(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111767731 A
(43)申请公布日 2020.10.13

(21)申请号 202010655492.3

(22)申请日 2020.07.09

(71)申请人 北京猿力未来科技有限公司
地址 100102 北京市朝阳区广顺南大街8号院1号楼6层F01-03、05-10单元

(72)发明人 何苏 王亮 赵薇 刘金龙 柳景明 郭常圳

(74)专利代理机构 北京智信禾专利代理有限公司 11637
代理人 刘晓楠

(51)Int.Cl.
G06F 40/289(2020.01)
G06N 3/04(2006.01)

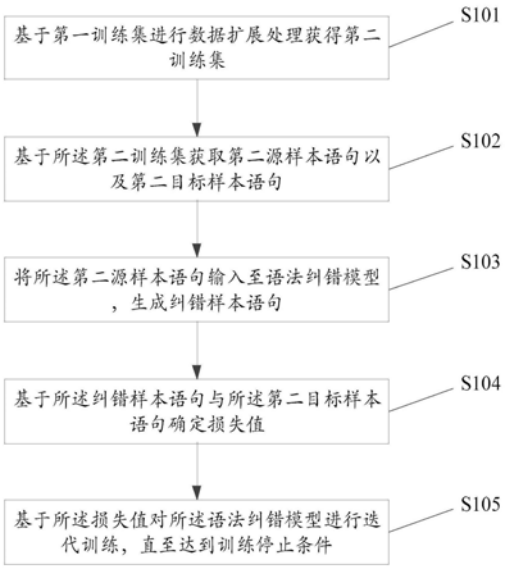
权利要求书3页 说明书20页 附图5页

(54)发明名称

语法纠错模型的训练方法及装置、语法纠错方法及装置

(57)摘要

本申请涉及一种语法纠错模型的训练方法及装置、语法纠错方法及装置、计算设备及计算机可读存储介质。训练方法包括:基于第一训练集进行数据扩展处理获得第二训练集;基于所述第二训练集获取第二源样本语句以及第二目标样本语句;将所述第二源样本语句输入至语法纠错模型,生成纠错样本语句;基于所述纠错样本语句与所述第二目标样本语句确定损失值;基于所述损失值对所述语法纠错模型进行迭代训练,直至达到训练停止条件。通过对已有的训练集进行数据增强处理,达到对训练集进行自动扩充的目的,有效地减少了人工劳动。



CN 111767731 A

1.一种语法纠错模型的训练方法,其特征在于,包括:

基于第一训练集进行数据扩展处理获得第二训练集;

基于所述第二训练集获取第二源样本语句以及第二目标样本语句;

将所述第二源样本语句输入至语法纠错模型,生成纠错样本语句;

基于所述纠错样本语句与所述第二目标样本语句确定损失值;

基于所述损失值对所述语法纠错模型进行迭代训练,直至达到训练停止条件。

2.根据权利要求1所述的训练方法,其特征在于,所述第一训练集包括第一源样本语句和第一目标样本语句;

所述基于第一训练集进行数据扩展处理获得第二训练集,包括:

对所述第一源样本语句和第一目标样本语句进行预处理;

基于所述第一训练集中词单元的出现频率,对所述词单元进行权重赋值,构建词典;

根据所述词典对所述第一训练集的源样本语句包含的语句进行腐化处理,获得数据扩展的第二源样本语句;根据所述第二源样本语句以及所述第二源样本语句对应的第二目标样本语句构建所述第二训练集。

3.根据权利要求2所述的训练方法,其特征在于,所述腐化处理包括词插入处理和/或词替代处理;

所述根据所述词典对所述第一训练集的源样本语句包含的语句进行腐化处理,获得数据扩展的第二源样本语句,包括:

根据所述词典对所述第一源样本语句进行词插入处理,获得数据扩展的第二源样本语句;和/或根据所述词典对所述第一源样本语句进行词替代处理,获得数据扩展的第二源样本语句。

4.根据权利要求3所述的训练方法,其特征在于,所述根据所述词典对所述第一源样本语句进行词插入处理,获得数据扩展的第二源样本语句,包括:

a1、获取所述第一源样本语句以及第一源样本语句的句长n;

a2、基于所述第一源样本语句的句长n生成对应的第一数组;

其中,所述第一数组中每个数值均为随机生成的(0,1)范围内的数值;

且所述第一数组中每个数值均具有与该数值在所述第一数组中位置顺序对应的下标i,所述下标i的取值范围是(0,n-1)范围内的整数;

a3、根据预设的第一阈值,获取所述第一数组中小于所述第一阈值的数值对应的下标i;

a4、基于权重随机选择所述词典中的一个词单元,插入所述第一源样本语句中的第i位置,生成词插入处理后数据扩展的第二源样本语句。

5.根据权利要求3所述的训练方法,其特征在于,所述根据所述词典对所述第一源样本语句进行词替代处理,获得数据扩展的第二源样本语句,包括:

b1、获取所述第一源样本语句以及第一源样本语句的句长n;

b2、基于所述第一源样本语句的句长生成对应的第二数组,其中,所述第二数组中每个数值均为随机生成的(0,1)范围内的数值;

且所述第二数组中每个数值均具有与该数值在所述第二数组中位置顺序对应的下标i,所述下标i的取值范围是(0,n-1)范围内的整数;

b3、根据预设的第二阈值，获取所述第二数组中小于所述第二阈值的数值对应的下标i；

b4、基于权重随机选择所述词典中的一个词单元，将所述第一源样本语句中第i位置的词单元替换为所述随机选择的词单元，生成词替代处理后数据扩展的第二源样本语句。

6.根据权利要求1所述的训练方法，其特征在于，所述第一训练集包括第一源样本语句和第一目标样本语句；

所述基于第一训练集进行数据扩展处理获得第二训练集，还包括：

c1、对所述第一源样本语句和第一目标样本语句进行预处理；

c2、基于所述第一源样本语句和第一目标样本语句，构建<第一目标样本语句，第一源样本语句>形式的反向训练集；

c3、基于所述反向训练集对所述语法纠错模型进行反向训练，其中，将所述第一目标样本语句作为所述语法纠错模型的输入，将所述第一源样本语句作为所述语法纠错模型的目标输出，进行预设代数的训练后，将所述语法纠错模型的参数固定；

c4、将所述反向训练集中的第一目标样本语句输入所述参数固定的语法纠错模型，通过集束搜索生成预设数量的候选纠错语句；

c5、对所述预设数量的候选纠错语句进行重排序，选取预设顺序的候选纠错语句作为第二源样本语句；

c6、根据所述第二源样本语句以及所述第二源样本语句对应第二目标样本语句构建所述第二训练集。

7.根据权利要求2或6所述的训练方法，其特征在于，所述对所述第一源样本语句和第一目标样本语句进行预处理，包括：

对所述第一源样本语句及所述第一目标样本语句进行分词处理，将每个词单元之间进行分隔处理；

去除所述第一训练集中句长大于预设阈值的语句；

去除所述第一源样本语句和第一目标样本语句中相同的语句。

8.根据权利要求1所述的训练方法，其特征在于，所述将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句，包括：

将所述第二源样本语句输入至所述语法纠错模型的编码器进行编码，生成编码向量；

将所述编码向量输入至所述语法纠错模型的解码器进行解码得到预设数量的候选纠错样本语句；

对所述预设数量的候选纠错样本语句进行重排序；

根据所述重排序结果，将分值最高的候选纠错样本语句作为纠错样本语句。

9.根据权利要求1所述的训练方法，其特征在于，所述基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件，包括：

判断所述损失值是否小于预设阈值；

若否，则继续获取待处理样本语句以及标签语句进行训练；

若是，则停止训练。

10.一种语法纠错的方法，其特征在于，包括：

获取源语句；

将所述源语句输入至语法纠错模型，生成语法纠错语句；

其中，所述语法纠错模型是通过权利要求1-9任一项所述的训练方法训练得到的。

11.根据权利要求1所述的语法纠错的方法，其特征在于，所述将所述源语句输入至语法纠错模型，生成语法纠错语句，包括：

将所述源语句输入至所述语法纠错模型的编码器进行编码，生成编码向量；

将所述编码向量输入至所述语法纠错模型的解码器进行解码，生成所述语法纠错语句。

12.一种语法纠错模型的训练装置，其特征在于，包括：

数据扩展模块，被配置为基于第一训练集进行数据扩展处理获得第二训练集；

获取模块，被配置为基于所述第二训练集获取第二源样本语句以及第二目标样本语句；

语法纠错模块，被配置为将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句；

损失确定模块，被配置为基于所述纠错样本语句与所述第二目标样本语句确定损失值；

迭代训练模块，被配置为基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件。

13.一种语法纠错的装置，其特征在于，包括：

获取模块，被配置为获取源语句；

语法纠错模块，被配置为将所述源语句输入至语法纠错模型，生成语法纠错语句；

其中，所述语法纠错模型是通过权利要求1-9任一项所述的训练方法训练得到的。

14.一种计算设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机指令，其特征在于，所述处理器执行所述指令时实现权利要求1-9或权利要求10-11任意一项所述方法的步骤。

15.一种计算机可读存储介质，其存储有计算机指令，其特征在于，该指令被处理器执行时实现权利要求1-9或权利要求10-11任意一项所述方法的步骤。

语法纠错模型的训练方法及装置、语法纠错方法及装置

技术领域

[0001] 本申请涉及计算机技术领域，特别涉及一种语法纠错模型的训练方法及装置、语法纠错方法及装置、计算设备及计算机可读存储介质。

背景技术

[0002] 在利用神经网络模型进行中文语法纠错时，往往需要大量的标注数据。对于缺少标注数据的情况，往往采用雇佣标注人员对数据进行标注，而人工标注数据往往耗时耗力。

[0003] 现有技术中存在的技术问题为：让机器自动的对有语法错误的中文语句进行纠错，往往达不到预期的效果，其中很重要的一个原因就是缺乏大量的标注数据。这是因为中文语法错误种类繁多，并且不同的标注人员可能对同一错误有不同的标注结果，所以这也就要求我们采取某种自动化的方式来扩充训练集。

发明内容

[0004] 有鉴于此，本申请提供了一种语法纠错模型的训练方法及装置、语法纠错方法及装置、计算设备及计算机可读存储介质，以解决现有技术中存在的技术缺陷。

[0005] 具体来说本申请提供了如下技术方案：

[0006] 本申请提供了一种语法纠错模型的训练方法，包括：

[0007] 基于第一训练集进行数据扩展处理获得第二训练集；

[0008] 基于所述第二训练集获取第二源样本语句以及第二目标样本语句；

[0009] 将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句；

[0010] 基于所述纠错样本语句与所述第二目标样本语句确定损失值；

[0011] 基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件。

[0012] 可选地，对于所述的训练方法，其中，所述第一训练集包括第一源样本语句和第一目标样本语句；

[0013] 所述基于第一训练集进行数据扩展处理获得第二训练集，包括：

[0014] 对所述第一源样本语句和第一目标样本语句进行预处理；

[0015] 基于所述第一训练集中词单元的出现频率，对所述词单元进行权重赋值，构建词典；

[0016] 根据所述词典对所述第一训练集的源样本语句包含的语句进行腐化处理，获得数据扩展的第二源样本语句；根据所述第二源样本语句以及所述第二源样本语句对应的第二目标样本语句构建所述第二训练集。

[0017] 可选地，对于所述的训练方法，其中，所述腐化处理包括词插入处理和/或词替代处理；

[0018] 所述根据所述词典对所述第一训练集的源样本语句包含的语句进行腐化处理，获得数据扩展的第二源样本语句，包括：

[0019] 根据所述词典对所述第一源样本语句进行词插入处理，获得数据扩展的第二源样

本语句;和/或根据所述词典对所述第一源样本语句进行词替代处理,获得数据扩展的第二源样本语句。

[0020] 可选地,对于所述的训练方法,其中,所述根据所述词典对所述第一源样本语句进行词插入处理,获得数据扩展的第二源样本语句,包括:

[0021] a1、获取所述第一源样本语句以及第一源样本语句的句长n;

[0022] a2、基于所述第一源样本语句的句长n生成对应的第一数组;

[0023] 其中,所述第一数组中每个数值均为随机生成的(0,1)范围内的数值;

[0024] 且所述第一数组中每个数值均具有与该数值在所述第一数组中位置顺序对应的下标i,所述下标i的取值范围是(0,n-1)范围内的整数;

[0025] a3、根据预设的第一阈值,获取所述第一数组中小于所述第一阈值的数值对应的下标i;

[0026] a4、基于权重随机选择所述词典中的一个词单元,插入所述第一源样本语句中的第i位置,生成词插入处理后数据扩展的第二源样本语句。

[0027] 可选地,对于所述的训练方法,其中,所述根据所述词典对所述第一源样本语句进行词替代处理,获得数据扩展的第二源样本语句,包括:

[0028] b1、获取所述第一源样本语句以及第一源样本语句的句长n;

[0029] b2、基于所述第一源样本语句的句长生成对应的第二数组,其中,所述第二数组中每个数值均为随机生成的(0,1)范围内的数值;

[0030] 且所述第二数组中每个数值均具有与该数值在所述第二数组中位置顺序对应的下标i,所述下标i的取值范围是(0,n-1)范围内的整数;

[0031] b3、根据预设的第二阈值,获取所述第二数组中小于所述第二阈值的数值对应的下标i;

[0032] b4、基于权重随机选择所述词典中的一个词单元,将所述第一源样本语句中第i位置的词单元替换为所述随机选择的词单元,生成词替代处理后数据扩展的第二源样本语句。

[0033] 可选地,对于所述的训练方法,其中,所述第一训练集包括第一源样本语句和第一目标样本语句;

[0034] 所述基于第一训练集进行数据扩展处理获得第二训练集,还包括:

[0035] c1、对所述第一源样本语句和第一目标样本语句进行预处理;

[0036] c2、基于所述第一源样本语句和第一目标样本语句,构建<第一目标样本语句,第一源样本语句>形式的反向训练集;

[0037] c3、基于所述反向训练集对所述语法纠错模型进行反向训练,其中,将所述第一目标样本语句作为所述语法纠错模型的输入,将所述第一源样本语句作为所述语法纠错模型的目标输出,进行预设代数的训练后,将所述语法纠错模型的参数固定;

[0038] c4、将所述反向训练集中的第一目标样本语句输入所述参数固定的语法纠错模型,通过集束搜索生成预设数量的候选纠错语句;

[0039] c5、对所述预设数量的候选纠错语句进行重排序,选取预设顺序的候选纠错语句作为第二源样本语句;

[0040] c6、根据所述第二源样本语句以及所述第二源样本语句对应第二目标样本语句构

建所述第二训练集。

[0041] 可选地，对于所述的训练方法，其中，所述对所述第一源样本语句和第一目标样本语句进行预处理，包括：

[0042] 对所述第一源样本语句及所述第一目标样本语句进行分词处理，将每个词单元之间进行分隔处理；

[0043] 去除所述第一训练集中句长大于预设阈值的语句；

[0044] 去除所述第一源样本语句和第一目标样本语句中相同的语句。

[0045] 可选地，对于所述的训练方法，其中，所述将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句，包括：

[0046] 将所述第二源样本语句输入至所述语法纠错模型的编码器进行编码，生成编码向量；

[0047] 将所述编码向量输入至所述语法纠错模型的解码器进行解码得到预设数量的候选纠错样本语句；

[0048] 对所述预设数量的候选纠错样本语句进行重排序；

[0049] 根据所述重排序结果，将分值最高的候选纠错样本语句作为纠错样本语句。

[0050] 可选地，对于所述的训练方法，其中，所述基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件，包括：

[0051] 判断所述损失值是否小于预设阈值；

[0052] 若否，则继续获取待处理样本语句以及标签语句进行训练；

[0053] 若是，则停止训练。

[0054] 本申请提供了一种语法纠错的方法，包括：

[0055] 获取源语句；

[0056] 将所述源语句输入至语法纠错模型，生成语法纠错语句；

[0057] 其中，所述语法纠错模型是通过权利要求1-9任一项所述的训练方法训练得到的。

[0058] 可选地，对于所述的语法纠错的方法，其中，所述将所述源语句输入至语法纠错模型，生成语法纠错语句，包括：

[0059] 将所述源语句输入至所述语法纠错模型的编码器进行编码，生成编码向量；

[0060] 将所述编码向量输入至所述语法纠错模型的解码器进行解码，生成所述语法纠错语句。

[0061] 本申请提供了一种语法纠错模型的训练装置，包括：

[0062] 数据扩展模块，被配置为基于第一训练集进行数据扩展处理获得第二训练集；

[0063] 获取模块，被配置为基于所述第二训练集获取第二源样本语句以及第二目标样本语句；

[0064] 语法纠错模块，被配置为将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句；

[0065] 损失确定模块，被配置为基于所述纠错样本语句与所述第二目标样本语句确定损失值；

[0066] 迭代训练模块，被配置为基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件。

[0067] 本申请提供了一种语法纠错的装置，包括：

[0068] 获取模块，被配置为获取源语句；

[0069] 语法纠错模块，被配置为将所述源语句输入至语法纠错模型，生成语法纠错语句；

[0070] 其中，所述语法纠错模型是通过权利要求1-9任一项所述的训练方法训练得到的。

[0071] 本申请还提供了一种计算设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机指令，其中，所述处理器执行所述指令时实现前述任一段所述方法的步骤。

[0072] 本申请还提供了一种计算机可读存储介质，其存储有计算机指令，其中，该指令被处理器执行时实现前述任一段所述方法的步骤。

[0073] 有益效果：

[0074] 本申请提供的一种纠错模型的训练方法，通过对已有的训练集进行数据增强处理，达到对训练集进行自动扩充的目的，有效地减少了人工劳动。

[0075] 本申请在进行数据增强处理过程中，将腐化语料处理以及反向翻译处理结合起来进行应用，有效地扩大了已有训练集的数据。

[0076] 并且本申请在进行腐化语料处理时，在词插入以及词替换过程中，对词典中的词单元进行权重赋值，通过基于词典中词单元的权重大小进行随机选择插入或者替换的词单元，词典中权重大的词单元更容易被选中，相比于直接随机选择一个词单元，基于权重进行选择可以考虑到词频的大小去挑选词，更符合语法错误的客观规律，进一步增加模型的准确性。

[0077] 本申请提供的语法纠错方法，利用训练好的语法纠错模型可以实现对于源语句的语法纠错，提高了语法纠错准确度。

[0078] 本申请提供的语法纠错模型的训练装置，通过对已有的训练集进行数据增强处理，达到对训练集进行自动扩充的目的，有效地减少了人工劳动；在本申请提供的训练装置中，将腐化语料单元以及反向翻译单元结合起来进行应用，有效地扩大了训练装置中训练集的数据；在训练装置中的词插入处理字单元以及词替代处理子单元中，基于词典中词单元的权重大小进行随机选择插入或者替换的词单元，词典中权重大的词单元更容易被选中，相比于直接随机选择一个词单元，基于权重进行选择词单元后进行插入或替代，更符合语法错误的客观规律，进一步增加训练装置的准确性。

[0079] 本申请提供的语法纠错的装置，利用训练好的语法纠错模型可以实现对于源语句的语法纠错，提高了语法纠错准确度。

附图说明

[0080] 图1是本申请实施例一提供的语法纠错模型训练方法的流程示意图；

[0081] 图2是本申请实施例二提供的语法纠错模型训练方法的流程示意图；

[0082] 图3是本申请实施例三中提供的语法纠错模型训练方法的流程示意图；

[0083] 图4是本申请实施例三中提供的语法纠错模型训练方法的模型示意图；

[0084] 图5是本申请实施例四中提供的语法纠错方法的流程示意图；

[0085] 图6是本申请实施例五提供的语法纠错模型的训练装置的模块示意图；

[0086] 图7是本申请实施例六提供的语法纠错装置的模块示意图；

[0087] 图8是本申请实施例七中提供的计算设备的结构示意图。

具体实施方式

[0088] 在下面的描述中阐述了很多具体细节以便于充分理解本申请。但是本申请能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况下做类似推广，因此本申请不受下面公开的具体实施的限制。

[0089] 在本说明书一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本说明书一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0090] 应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一也可以被称为第二，类似地，第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在.....时”或“当.....时”或“响应于确定”。

[0091] 首先，对本发明一个或多个实施例涉及的名词术语进行解释。

[0092] Transformer模型：其本质上是一个编码器(Encoder)-解码器(Decoder)的结构，编码器由6个编码层依次连接组成，解码器是6个解码层依次连接组成。与所有的生成模型相同的是，编码器接收原始输入的文本，并输出编码向量至解码器，解码器生成解码向量并得到最终的输出文本。

[0093] 集束搜索(beam search)：应用于机器翻译领域，在寻找最佳翻译结果(概率最大的结果)时常用的算法是集束搜索(beam search)。集束搜索有一个参数：集束宽(beam width)，表示在生成每个翻译结果时会考虑集束宽个候选结果。

[0094] 词单元(token)：对输入文本做任何实际处理前，都需要将其分割成诸如字、标点符号、数字或字母等语言单元，这些单元被称为词单元。对于英文文本，词单元可以是一个单词、一个标点符号、一个数字等，对于中文文本，最小的词单元可以是一个字、一个标点符号、一个数字等。

[0095] 句长：句子长度(词数)，是指句子中包含的词单元个数。

[0096] 数组(array)：数组的作用就是存储一组数据，数组中存放的每个元素类型必须相同，所以每个元素所占用的内存大小也一致。数组中的值通过数组名和下标组合起来进行访问。数组一经创建，其长度就不可改变，数组元素的有效下标范围为0～n-1。

[0097] 编码器(encoder)：将待翻译语句由文字转化为编码向量。

[0098] 解码器(decoder)：将编码向量生成解码向量，并将解码向量转换为答案。

[0099] 语言模型(Language Model)：是用来计算一个句子的概率的概率模型。利用语言模型，可以确定哪个词序列的可能性更大，或者给定若干个词，可以预测下一个最可能出现的词语。

[0100] 重排序(Rerank)：是指基于语言模型对多条候选语句进行概率值大小的排序。

[0101] 在本申请中，提供了一种语法纠错模型的训练方法及装置、语法纠错方法及装置、

计算设备及计算机可读存储介质，在下面的实施例中逐一进行详细说明。

[0102] 实施例一

[0103] 本实施例提供了一种纠错模型的训练方法，参见图1所示，所述训练方法包括步骤S101～步骤S105，下面将对每一步骤进行详细介绍。

[0104] S101、基于第一训练集进行数据扩展处理获得第二训练集。

[0105] 在本实施例中，所述第一训练集为已经存在的、数据较少的训练集。其分为source端(源语句端)和target端(目标语句端)。在第一训练集中，source端(源语句端)包括第一源样本语句，target端(目标语句端)包括第一目标样本语句。所述第一源样本语句为具有语法错误的语句，所述第一目标样本语句为与所述第一源样本语句一一对应的语法正确的目标样本语句。

[0106] 例如，在第一训练集中，第一源样本语句为“无论有没有事情，都来上学习”，其中“上学习”存在语法错误；与该第一源样本语句对应的第一目标样本语句为“无论有没有事情，都来学习”。

[0107] 具体地，在所述第一训练集中，第一源样本语句与第一目标样本语句以语句对的形式存在，即<第一源样本语句，第一目标样本语句>。在本领域技术人员进行语法纠错模型训练时，普遍将source端(源语句端)的第一源样本语句作为纠错模型输入，将target端(目标语句端)的第一目标样本语句作为纠错模型的输出标签。

[0108] 具体地，在第一训练集中包含的<第一源样本语句，第一目标样本语句>语句对的数量有限。

[0109] 进一步地，为了扩大训练集的数据数量，基于第一训练集进行数据扩展处理获得第二训练集。

[0110] 具体地，数据扩展处理的操作包括但不限于：对所述第一训练集中的source端(源语句端)包含的第一源样本语句进行语料腐化处理(例如对第一源样本语句随机进行词插入、词替换处理等)，或者还可以基于第一训练集中的语句对进行反向翻译，将反向翻译后得到的语句对第一训练集进行数据扩展。

[0111] 进一步地，在本实施例中，将数据扩展后得到的第二源样本语句以及与第二样本语句对应的第二目标样本语句共同构成第二训练集。

[0112] 具体地，在本实施例中，由于基于第一训练集进行数据扩展处理时，可以对同一个第一源样本语句分别进行词插入、词替换、以及反向翻译，即得到三个新语句，则该第一源样本语句对应的第一目标样本语句一共对应4个源样本语句，则将数据扩展后的四个源样本语句作为第二源样本语句，并且每一个第二源样本语句分别对应的相同的四个目标样本语句均第二样本语句。

[0113] 例如，第一训练集中原有的语句对为<我每天早上五点半起床后，我就学习了一些中文，我每天早上五点半起床后，就开始学一些中文>，然后通过对其中的第一源样本语句(记为w1)分别进行词插入、词替换、以及反向翻译处理，得到的新语句分别为：

[0114] w2、我每天早上五点半起床后，我就好学习了一些中文，其中“好”词单元为插入的词单元；

[0115] w3、我每天早上五点半起床后，我就学习了一些书文，将原来的“文”字采用“书”字进行了替代；

[0116] w4、我五点半每天起床早上后，学习了我就一些中文。

[0117] 然后将w1～w4分别与原第一目标样本语句(p1)进行组合，形成的语句对如下：

[0118] <w1,p1>、<w2,p1'>、<w3,p1">、<w4,p1"'>，其中p1'～p1"'与p1相同。

[0119] 则w1～w4为第二源样本语句，p1～p1"'为第二源样本语句对应的第二目标样本语句。

[0120] 进一步地，所述第二源样本语句与对应的第二目标样本语句构成第二训练集。

[0121] S102、基于所述第二训练集获取第二源样本语句以及第二目标样本语句。

[0122] 具体地，基于数据扩展后得到的第二训练获取语句对，所述语句对的形式为<第二源样本语句,第二目标样本语句>。

[0123] 进一步地，所述第二源样本语句作为语法纠错模型的输入，所述第二目标样本语句作为语法纠错模型的输出的目标样本标签语句。在本实施例中，将数据扩展后得到的第二训练集中语句对中的第二目标样本语句作为模型输出的标签语句。

[0124] S103、将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句。

[0125] 在本实施例中，所述语法纠错模型具体形式为Transformer神经网络模型。所述Transformer神经网络模型包括编码器(Encoder)和解码器(Decoder)两部分。

[0126] 具体地，在本实施例中，将第二训练集包括的语句对中的第二源样本语句输入至语法纠错模型中，经过模型的处理生成了经过纠错的纠错样本语句。

[0127] S104、基于所述纠错样本语句与所述第二目标样本语句确定损失值。

[0128] 具体地，将前述语言纠错模型生成的纠错样本语句与输入模型的第二源样本语句对应的第二目标样本语句(即标签语句)进行对比，通过损失函数计算损失值。

[0129] 在实际应用中，损失函数可以为如分类交叉熵、最大熵函数等，本申请对此不作限制。

[0130] S105、基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件。

[0131] 具体地，可以通过预先设定所述损失值的阈值作为需训练停止的条件。例如设定阈值为0.2。本申请对此不作限制。

[0132] 本实施例通过上述步骤S101～S105，提供了一种语法纠错模型的训练方法，其中，通过数据扩展处理，在已有的、数量有限的第一训练集的基础上生成了更多数据的第二训练集，可以有效扩充模型的训练数据，同时节省了大量人工标注所需要的劳动力，进一步提高了模型训练的效果。

[0133] 实施例二

[0134] 本实施例提供了一种语法纠错模型的训练方法，参见图2所示，具体包括步骤S201～步骤S207。

[0135] S201、基于第一训练集进行数据扩展处理获得第二训练集。

[0136] 在本实施例中，所述第一训练集包括第一源样本语句和第一目标样本语句。

[0137] 所述第一训练集为已经存在的、数据较少的训练集。其分为source端(源语句端)和target端(目标语句端)。在第一训练集中，source端(源语句端)包括第一源样本语句，target端(目标语句端)包括第一目标样本语句。所述第一源样本语句为具有语法错误的语句，所述第一目标样本语句为与所述第一源样本语句一一对应的语法正确的目标样本语

句。

[0138] 具体地，在所述第一训练集中，第一源样本语句与第一目标样本语句以语句对的形式存在，即<第一源样本语句，第一目标样本语句>。在本申请中，基于已有的第一训练集，进行语法纠错模型训练过程中，将source端(源语句端)的第一源样本语句作为纠错模型输入，将target端(目标语句端)的第一目标样本语句作为纠错模型的输出标签。

[0139] 进一步地，在本实施例中，首先对所述第一源样本语句和第一目标样本语句进行预处理。

[0140] 所述预处理具体为：对所述第一源样本语句及所述第一目标样本语句进行分词处理，将每个词单元之间进行分隔处理；

[0141] 去除所述第一训练集中句长大于预设阈值的语句；

[0142] 去除所述第一源样本语句和第一目标样本语句中相同的语句。

[0143] 进一步地，将所述第一源样本语句及所述第一目标样本语句中包括的所有语句中每个词单元之间采用空格彼此分隔开；然后将第一训练集中过长或者过短语句进行删除，例如词单元为30以上为过长句子，词单元为5个以下过短语句；并且，在第一训练集的第一源样本语句及第一目标样本语句中存在语句相同的情况，将相同的语句对进行删除。

[0144] 本实施例通过上述方法对已有的第一训练集进行预处理，可以避免过长或者过短句子对模型的训练效果产生不好的影响，提高模型训练的准确度；并且由于在源语句端以及目标语句端存在句子相同的情况，这样并不能达到对模型语法纠错能力的训练，还会增加模型训练过程中的负担，因此将第一训练集中相同的第一源样本语句及第一目标样本语句进行删除，可以使模型训练更准确。

[0145] 进一步地，在本实施例中，基于预处理后的第一训练集进行数据扩展处理得到第二训练集。

[0146] 具体地，所述数据扩展处理包括：腐化语料处理以及反向翻译处理，并且所述两种处理方法是完全独立的，在实际应用过程中可以交叉使用，例如先进行腐化语料处理，然后将腐化语料得到的语句对再进行反向翻译处理；也可以仅进行腐化语料处理或者仅进行反向翻译处理等。本申请对此不作限制。

[0147] 具体地，所述基于第一训练集进行数据扩展处理获得第二训练集，包括：

[0148] 对所述第一源样本语句和第一目标样本语句进行预处理；

[0149] 基于所述第一训练集中词单元的出现频率，对所述词单元进行权重赋值，构建词典；

[0150] 根据所述词典对所述第一训练集的源样本语句包含的语句进行腐化处理，获得数据扩展的第二源样本语句；根据所述第二源样本语句以及所述第二源样本语句对应的第二目标样本语句构建所述第二训练集。

[0151] 其中，所述对所述第一源样本语句和第一目标样本语句进行预处理已经在前述内容记载，在此便不再赘述。

[0152] 在本实施例中，将经过预处理后的第一训练集中包含的所有语句中出现的词单元构建词典，并按词单元在第一训练集中出现的频率进行权重赋值：出现次数越多，权重越大，即词单元的权重越大，该词单元被使用的概率越大。例如，“了”、“我”词单元在第一训练集中出现的频率比“竞”、“影”词单元出现的频率大，则“了”、“我”词单元的权重赋值大于

“竞”、“影”词单元的权重赋值。

[0153] 进一步地，所述第二训练集中语句对的存在形式为<第二源样本语句,第二目标样本语句>。具体的，在第二训练集中，存在若干相同的第二目标样本语句，并且相同的第二目标样本语句分别对应不同的第二源样本语句。

[0154] 进一步地，在本实施例中，所述根据所述词典对所述第一训练集的源样本语句包含的语句进行腐化处理，获得数据扩展的第二源样本语句，包括：

[0155] 根据所述词典对所述第一源样本语句进行词插入处理，获得数据扩展的第二源样本语句；和/或根据所述词典对所述第一源样本语句进行词替代处理，获得数据扩展的第二源样本语句。

[0156] 进一步地，所述根据所述词典对所述第一源样本语句进行词插入处理，获得数据扩展的第二源样本语句，包括：

[0157] a1、获取所述第一源样本语句以及第一源样本语句的句长n；

[0158] a2、基于所述第一源样本语句的句长n生成对应的第一数组；

[0159] 其中，所述第一数组中每个数值均为随机生成的(0,1)范围内的数值；

[0160] 且所述第一数组中每个数值均具有与该数值在所述第一数组中位置顺序对应的下标i，所述下标i的取值范围是(0,n-1)范围内的整数；

[0161] a3、根据预设的第一阈值，获取所述第一数组中小于所述第一阈值的数值对应的下标i；

[0162] a4、基于权重随机选择所述词典中的一个词单元，插入所述第一源样本语句中的第i位置，生成词插入处理后数据扩展的第二源样本语句。

[0163] 进一步地，所述根据所述词典对所述第一源样本语句进行词替代处理，获得数据扩展的第二源样本语句，包括：

[0164] b1、获取所述第一源样本语句以及第一源样本语句的句长n；

[0165] b2、基于所述第一源样本语句的句长生成对应的第二数组，其中，所述第二数组中每个数值均为随机生成的(0,1)范围内的数值；

[0166] 且所述第二数组中每个数值均具有与该数值在所述第二数组中位置顺序对应的下标i，所述下标i的取值范围是(0,n-1)范围内的整数；

[0167] b3、根据预设的第二阈值，获取所述第二数组中小于所述第二阈值的数值对应的下标i；

[0168] b4、基于权重随机选择所述词典中的一个词单元，将所述第一源样本语句中第i位置的词单元替换为所述随机选择的词单元，生成词替代处理后数据扩展的第二源样本语句。

[0169] 进一步地，在本实施例中，上述词插入以及词替代处理中，所述“基于权重随机选择所述词典中的一个词单元”具体是指：在随机选择词单元的过程中，词典中权重越大的词单元越容易被选中，例如权重0.9的词单元比权重0.1的词单元更容易被选中。

[0170] 例如，第一源样本语句为：我水平不高，但是我要跟别的学生竞争(记为m_0)；

[0171] 第一目标样本语句为虽然我水平不高，但是我要跟别的学生竞争(记为p_0)。

[0172] 在进行词插入处理时，其中第一源样本语句句长为17，然后基于句长随机生成了第一数组(0.5,0.5,0.1,0.7,0.5,....0.6)，其中数组中数值的下标为0～16，预设的阈值

为0.3,所生成的第一数组中“0.1”小于第一阈值,则“0.1”数值下标i为2,则从构建的词典中基于权重随机选择出一个词单元“不”,则将该词单元插入到原第一源样本语句的第2位置,生成词插入处理后数据扩展的第二源样本语句为“我不水平不高,但是我要跟别的学生竞争”(即为m_1)。

[0173] 进一步地,在本实施例中,当基于第一源样本语句生成的第一数组中包括多个小于预设阈值的数值,则可以基于所有小于预设第一阈值数值的下标,将从词典中选取的词单元同时插入第一源样本语句,也可以分批插入,或者一个一个插入。如果是分批插入,或者一个一个插入,则每插入一批词单元或一个词单元,则再进行新的词单元插入时,第i位置随已经插入的词单元进行更新。

[0174] 在进行词替代处理时,其中第一源样本语句句长为17,然后基于句长随机生成了第一数组(0.5,0.5,0.1,0.7,0.5,....0.6),其中数组中数值的下标为0～16,预设的阈值为0.3,所生成的第一数组中“0.1”小于第一阈值,则“0.1”数值下标i为2,则从构建的词典中基于权重随机选择出一个词单元“你”,则将原第一源样本语句中第2位置的词单元替换为“你”,生成词替换处理后数据扩展的第二源样本语句为“我不平不高,但是我要跟别的学生竞争”(即为m_2)。

[0175] 在本实施例中通过上述词插入、词替换处理,得到了数据扩展的第二源样本语句m_1和m_2,并且m_1和m_2分别对应的第二目标样本语句与p_0相同。

[0176] 在本实施例提供的语法纠错模型训练方法过程中,在构建词典过程中,基于词单元出现的频率对词单元进行权重赋值,并且在腐化处理进行词插入和词替代过程中,在词典中基于词单元的权重随机选择进行插入或者替换的词单元,词典中权重大的更容易被选择到,符合语法错误的客观规律,即越常出现的词,越容易在语法犯错的过程中被使用,进一步提高了模型训练效果。

[0177] 进一步地,所述基于第一训练集进行数据扩展处理获得第二训练集,还包括:

[0178] c1、对所述第一源样本语句和第一目标样本语句进行预处理;

[0179] c2、基于所述第一源样本语句和第一目标样本语句,构建<第一目标样本语句,第一源样本语句>形式的反向训练集;

[0180] c3、基于所述反向训练集对所述语法纠错模型进行反向训练,其中,将所述第一目标样本语句作为所述语法纠错模型的输入,将所述第一源样本语句作为所述语法纠错模型的目标输出,进行预设代数的训练后,将所述语法纠错模型的参数固定;

[0181] c4、将所述反向训练集中的第一目标样本语句输入所述参数固定的语法纠错模型,通过集束搜索生成预设数量的候选纠错语句;

[0182] c5、对所述预设数量的候选纠错语句进行重排序,选取预设顺序的候选纠错语句作为第二源样本语句;

[0183] c6、根据所述第二源样本语句以及所述第二源样本语句对应第二目标样本语句构建所述第二训练集。

[0184] 其中,对所述第一源样本语句和第一目标样本语句进行预处理已在前述内容详细叙述。

[0185] 具体地,步骤c5中对对所述预设数量的候选纠错语句进行重排序采用的是已经训练好的语言模型(LM)。

[0186] 具体地，所述语言模型为：对于任意的词序列，语言模型能够计算出这个序列是一句话的概率。例如，词序列A：“今天|的|天气|真|好|啊”，可以看出词序列A明显是一句话，如果采用一个训练比较好的语言模型，其对于词序列A会给出很高的概率；又比如词序列B：“今天|的|水果|学习|不如|”，可以明显看出词序列B不是一句话，如果语言模型训练得好，则其对于序列B给出的概率就极小。

[0187] 进一步地，假设为中文创建一个语言模型，V表示词典，V＝{太阳，日出，月亮，乌云，的，人类，...}。在实际应用中，V的维度非常高，可达到几万维、十几万维。

[0188] 另外有一个句子，是由单词组成的表示为：$w_1w_2w_3...w_n$，其中w_i属于词典V。

[0189] 则语言模型的作用为：给定词典V，能够计算出任意单词序列是一句话的概率$p(w_1w_2w_3...w_n)$，其中，$p\geq 0$。因此根据语言模型可以对每一给定的词序列计算出对应的$p(w_1w_2w_3...w_n)$，然后基于概率p对多个词序列进行重排序。

[0190] 进一步地，语言模型用数据学习出$p(w_1w_2w_3...w_n)$的最简单的方法是“数数”，具体为：假设训练集中共有N个句子，语言模型可以数出在训练集中$(w_1w_2w_3...w_n)$出现的次数，假设为n，则$p(w_1w_2w_3...w_n)=n/N$。但是该方法的预测能力几乎为0，一旦单词序列没有在训练集中出现过，模型的输出概率就是0，显然相当不合理。

[0191] 因此，为了更合理地学习出$p(w_1w_2w_3...w_n)$，可以采用n-gram语言模型。所述n-gram语言模型原理为：根据链式法则(chain rule)把p展开：

[0192] $$p(w_1w_2w_3...w_n)=p(w_i)\prod_{i=2}^{n}p(w_i|w_{1,...,}w_{i-1})；$$

[0193] 进一步地，为了简化后验概率$p(w_i|w_1,...,w_{i-1})$的计算，引入一阶马尔可夫假设(first-order Markov assumption)：每个词只依赖前一个词；

[0194] 则$p(w_i|w_1,...,w_{i-1})=p(w_i|w_{i-1})$；

[0195] 此时，$$p(w_1w_2w_3...w_n)=p(w_1)\prod_{i=2}^{n}p(w_i|w_{1,...,}w_{i-1})=p(w_1)\prod_{i=2}^{n}p(w_i|w_{i-1})。$$

[0196] 进一步地，也可以引入二阶马尔可夫假设：每个词依赖前两个词，具体原理引入一阶马尔可夫假设相似，在此不再赘述。最终得到的$p(w_1w_2w_3...w_n)=$

$$p(w_1)p(w_2|w_1)\prod_{i=3}^{n}p(w_i|w_{i-2,}w_{i-1})$$

[0197] 进一步地，$$p(w_i|w_{i-2,}w_{i-1})=\frac{count(w_{i-2,}w_{i-1,}w_i)}{count(w_{i-2,}w_{i-1})}；$$

[0198] 其中，count(*)表示在训练集中出现的次数。需要注意的是，由于n-gram的参数太多，有许多$|V|^n$，实际上很多参数并没有在训练集中出现过，也就是$count(w_{i-N+1},....,w_{i-1},w_i)=0$，导致模型在进行预测时，很多句子的概率都是0，为了避免这种情况发生，需要对count(*)＝0的情况做一些平滑处理，最简单的方法是所有词组出现次数加1。

[0199] 上述内容就是n-gram语言模型的具体原理。

[0200] 综上所述，通过上述反向翻译的过程，可以基于已经存在的第一训练集中的第一目标样本语句生成数据扩展的第二源样本语句，进一步扩大了训练集。

[0201] 例如，将第一训练集中的语句对$<m_1,p_1>$的第一源样本语句和第一目标样本语句调

换，生成反向训练集<p1,m1>，所述反向训练集包括多个形式为<p_1,m_1>的语句对；

[0202] 然后将所述反向训练集中的第一目标样本语句p_1输入至语法纠错模型进行反向翻译训练，并以第一源样本语句m_1作为反向训练的输出目标标签，经过一定的代数训练后，将反向翻译的语法纠错模型的参数固定；

[0203] 再将反向训练集语句对中的第一源样本语句m_1输入至参数固定的反向语法纠错模型中，通过集束搜索生成预设数量的候选纠错语句，例如设置集束大小为12，则可以生成与输入的第一目标样本语句对应的12条候选纠错语句；

[0204] 对于生成的12条候选纠错语句采用已经训练好的语言模型进行排序，然后选取预设顺序的候选纠错语句作为第二源样本语句，例如选择排序第4的候选纠错语句作为第二源样本语句。

[0205] 进一步地，在本实施例中根据候选纠错语句的排序可以选择一条也可以选择多条，本申请对此不作限制。

[0206] 进一步地，根据所述第二源样本语句以及所述第二源样本语句对应第二目标样本语句构建所述第二训练集。其中所述第二目标样本语句为反向翻译过程中进行输入的语句。

[0207] 进一步地，通过前述腐化语料处理(词插入、词替代)以及反向翻译处理对已有的第一训练集的数据进行扩展，得到了第二训练集。

[0208] 在本实施例中，通过数据扩展处理，基于第一训练集可自动生成第二训练集，不仅扩大了训练集，并且不需要额外的人工进行标注节省时间和人力成本；并且在进行腐化语料处理时，基于词典中词单元的权重进行随机选择，更贴近实际生活中的犯错习惯，因此可以进一步提高模型训练的准确度和置信度。

[0209] S202、基于所述第二训练集获取第二源样本语句以及第二目标样本语句。

[0210] 具体地，基于数据扩展后得到的第二训练集获取语句对，所述语句对的形式为<第二源样本语句,第二目标样本语句>。

[0211] 进一步地，所述第二源样本语句作为语法纠错模型的样本语句，所述第二目标样本语句作为语法纠错模型的标签语句。

[0212] S203、将所述第二源样本语句输入至所述语法纠错模型的编码器进行编码，生成编码向量。

[0213] 具体地，首先将第二源样本语句输入至所述语法纠错解题模型的嵌入层，生成嵌入向量；

[0214] 将所述嵌入向量输入至所述语法纠错模型的编码器进行编码，生成编码向量。

[0215] S204、将所述编码向量输入至所述语法纠错模型的解码器进行解码得到预设数量的候选纠错样本语句。

[0216] S205、对所述预设数量的候选纠错样本语句进行重排序。

[0217] 具体地，采用已经训练好的语言模型对所述预设数量的候选纠错样本语句进行重排序。

[0218] 其中，对于语言模型进行重排序的过程，在前述内容已经详述，在此便不再赘述。

[0219] S206、根据所述重排序结果，将分值最高的候选纠错样本语句作为纠错样本语句。

[0220] 例如，将第二源样本语句“你知道去什么时候上课吗?”通过语法纠错模型的嵌入

层(embedding层)进行嵌入生成嵌入向量,然后输入至语法纠错模型的编码器(Encoder)生成编码向量;之后将编码向量输入至语法纠错模型的解码器(Decoder),生成预设数量的候选纠错样本语句,例如生成12条候选纠错样本语句;然后采用训练好的语言模型对所述12条候选纠错样本语句进行重排序,并根据重排序结果,将分值最高的候选纠错样本语句作为纠错样本语句,例如12条候选纠错样本语句中第5条的分值最高,则将第5条候选纠错样本语句作为语法纠错模型的纠错样本语句。

[0221] S207、基于所述损失值对所述语法纠错模型进行迭代训练,直至达到训练停止条件。

[0222] 具体为:判断所述损失值是否小于预设阈值;

[0223] 若否,则继续获取待处理样本语句以及标签语句进行训练;

[0224] 若是,则停止训练。

[0225] 例如预设值为0.2,则当损失值小于0.2时,停止训练。

[0226] 本实施例提供了一种语法纠错模型的训练方法,通过对已有的训练集进行数据增强处理,达到对训练集进行自动扩充的目的,有效地减少了人工劳动。

[0227] 本申请在进行数据增强处理过程中,将腐化语料处理以及反向翻译处理结合起来进行应用,有效地扩大了已有训练集的数据。

[0228] 并且本申请在进行腐化语料处理时,在词插入以及词替换过程中,对词典中的词单元进行权重赋值,通过基于词典中词单元的权重大小进行随机选择插入或者替换的词单元,词典中权重大的词单元更容易被选中,相比于直接随机选择一个词单元,基于权重进行选择,更符合语法错误客观规律,进一步增加模型训练的准确性。

[0229] 实施例三

[0230] 本实施例提供了一种语法纠错模型训练方法,参见图3所示,包括以下步骤:

[0231] S301、数据预处理。

[0232] 具体地,对已有的第一训练集中的第一源样本语句和第一目标样本语句进行预处理,包括:

[0233] 对所述第一源样本语句及所述第一目标样本语句进行分词处理,将每个词单元之间进行分隔处理;

[0234] 去除所述第一训练集中句长大于预设阈值的语句;

[0235] 去除所述第一源样本语句和第一目标样本语句中相同的语句。

[0236] 进一步地,将所述第一源样本语句及所述第一目标样本语句中包括的所有语句中每个词单元之间采用空格彼此分隔开;然后将第一训练集中过长或者过短语句进行删除,例如词单元为35以上为过长句子,词单元为5个以下过短语句;并且,在第一训练集的第一源样本语句及第一目标样本语句中存在语句相同的情况,将相同的语句对进行删除。

[0237] S302、数据增强。

[0238] 具体地,基于预处理后的第一训练集进行数据扩展处理获得第二训练集。

[0239] 在本实施例中,所述数据增强包括以下方式:反向翻译、腐化语料。

[0240] 进一步地,所述腐化语料包括:词插入、词替代。

[0241] 进一步地,在本实施例提供的语法纠错模型的训练方法中,腐化语料还可能包括词交换等操作。

[0242] 具体地，上述两种处理反向翻译、腐化语料两种方式是完全独立的，在实际应用过程中可以交叉使用，例如先进行腐化语料处理，然后将腐化语料得到的语句对再进行反向翻译处理；也可以仅进行腐化语料处理或者仅进行反向翻译处理等。本申请对此不作限制。

[0243] 通过上述数据增强，本实施例在已有的第一训练集的基础上获得了第二训练集。

[0244] S303、基于Transformer的中文语法纠错模型生成候选纠错样本语句。

[0245] 基于上述得到的第二训练集，采用其中的第二源样本语句输入至中文语法纠错模型，生成候选纠错样本语句。

[0246] 进一步地，所述中文语法纠错模型的基本结构是Transformer模型结构。

[0247] 所述基于Transformer的中文语法纠错模型生成候选纠错样本语句过程如图4虚线框内的结构所示。

[0248] 具体为，首先，将第二源样本语句“你很喜欢哪部电影?”通过嵌入层(Embedding层)进行嵌入，生成嵌入向量；

[0249] 然后将所述嵌入向量输入至编码器(Encoder)中进行编码，生成编码向量；

[0250] 然后将所述编码向量输入至解码器(Decoder)进行解码，生成预设条数的候选纠错样本语句，例如对应于输入的第二源样本语句“你很喜欢哪部电影?”生成10条候选纠错样本语句p_1-p_{10}。

[0251] S304、语言模型进行重排序。

[0252] 如图4中虚线框外的结构所示，采用已经训练好的语言模型对上述候选纠错样本语句p_1-p_{10}进行重排序(Rerank)；其中，对于语言模型进行重排序的过程，在前述内容已经详述，在此便不再赘述。

[0253] 最后基于重排序的结果，选取p_1-p_{10}中分值最高的作为纠错样本语句。例如在p_1-p_{10}中p_7的分值最高，即以p_7作为纠错样本语句。

[0254] S305、迭代训练。

[0255] 将上述获得的纠错样本语句(如p_7)与步骤S303中第二源样本语句对应的第二目标样本语句进行损失值计算；

[0256] 然后基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件。

[0257] 本实施例通过上述步骤提供了一种语法纠错模型的训练方法，通过对已有的训练集进行数据增强处理，达到对训练集进行自动扩充的目的，有效地减少了人工劳动。本申请在进行数据增强处理过程中，将腐化语料处理以及反向翻译处理结合起来进行应用，有效地扩大了已有训练集的数据。

[0258] 实施例四

[0259] 基于前述提供的训练方法得到的语法纠错模型，本实施例提供了一种语法纠错的方法，参见图5所示，包括下述步骤：

[0260] S501、获取源语句。

[0261] 具体地，源语句的获取方式包括：从各种论坛网站获取、从学生提交的语文试卷的作文中获取等。本申请对此不作限制。

[0262] S502、将所述源语句输入至语法纠错模型，生成语法纠错语句。

[0263] 进一步地，所述将所述源语句输入至语法纠错模型，生成语法纠错语句，包括：

[0264] 将所述源语句输入至所述语法纠错模型的编码器进行编码,生成编码向量;

[0265] 将所述编码向量输入至所述语法纠错模型的解码器进行解码,生成所述语法纠错语句。

[0266] 进一步地,所述将所述源语句输入至语法纠错模型,生成语法纠错语句,包括:

[0267] 将所述源语句输入至语法纠错模型的嵌入层进行嵌入,生成所述源语句的嵌入向量;

[0268] 再将所述嵌入向量输入至语法纠错模型的编码器中进行编码,生成编码向量;

[0269] 将所述编码向量输入至语法纠错模型的解码器中进行解码,得到语法纠错语句。

[0270] 具体地,在本申请中采用的语法纠错模型的基本结构是Transformer模型结构。

[0271] 对于所述Transformer模型包括嵌入层、编码器以及解码器。

[0272] (1)对源语句进行嵌入层处理,更具体的是将源语句进行切分得到多个词单元,然后对每个词单元进行词嵌入处理,最后得到每个词单元的词向量。

[0273] 词嵌入实际上是一种将各个词单元在预定的向量空间中表示为实值向量的一类技术。每个词单元被映射成一个向量(初始随机化)。

[0274] 使用嵌入层通常步骤一般是先预处理源语句,将每个词单元转化成one-hot形式的编码。而此词单元对应的词向量其实是算法模型的其中一部分,词向量用预定义的维度来表示,大小随机初始化。在这里,嵌入层其实就是语法纠错模型的输入层。

[0275] (2)具体地,所述Transformer模型的编码器共包括六个编码层。

[0276] 具体地,编码器中的每个编码层包括1个多头注意力层(multi-head self-attention)和1个全连接层(fully connected feed-forward network,FFN)。

[0277] 对于输入至多头注意力层的编码向量,每个词单元对应有3个不同的向量,分别是词向量Q(Query)、K(Key)、V(Value)。多头注意力层的计算是通过h个不同的线性变换对词向量Q,K,V进行投影,最后将不同的attention结果拼接起来。

[0278] 其中,在编码器的注意力计算中,词向量Q(Query)、K(Key)、V(Value)都彼此相等,他们是上一个编码层输出的第一编码向量。对于第一个编码层,词向量Q、K、V是嵌入层(word embedding)输出的向量乘以权重矩阵得到。

[0279] 具体地,多头注意力层的计算公式如下:

[0280] $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ (1)

[0281] $head1,...,headh)W^O$ (2)

[0282] 其中,Q、K、V为输入的编码向量对应的词向量;

[0283] $head_i$为多头注意力层的每个头部(head)的自注意力结果;

[0284] Multihead为多头注意力层的输出结果;

[0285] Concat为拼接函数;

[0286] W_i^Q、W_i^K、W_i^V为每次词向量Q、K、V进行线性变换的权值矩阵,例如,每个词单元对应三个不同的词向量Q、K、V,均为64维,他们通过3个不同的权值矩阵由嵌入向量乘以三个不同的权值矩阵W_i^Q、W_i^K、W_i^V而得到,三个矩阵均为512*64维。

[0287] W^O为线性变换需要的权值矩阵,大小为512*512维。

[0288] 然后,多头注意力层的输出结果会输入至全连接层(FFN)。全连接层(FFN)的计算公式如下:

[0289] x＝max(0,xH1+b1)H2+b2 (3)

[0290] 其中，H_1、H_2为参数矩阵，训练得到；

[0291] b_1、b_2为参数，训练得到；

[0292] x为多头注意力层的输出结果；

[0293] FFN(x)为全连接层的输出结果。

[0294] 经过全连接层的输出结果，得到第j个所述编码层输出的编码向量。

[0295] (3)具体地，所述Transformer模型的解码器共包括六个解码层。

[0296] 对于每个解码层，包括三个层，第一层是掩盖多头注意力层(masked multi-head self-attention)，第二层是多头注意力层(multi-head self-attention)，第三层是一个前馈层(feed-forward network)。其中，对于多头注意力层和前馈层，在前述编码层生成编码向量的过程中已经详细介绍，在此便不再赘述。

[0297] 需要说明的是，在解码层的自注意力计算中，词向量Q、K、V的维度相等，对于第一个解码层，词向量Q对应于输入解码器中的参考向量，词向量K和V来自于编码器输出的源语句对应的编码向量；对于除去第一个解码层的其他解码层，词向量Q来自于上一个解码层输出的解码向量，词向量K和V来自于编码器输出的源语句对应的编码向量。

[0298] 由上述解码层的结构和编码层的结构对比可知，解码层比编码层多了一个掩盖多头注意力层(masked multi-head self-attention)。掩盖多头注意力层与多头注意力层的运算基本一致，所不同的是增加了mask的操作。mask的作用就是防止在模型在执行语法纠错任务的过程中使用未来的输出的单词。第一个单词是不能参考第二个单词的生成结果的。Mask就会把这个信息变成0，用来保证每个位置i的输出，只会依赖于i位之前(不包括i位，因为右移一位和mask)。

[0299] 进一步地，所述语法纠错模型通过前述实施例提供的训练方法训练得到。所述语法纠错模型的训练过程在前述实施例中已经进行了详细说明，在本实施例中便不再赘述。

[0300] 本实施例提供了一种语法纠错的方法，利用训练好的语法纠错模型可以实现对于源语句的语法纠错，提高了语法纠错准确度。

[0301] 实施例五

[0302] 本实施例提供了一种语法纠错模型的训练装置，参见图6所示，包括以下模块：

[0303] 数据扩展模块610，被配置为基于第一训练集进行数据扩展处理获得第二训练集。

[0304] 其中，所述第一训练集包括第一源样本语句和第一目标样本语句。

[0305] 具体地，所述数据扩展模块610包括：语料腐化单元以及反向翻译单元。

[0306] 所述语料腐化单元被配置为：

[0307] 对所述第一源样本语句和第一目标样本语句进行预处理；

[0308] 基于所述第一训练集中词单元的出现频率，对所述词单元进行权重赋值，构建词典；

[0309] 根据所述词典对所述第一训练集的源样本语句包含的语句进行腐化处理，获得数据扩展的第二源样本语句；根据所述第二源样本语句以及所述第二源样本语句对应的第二目标样本语句构建所述第二训练集。

[0310] 所述语料腐化单元包括词插入处理子单元和/或词替代处理子单元。

[0311] 所述词插入处理子单元被配置为：根据所述词典对所述第一源样本语句进行词插

入处理，获得数据扩展的第二源样本语句。

[0312] 所述词插入处理子单元进一步被配置为：

[0313] 获取所述第一源样本语句以及第一源样本语句的句长n；

[0314] 基于所述第一源样本语句的句长n生成对应的第一数组；

[0315] 其中，所述第一数组中每个数值均为随机生成的(0,1)范围内的数值；

[0316] 且所述第一数组中每个数值均具有与该数值在所述第一数组中位置顺序对应的下标i，所述下标i的取值范围是(0,n-1)范围内的整数；

[0317] 根据预设的第一阈值，获取所述第一数组中小于所述第一阈值的数值对应的下标i；

[0318] 随机选择所述词典中的一个词单元，插入所述第一源样本语句中的第i位置，生成词插入处理后数据扩展的第二源样本语句。

[0319] 所述词替代处理子单元被配置为：根据所述词典对所述第一源样本语句进行词替代处理，获得数据扩展的第二源样本语句。

[0320] 所述词替代处理子单元进一步被配置为：

[0321] 获取所述第一源样本语句以及第一源样本语句的句长n；

[0322] 基于所述第一源样本语句的句长生成对应的第二数组，其中，所述第二数组中每个数值均为随机生成的(0,1)范围内的数值；

[0323] 且所述第二数组中每个数值均具有与该数值在所述第二数组中位置顺序对应的下标i，所述下标i的取值范围是(0,n-1)范围内的整数；

[0324] 根据预设的第二阈值，获取所述第二数组中小于所述第二阈值的数值对应的下标i；

[0325] 随机选择所述词典中的一个词单元，将所述第一源样本语句中第i位置的词单元替换为所述随机选择的词单元，生成词替代处理后数据扩展的第二源样本语句。

[0326] 所述反向翻译单元被配置为：

[0327] 对所述第一源样本语句和第一目标样本语句进行预处理；

[0328] 基于所述第一源样本语句和第一目标样本语句，构建<第一目标样本语句，第一源样本语句>形式的反向训练集；

[0329] 基于所述反向训练集对所述语法纠错模型进行反向训练，其中，将所述第一目标样本语句作为所述语法纠错模型的输入，将所述第一源样本语句作为所述语法纠错模型的目标输出，进行预设代数的训练后，将所述语法纠错模型的参数固定；

[0330] 将所述反向训练集中的第一目标样本语句输入所述参数固定的语法纠错模型，通过集束搜索生成预设数量的候选纠错语句；

[0331] 对所述预设数量的候选纠错语句进行重排序，选取预设顺序的候选纠错语句作为第二源样本语句；

[0332] 根据所述第二源样本语句以及所述第二源样本语句对应第二目标样本语句构建所述第二训练集。

[0333] 所述语料腐化单元以及所述反向翻译单元进一步被配置为：

[0334] 对所述第一源样本语句及所述第一目标样本语句进行分词处理，将每个词单元之间进行分隔处理；

[0335] 去除所述第一训练集中句长大于预设阈值的语句；

[0336] 去除所述第一源样本语句和第一目标样本语句中相同的语句。

[0337] 获取模块620，被配置为：基于所述第二训练集获取第二源样本语句以及第二目标样本语句。

[0338] 语法纠错模块630，被配置为将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句。

[0339] 所述语法纠错模块630进一步被配置为：

[0340] 将所述第二源样本语句输入至所述语法纠错模型的编码器进行编码，生成编码向量；

[0341] 将所述编码向量输入至所述语法纠错模型的解码器进行解码得到预设数量的候选纠错样本语句；

[0342] 对所述预设数量的候选纠错样本语句进行重排序；

[0343] 根据所述重排序结果，将分值最高的候选纠错样本语句作为纠错样本语句。

[0344] 损失确定模块640，被配置为基于所述纠错样本语句与所述第二目标样本语句确定损失值。

[0345] 迭代训练模块650，被配置为基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件。

[0346] 本实施例提供的一种语法纠错模型的训练装置，通过对已有的训练集进行数据增强处理，达到对训练集进行自动扩充的目的，有效地减少了人工劳动，并且在本申请提供的训练装置中，将腐化语料单元以及反向翻译单元结合起来进行应用，有效地扩大了训练装置中训练集的数据；并且在训练装置中的词插入处理字单元以及词替代处理子单元中，基于词典中词单元的权重大小进行随机选择插入或者替换的词单元，词典中权重大的词单元更容易被选中，相比于直接随机选择一个词单元，基于权重进行选择词单元进行插入或替代，更加符合语法错误的客观规律，进一步增加训练装置的准确性。

[0347] 实施例六

[0348] 本实施例提供了一种语法纠错的装置，参见图7所示，包括下述模块：

[0349] 获取模块710，被配置为获取源语句；

[0350] 语法纠错模块720，被配置为将所述源语句输入至语法纠错模型，生成语法纠错语句；

[0351] 其中，所述语法纠错模型是通过前述实施例提供的训练方法训练得到的。本实施例不再赘述。

[0352] 所述语法纠错模块720，进一步被配置为：

[0353] 将所述源语句输入至所述语法纠错模型的编码器进行编码，生成编码向量；

[0354] 将所述编码向量输入至所述语法纠错模型的解码器进行解码，生成所述语法纠错语句。

[0355] 本实施例提供了一种语法纠错模型，利用训练好的语法纠错模型可以实现对于源语句的语法纠错，提高了语法纠错准确度。

[0356] 实施例七

[0357] 本实施例还提供了一种计算设备800，参见图8所示。

[0358] 图8是示出了根据本说明书一实施例的计算设备800的结构框图。该计算设备800的部件包括但不限于存储器810和处理器820。处理器820与存储器810通过总线830相连接，数据库850用于保存数据。

[0359] 计算设备800还包括接入设备840，接入设备840使得计算设备800能够经由一个或多个网络860通信。这些网络的示例包括公用交换电话网(PSTN)、局域网(LAN)、广域网(WAN)、个域网(PAN)或诸如因特网的通信网络的组合。接入设备840可以包括有线或无线的任何类型的网络接口(例如，网络接口卡(NIC))中的一个或多个，诸如IEEE802.11无线局域网(WLAN)无线接口、全球微波互联接入(Wi-MAX)接口、以太网接口、通用串行总线(USB)接口、蜂窝网络接口、蓝牙接口、近场通信(NFC)接口，等等。

[0360] 在本说明书的一个实施例中，计算设备800的上述部件以及图8中未示出的其他部件也可以彼此相连接，例如通过总线。应当理解，图1所示的计算设备结构框图仅仅是出于示例的目的，而不是对本说明书范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

[0361] 计算设备800可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备(例如，平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等)、移动电话(例如，智能手机)、可佩戴的计算设备(例如，智能手表、智能眼镜等)或其他类型的移动设备，或者诸如台式计算机或PC的静止计算设备。计算设备100还可以是移动式或静止式的服务器。

[0362] 其中，处理器820可以执行前述实施例提供的语法纠错模型训练方法的步骤或语法纠错的方法的步骤。具体步骤在本实施例不再赘述。

[0363] 本申请一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现如前所述语法纠错的训练方法或语法纠错方法的步骤。

[0364] 上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的语法纠错的训练方法或语法纠错方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述语法纠错的训练方法或语法纠错方法的技术方案的描述。

[0365] 所述计算机指令包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括电载波信号和电信信号。

[0366] 需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定都是本申请所必须的。

[0367] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部

分，可以参见其它实施例的相关描述。

[0368] 以上公开的本申请优选实施例只是用于帮助阐述本申请。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书的内容，可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本申请的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本申请。本申请仅受权利要求书及其全部范围和等效物的限制。

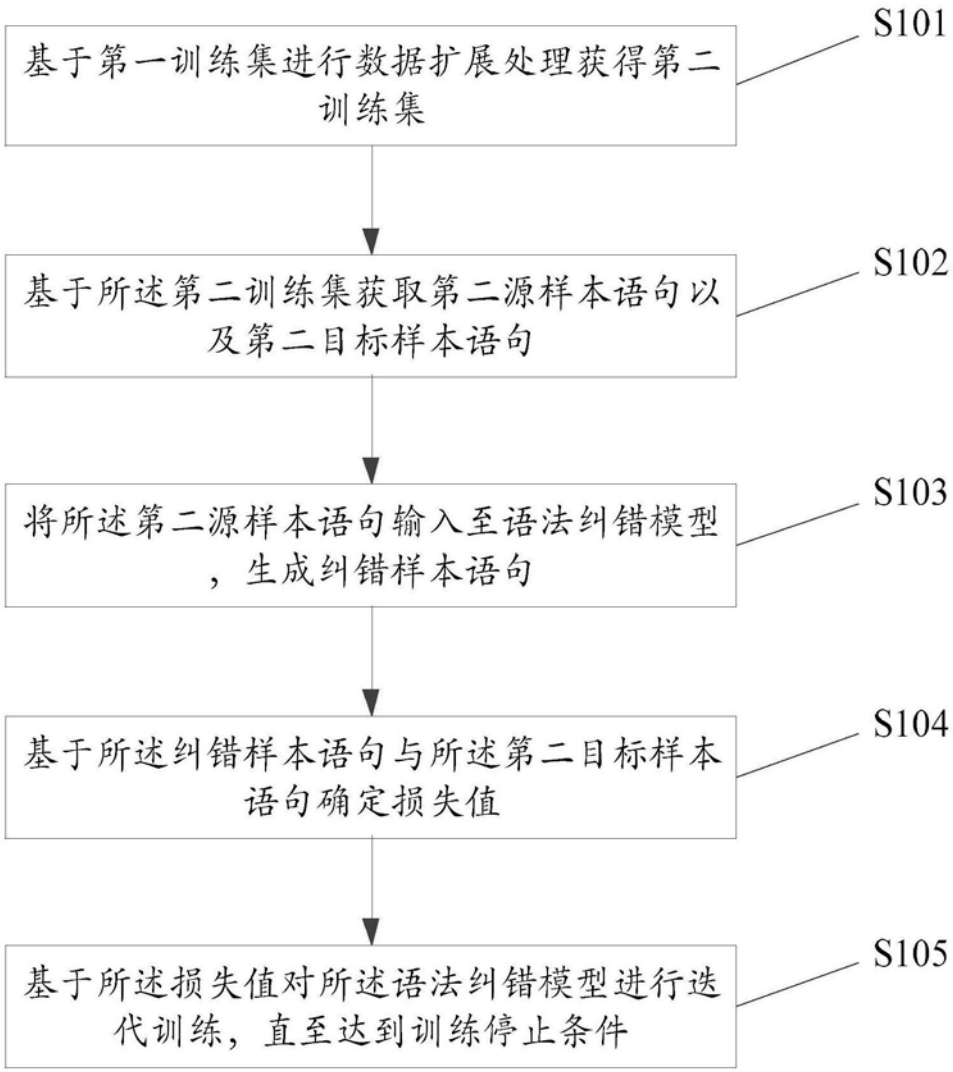

基于第一训练集进行数据扩展处理获得第二训练集
S101
基于所述第二训练集获取第二源样本语句以及第二目标样本语句
S102
将所述第二源样本语句输入至语法纠错模型，生成纠错样本语句
S103
基于所述纠错样本语句与所述第二目标样本语句确定损失值
S104
基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件
S105

图1

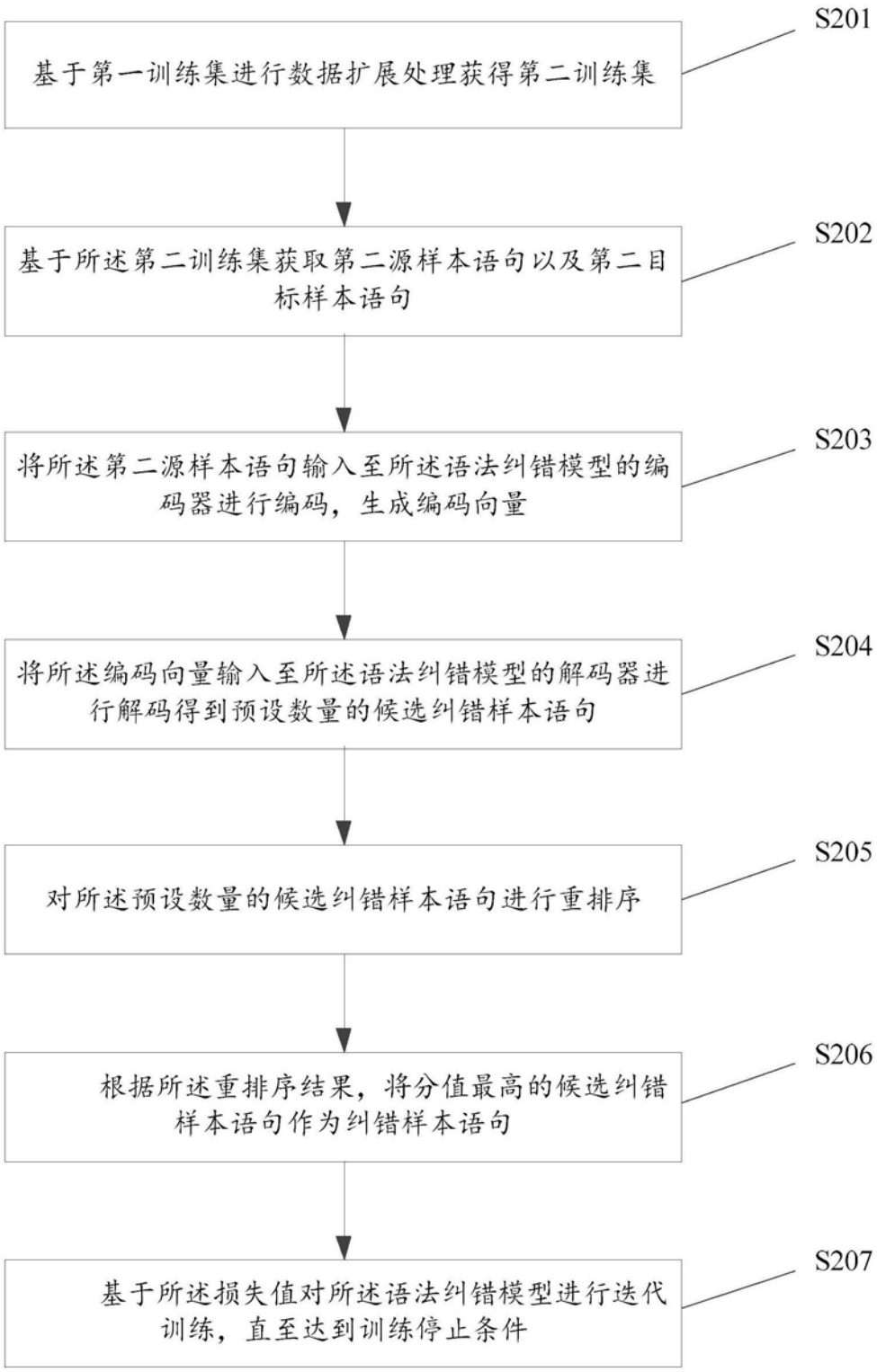

S201
基于第一训练集进行数据扩展处理获得第二训练集
S202
基于所述第二训练集获取第二源样本语句以及第二目标样本语句
S203
将所述第二源样本语句输入至所述语法纠错模型的编码器进行编码，生成编码向量
S204
将所述编码向量输入至所述语法纠错模型的解码器进行解码得到预设数量的候选纠错样本语句
S205
对所述预设数量的候选纠错样本语句进行重排序
S206
根据所述重排序结果，将分值最高的候选纠错样本语句作为纠错样本语句
S207
基于所述损失值对所述语法纠错模型进行迭代训练，直至达到训练停止条件

图2

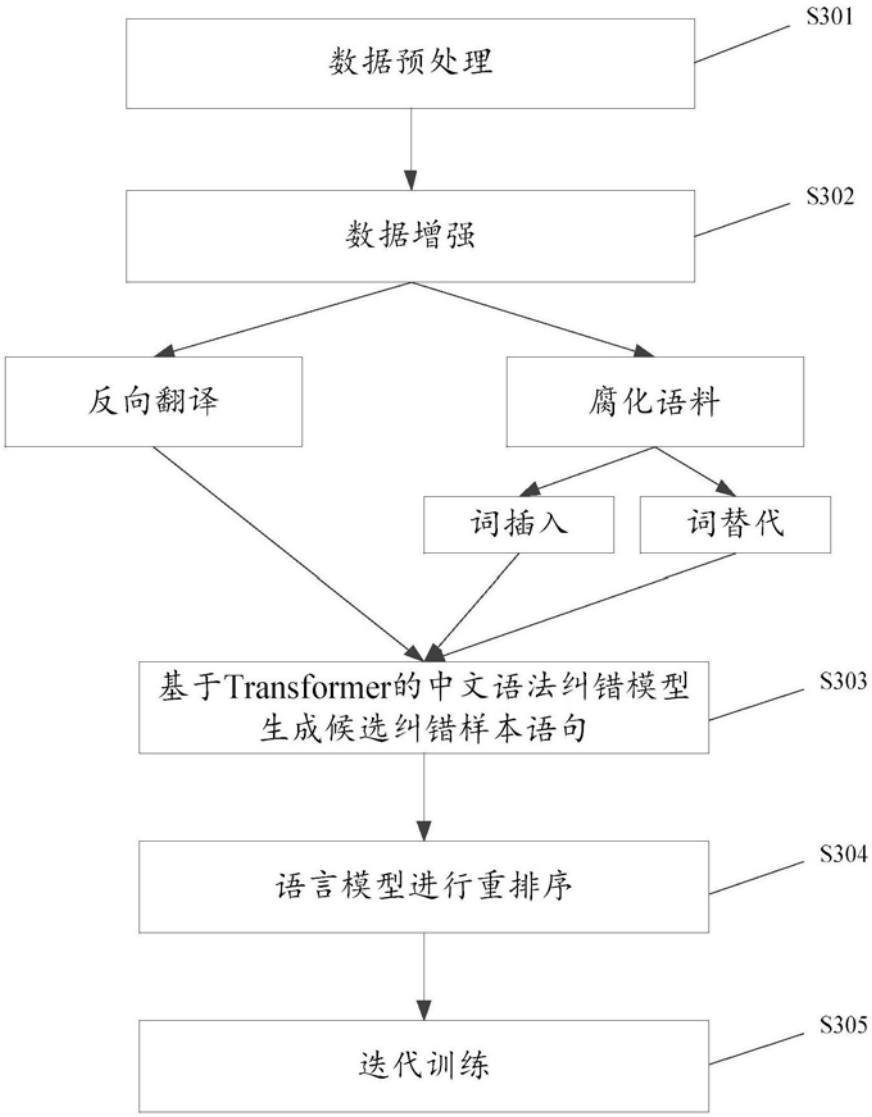

S301
数据预处理
S302
数据增强
反向翻译
腐化语料
词插入
词替代
基于Transformer的中文语法纠错模型
生成候选纠错样本语句
S303
语言模型进行重排序
S304
迭代训练
S305

图3

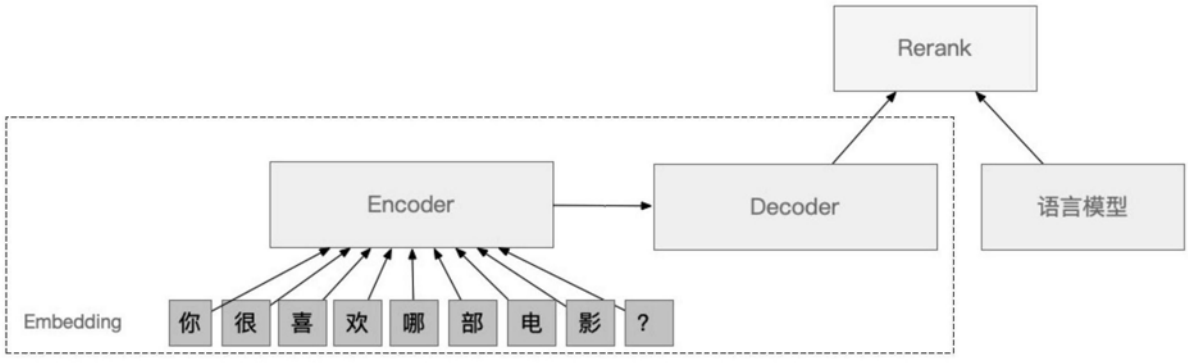

Rerank
Encoder
Decoder
语言模型
Embedding
你
很
喜
欢
哪
部
电
影
?

图4

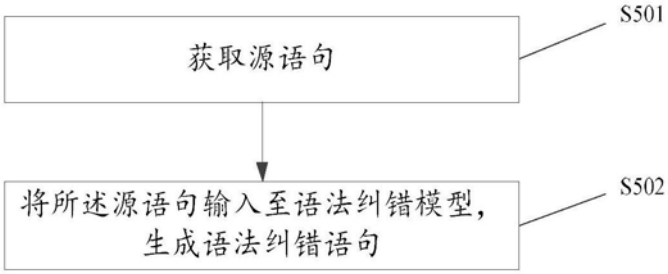

获取源语句
S501
将所述源语句输入至语法纠错模型，生成语法纠错语句
S502

图5


数据扩展模块610
获取模块620
语法纠错模块630
损失确定模块640
迭代训练模块650

图6


获取模块710
语法纠错模块720

图7

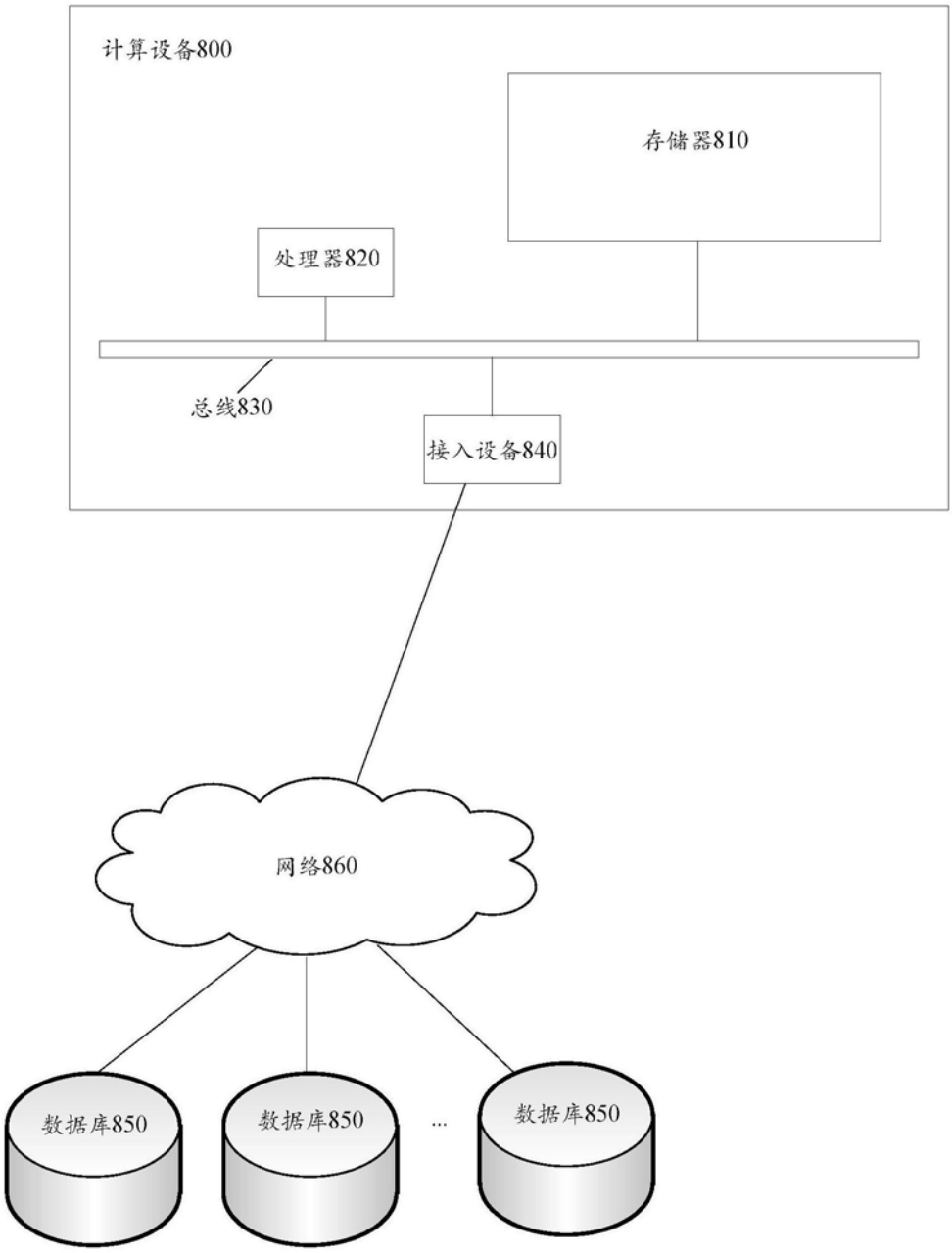

计算设备800
存储器810
处理器820
总线830
接入设备840
网络860
数据库850
数据库850
...
数据库850

图8