值调整模型参数得到，文本集合的损失值基于文本集合中的第一文本组的损失值和第二文本组的损失值得到，第一文本组中的各个文本以第一文本标注形式标注，第二文本组中的各个文本以第二文本标注形式标注，在得到第一文本组的损失值和第二文本组的损失值之前，采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式，具体过程请参见上述实施例。

[0088] 在获取到第一待处理文本和第二待处理文本后，第一待处理文本和第二待处理文本作为文本相似度模型的输入，通过文本相似度模型对第一待处理文本和第二待处理文本进行编码，得到第一待处理文本的文本向量和第二待处理文本的文本向量，然后通过文本相似度模型对第一待处理文本的文本向量和第二处理文本的文本向量进行相似度确定，如利用上述负指数相似度的方式确定第一待处理文本的文本向量和第二处理文本的文本向量之间的相似度。

[0089] 303：获得文本相似度模型输出的指示第一待处理文本和第二待处理文本是否相似的处理结果。如针对文本相似度模型设置一个用于确定第一待处理文本和第二待处理文本是否相似的阈值，如果文本相似度模型输出的相似度大于阈值，确定第一待处理文本和第二待处理文本相似，否则确定第一待处理文本和第二待处理文本不相似，其中阈值的取

值本实施例不进行限定。

[0090] 上述文本相似度模型确定方法，调用上述文本相似度模型确定第一待处理文本和第二待处理文本是否相似，因上述相似度模型能够基于每种标注形式各自对应的损失函数获取损失值，使得在调整模型参数过程中保留每种标注形式的优势，提高文本相似度模型的准确度，所以调用上述文本相似度模型确定第一待处理文本和第二待处理文本是否相似过程中，可提高确定是否相似的准确度。

[0091] 对于前述的各方法实施例，为了简单描述，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他顺序或者同时进行。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是本申请所必须的。

[0092] 与上述方法实施例相对应，本申请实施例提供一种模型训练装置，其可选结构如图5所示，可以包括：获取单元10、标注单元20、损失值确定单元30和调整单元40。

[0093] 获取单元10，用于获取用于对文本相似度模型进行训练的文本集合，文本集合包括第一文本组和第二文本组，第一文本组中的各个文本以第一文本标注形式标注，第二文本组中的各个文本以第二文本标注形式标注，以通过多种不同标注形式的文本组对文本相似度模型进行训练。

[0094] 其中第一文本标注形式和第二文本标注形式的关系是但不限于是：第一文本标注形式在标注过程中的持续一致性强于第二文本标注形式在标注过程中的持续一致性，但第一文本标注形式的监督性能差于第二文本标注形式的监督性能。例如第一文本标注形式为三元组标注形式，第二文本标注形式为0-1标注形式。

[0095] 标注单元20，用于采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式。目标文本标注形式可以是集成第一文本标注形式和第二文本标注形式各自优点的一种文本标注形式，类似于上述三元组标注形式和0-1标注形式，目标文本标注形式规定一个文本组中文本的关系，因此可以直接利用目标文本标注形式对第一文本组和第二文本组中至少一个文本组中的各个文本进行标注。

[0096] 目标文本标注形式还可以是第一文本标注形式和第二文本标注形式中的任意一种，例如目标文本标注形式为第一文本标注形式，如目标文本标注形式为三元组标注形式。在目标文本标注形式为三元组标注形式的情况下，针对0-1标注形式标注的第二文本组来说，其形式转化方式如下：

[0097] 如果第二文本组中两个文本对应0标签，即第二文本组的标注为(A,B,0)，则将第二文本组以(A,B,A)形式进行标注，A和B为第二文本组中的两个文本；如果第二文本组中两个文本对应1标签，即第二文本组的标注为(A,B,0)，则将第二文本组以(A,A,B)形式进行标注，A和B为第二文本组中的两个文本，通过上述标注形式转化，第一文本组和第二文本组被一致地转化为利用三元组进行表示，如通过(A,P,N)、(A,B,A)、(A,A,B)分别表示第一文本组、具有0标签的第二文本组和具有1标签的第二文本组。

[0098] 损失值确定单元30，用于基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值；基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值。

[0099] 第一损失函数与第一文本标注形式对应，第二损失函数与第二文本标注形式对

应，使得每个文本组能够基于各自对应的损失函数得到损失值，以符合每个文本组的文本标注需求。

[0100] 在本实施例中，损失值确定单元30获取第一文本组的损失值和第二文本组的损失值的过程如下：

[0101] 获取以目标文本标注形式标注的第一文本组中各个文本的第一文本向量，获取以目标文本标注形式标注的第二文本组中各个文本的第二文本向量；基于第一文本向量，确定第一文本组中各个文本之间的相似度，基于第二文本向量，确定第二文本组中各个文本之间的相似度；基于第一损失函数和第一文本向量，获取第一文本组的损失值；基于第二损失函数和第二文本向量，获取第二文本组的损失值。

[0102] 其中基于第一文本向量，确定第一文本组中各个文本之间的相似度包括：对第一文本组中的任意两个文本，利用公式 $sim(v_1, v_2) = e^{-\frac{\|v_1 - v_2\|_2}{\sigma}}$ 得到任意两个文本之间的相似度，$\|.\|_2$ 表示2-范数，$\sigma > 0$ 表示2-范数标准化因子，v_1 和 v_2 表示任意两个文本各自的第一文本向量；

[0103] 基于第二文本向量，确定第二文本组中各个文本之间的相似度包括：对第二文本组中的任意两个文本，利用公式 $sim(y_1, y_2) = e^{-\frac{\|y_1 - y_2\|_2^2}{\sigma}}$ 得到任意两个文本之间的相似度，$\|.\|_2$ 表示2-范数，$\sigma > 0$ 表示2-范数标准化因子，y_1 和 y_2 表示任意两个文本各自的第二文本向量。

[0104] 如果第一文本标注形式为三元组标注形式，第二文本标注形式为0-1标注形式，则损失值确定单元30获取第一文本组的损失值和第二文本组的损失值的过程如下：

[0105] 对以三元组标注形式标注的第一文本组（A，P，N），利用 $loss_{triplet} = \frac{1}{n}\sum_{(A,P,N)} max(sim(v_A - v_N) - sim(v_A - v_P) + \delta, 0)$ 获取第一文本组的损失值，A，P和N为第一文本组中的三个文本，n为第一文本组的总数，v_A 为第一文本组中文本A的文本向量，v_P 为第一文本组中文本P的文本向量，v_N 第一文本组中文本N的文本向量，$\sigma > 0$ 表示第一文本组中相似度高的文本对与相似度低的文本对之间的相似度差异阈值；

[0106] 对具有0标签且以三元组标注形式标注的第二文本组（A，B，A），利用

$loss_0 = \frac{1}{n}\sum_{(A,B,A)} sim(v_A, v_B)$ 获取具有0标签且以三元组标注形式标注的第二文本组的损失值，A，B和A为具有0标签且以三元组标注形式标注的第二文本组中的三个文本，n为具有0标签且以三元组标注形式标注的第二文本组的总数，v_A 为具有0标签且以三元组标注形式标注的第二文本组中文本A的文本向量，v_B 为具有0标签且以三元组标注形式标注的第二文本组中文本B的文本向量；

[0107] 对具有1标签且以三元组标注形式标注的第二文本组（A，A，B），利用 $loss_1 = \frac{1}{n}\sum_{(A,A,B)} (1 - sim(v_A, v_B))$ 获取具有1标签且以三元组标注形式标注的第二文本组的损失值，A，A和B为具有1标签且以三元组标注形式标注的第二文本组中的三个文本，n为具有1标签且以三元组标注形式标注的第二文本组的总数，v_A 为具有1标签且以三元组标注形式标注的第二文本组中文本A的文本向量，v_B 为具有1标签且以三元组标注形式标注的第二文本组中文本B的文本向量。

[0108] 相对应的，确定文本集合的损失值的过程包括：利用 $loss = loss_{triplet} + \lambda_0 \cdot loss_0 + \lambda_1 \cdot loss_1$ 得到文本集合的损失值 $loss$，λ_0 为 $loss_0$ 的权重，λ_1 为 $loss_1$ 的权重。

[0109] 对于损失值确定单元30的详细说明，请参见上述方法实施例，此处不再赘述。

[0110] 调整单元40，用于利用文本集合的损失值，调整文本相似度模型的模型参数，以对文本相似度模型进行训练，文本相似度模型的模型参数调整过程与目前利用损失值进行模型参数调整相同，本实施例不再赘述。

[0111] 上述模型训练装置，获取用于对文本相似度模型进行训练的文本集合，文本集合包括第一文本组和第二文本组，第一文本组中的各个文本以第一文本标注形式标注，第二文本组中的各个文本以第二文本标注形式标注；采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式；基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值；基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值；利用文本集合的损失值，调整文本相似度模型的模型参数，以对文本相似度模型进行训练，实现基于多种标注形式的文本组对文本相似度模型进行训练，这样在进行模型训练过程中能够兼容不同标注形式的优点调整文本集合，使得文本相似度模型的训练方式灵活便捷，从而使得文本相似度模型的训练方法具备更强的兼容性。并且基于每种标注形式各自对应的损失函数获取损失值，使得在调整模型参数过程中保留每种标注形式的优势，提高文本相似度模型的准确度。

[0112] 请参见图6，其示出了本申请实施例提供的另一种模型训练装置的可选结构，还可

以包括:编码单元50和确定单元60。

[0113] 获取单元10,还用于获取测试文本组。在本实施例中测试文本组中各文本可以以第一文本标注形式和第二文本标注形式中的任意一种进行标注,并且可以同时获取多个测试文本组,通过多个测试文本组对训练得到的文本相似度模型进行测试,对训练得到的文本相似度模型进行测试则是对文本相似度模型的模型参数进行调优的过程。

[0114] 编码单元50,用于利用训练得到的文本相似度模型对测试文本组中的文本进行编码,得到测试文本组中各文本的文本向量。

[0115] 确定单元60,用于基于测试文本组中各文本的文本向量,确定测试文本组中各文本的相似结果。

[0116] 标注单元20,还用于如果相似结果与测试文本组的已知测试结果不一致,则以目标文本标注形式对测试文本组进行标注。

[0117] 如果相似结果与测试文本组的已知测试结果不一致,说明相似结果有误,因为相似结果基于文本相似度模型编码出的文本向量得到,说明文本相似度模型编码出的文本向量有误,确定文本相似度模型存在问题,进而需要对文本相似度模型的模型参数进行调优。

[0118] 在对文本相似度模型的模型参数进行调优,首先以目标文本标注形式对测试文本组进行标注,如以三元组标注形式对测试文本组进行标注。

[0119] 调整单元40,还用于利用标注后的测试文本组中的各文本,对训练得到的文本相似度模型的模型参数进行调整。如将标注后的测试文本组加入到文本集合中,利用上述图1所示的方式得到文本集合的损失值,然后利用文本集合的损失值重新调整文本相似度模型的模型参数。

[0120] 上述模型训练装置,在训练得到文本相似度模型后,通过测试文本组可继续对文本相似度模型的模型参数进行调整,以优化文本相似度模型的模型参数。

[0121] 请参见图7,其示出了本申请实施例提供的一种文本相似度确定装置的可选结构,可以包括:文本获取单元100、调用单元200和结果获得单元300。

[0122] 文本获取单元100,用于获取第一待处理文本和第二待处理文本。其中第一待处理文本和第二待处理文本是用于确定是否相似的两个文本,对于第一待处理文本和第二待处理文本的来源和获取方式本实施例不进行限定。

[0123] 调用单元200,用于调用文本相似度模型,文本相似度模型是通过文本集合的损失值调整模型参数得到,文本集合的损失值基于文本集合中的第一文本组的损失值和第二文本组的损失值得到,第一文本组中的各个文本以第一文本标注形式标注,第二文本组中的各个文本以第二文本标注形式标注,在得到第一文本组的损失值和第二文本组的损失值之前,采用目标文本标注形式对各个文本进行标注,以统一第一文本组和第二文本组中各个文本的标注形式,具体过程请参见上述实施例。

[0124] 在获取到第一待处理文本和第二待处理文本后,第一待处理文本和第二待处理文本作为文本相似度模型的输入,通过文本相似度模型对第一待处理文本和第二待处理文本进行编码,得到第一待处理文本的文本向量和第二待处理文本的文本向量,然后通过文本相似度模型对第一待处理文本的文本向量和第二处理文本的文本向量进行相似度确定,如利用上述负指数相似度的方式确定第一待处理文本的文本向量和第二处理文本的文本向量之间的相似度。

[0125] 结果获得单元300，用于获得文本相似度模型输出的指示第一待处理文本和第二待处理文本是否相似的处理结果。

[0126] 上述文本相似度模型确定装置，调用上述文本相似度模型确定第一待处理文本和第二待处理文本是否相似，因上述相似度模型能够基于每种标注形式各自对应的损失函数获取损失值，使得在调整模型参数过程中保留每种标注形式的优势，提高文本相似度模型的准确度，所以调用上述文本相似度模型确定第一待处理文本和第二待处理文本是否相似过程中，可提高确定是否相似的准确度。

[0127] 本申请实施例还提供一种电子设备，包括：处理器和存储器。

[0128] 存储器，用于存储可执行指令。处理器配置为经由执行可执行指令来执行上述模型训练方法和/或上述文本相似度确定方法。

[0129] 本申请实施例还提供一种存储介质，存储介质中存储有计算机程序代码，计算机程序代码被运行时执行上述模型训练方法和/或上述文本相似度确定方法。

[0130] 需要说明的是，本说明书中的各个实施例可以采用递进的方式描述、本说明书中各实施例中记载的特征可以相互替换或者组合，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。对于装置类实施例而言，由于其与方法实施例基本相似，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。

[0131] 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0132] 对所公开的实施例的上述说明，使本领域技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

[0133] 以上所述仅是本申请的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

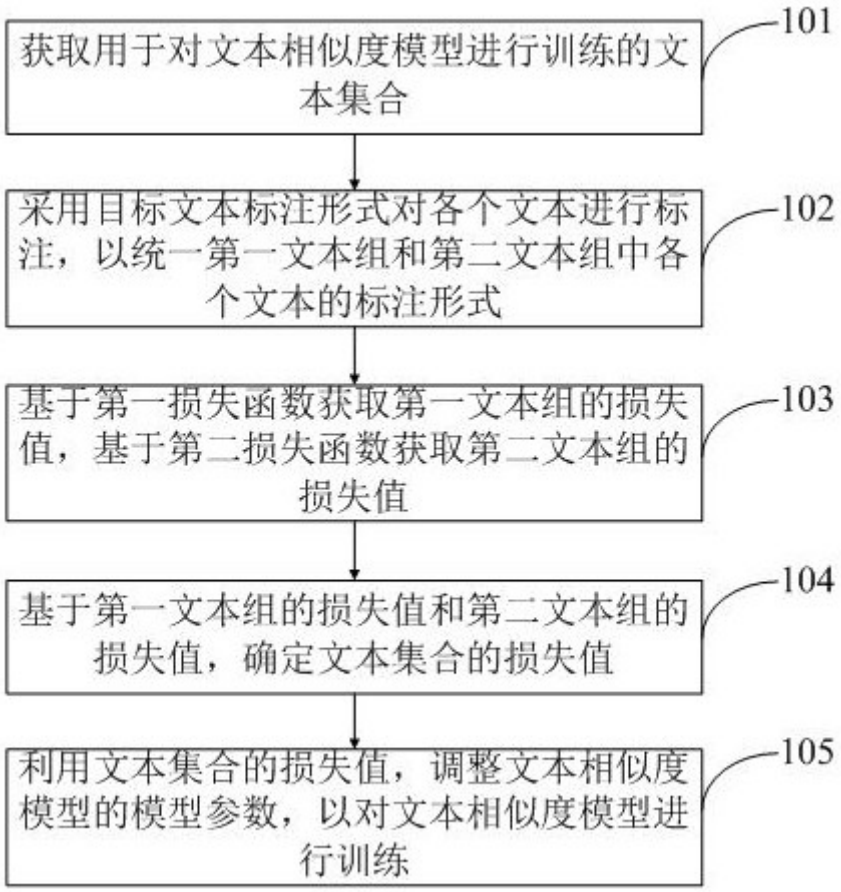

获取用于对文本相似度模型进行训练的文本集合
101
采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式
102
基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值
103
基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值
104
利用文本集合的损失值，调整文本相似度模型的模型参数，以对文本相似度模型进行训练
105

图1

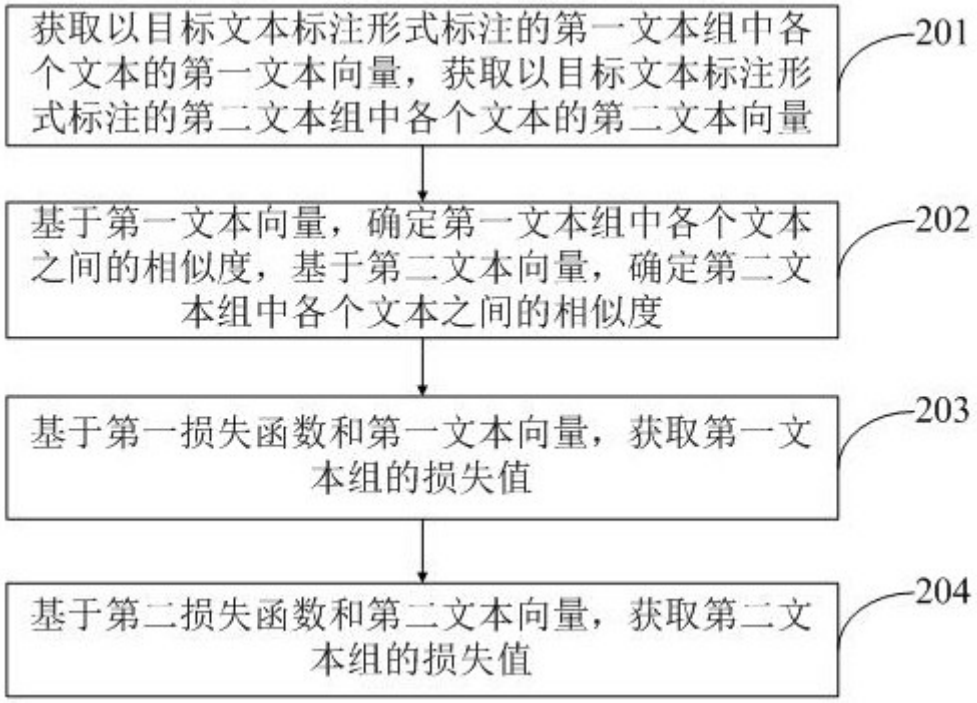

获取以目标文本标注形式标注的第一文本组中各个文本的第一文本向量，获取以目标文本标注形式标注的第二文本组中各个文本的第二文本向量
201
基于第一文本向量，确定第一文本组中各个文本之间的相似度，基于第二文本向量，确定第二文本组中各个文本之间的相似度
202
基于第一损失函数和第一文本向量，获取第一文本组的损失值
203
基于第二损失函数和第二文本向量，获取第二文本组的损失值
204

图2

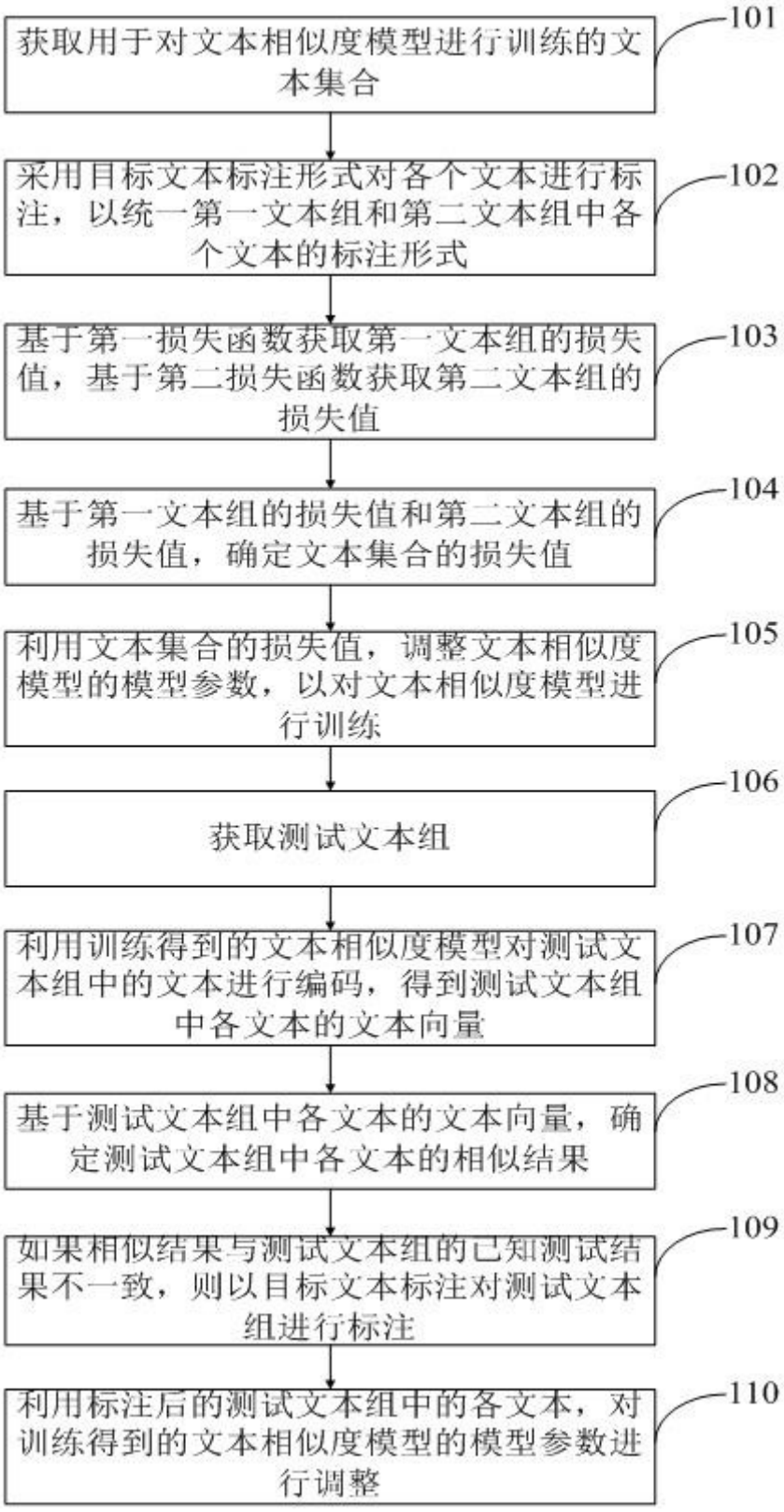

获取用于对文本相似度模型进行训练的文本集合
101
采用目标文本标注形式对各个文本进行标注，以统一第一文本组和第二文本组中各个文本的标注形式
102
基于第一损失函数获取第一文本组的损失值，基于第二损失函数获取第二文本组的损失值
103
基于第一文本组的损失值和第二文本组的损失值，确定文本集合的损失值
104
利用文本集合的损失值，调整文本相似度模型的模型参数，以对文本相似度模型进行训练
105
获取测试文本组
106
利用训练得到的文本相似度模型对测试文本组中的文本进行编码，得到测试文本组中各文本的文本向量
107
基于测试文本组中各文本的文本向量，确定测试文本组中各文本的相似结果
108
如果相似结果与测试文本组的已知测试结果不一致，则以目标文本标注对测试文本组进行标注
109
利用标注后的测试文本组中的各文本，对训练得到的文本相似度模型的模型参数进行调整
110

图3

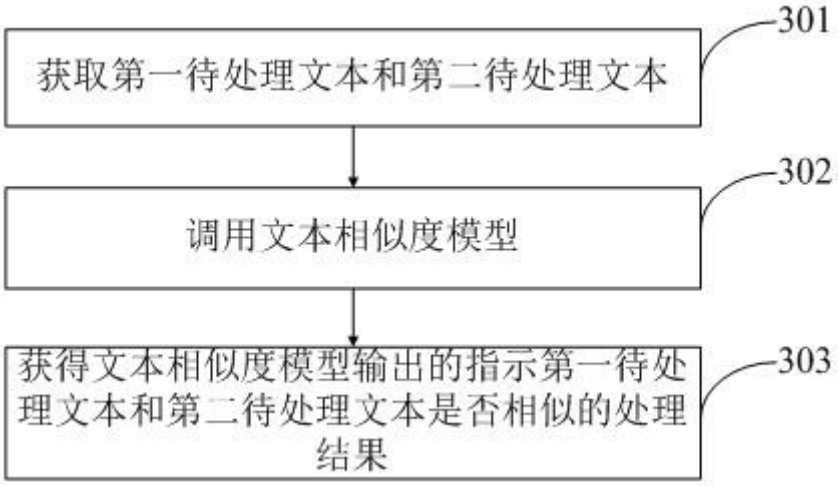

301
获取第一待处理文本和第二待处理文本
302
调用文本相似度模型
303
获得文本相似度模型输出的指示第一待处理文本和第二待处理文本是否相似的处理结果

图4

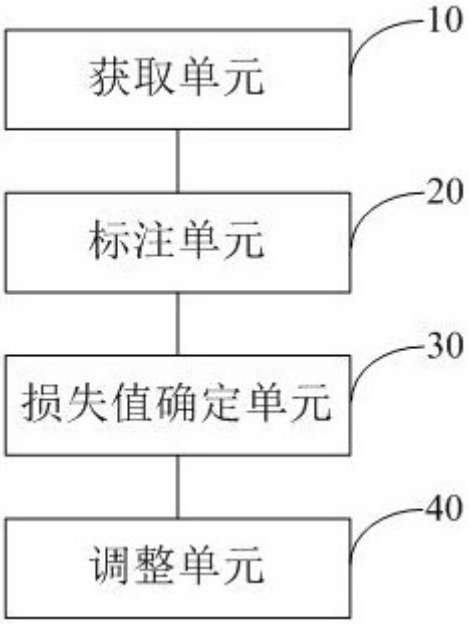

10
获取单元
20
标注单元
30
损失值确定单元
40
调整单元

图5

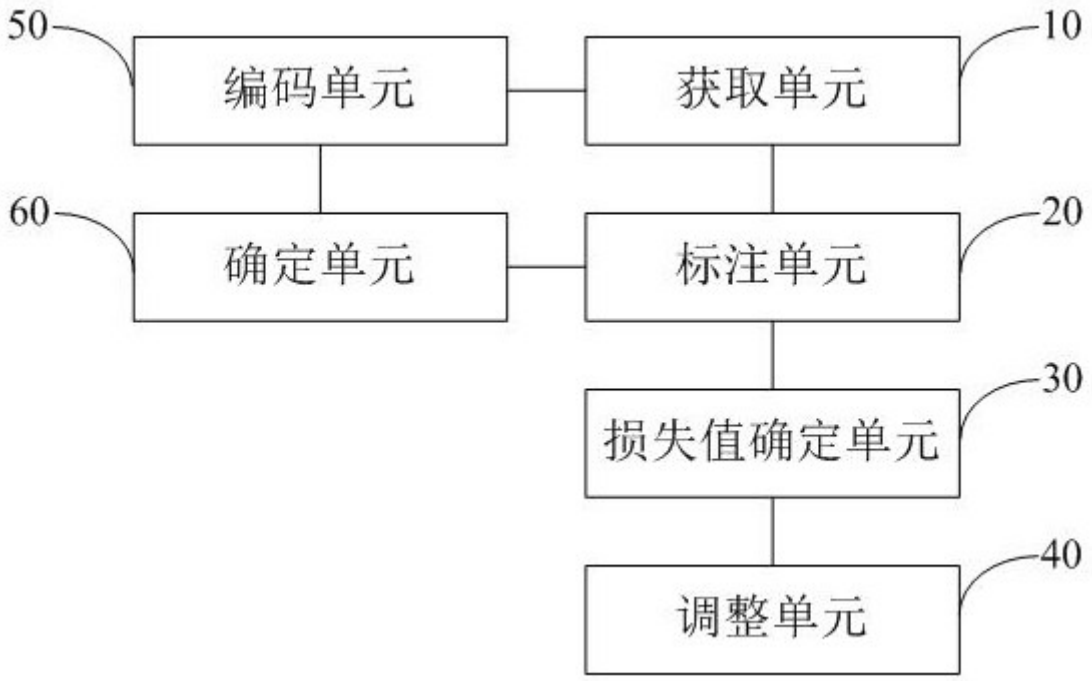

50
编码单元
获取单元
10
60
确定单元
标注单元
20
30
损失值确定单元
40
调整单元

图6

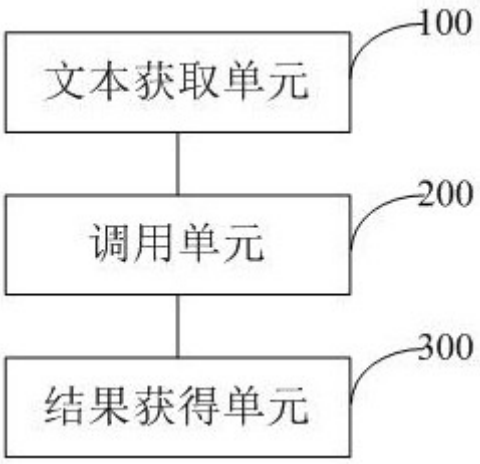

100
文本获取单元
200
调用单元
300
结果获得单元

图7