(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2021 年 9 月 30 日 (30.09.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/189890 A1

(51) 国际专利分类号:
G06F 40/30 (2020.01)

(21) 国际申请号: PCT/CN2020/131757

(22) 国际申请日: 2020 年 11 月 26 日 (26.11.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011139506.2 2020年10月22日 (22.10.2020) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

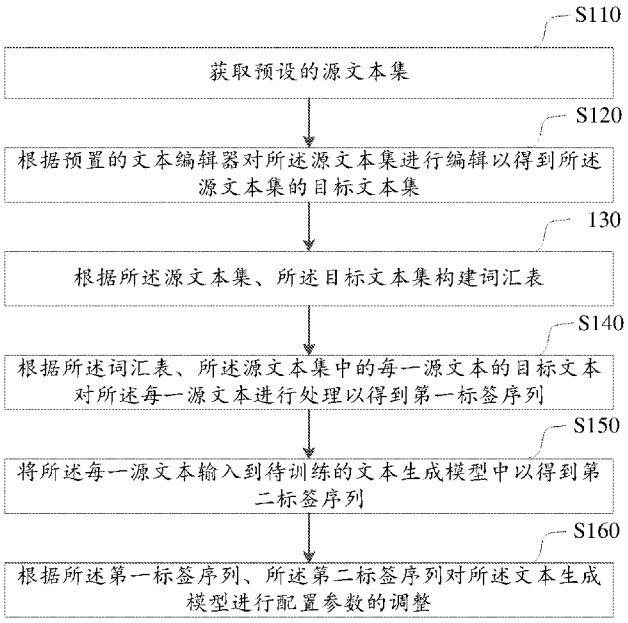
(72) 发明人: 孙超 (SUN, Chao); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 王健宗 (WANG, Jianzong); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 吴天博 (WU, Tianbo); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 程宁 (CHENG, Ning); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市精英专利事务所 (SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路6009号绿景广场B栋20层B, Guangdong 518000 (CN)。

WO 2021/189890 A1

(54) **Title:** TEXT GENERATION MODEL TRAINING METHOD AND APPARATUS BASED ON TEXT EDITING TECHNOLOGY

(54) **发明名称:** 基于文本编辑技术的文本生成模型的训练方法及装置



S110 Acquire a preset source text set
S120 Perform editing on the source text set according to a preset text editor, so as to obtain a target text set of the source text set
S130 Construct a word list according to the source text set and the target text set
S140 According to the word list and target text of each piece of source text in the source text set, process each piece of source text so as to obtain a first label sequence
S150 Input each piece of source text into a text generation model to be trained, so as to obtain a second label sequence
S160 According to the first label sequence and the second label sequence, adjust a configuration parameter of the text generation model

图 1

(57) **Abstract:** A text generation model training method and apparatus (100) based on text editing technology. The method comprises: acquiring a preset source text set (S110); performing editing on the source text set according to a preset text editor, so as to obtain a target text set of the source text set (S120); constructing a word list according to the source text set and the target text set (S130); according to the word list and target text of each piece of source text in the source text set, processing each piece of source text so as to obtain a first label sequence (S140); inputting each piece of source text into a text generation model to be trained, so as to obtain a second label sequence (S150); and according to the first label sequence and the second label sequence, adjusting a configuration parameter of the text generation model (S160). By means of the method, training a text generation model not only greatly improves the efficiency of training the text generation model, but also improves the accuracy of high-semantic text generated by the text generation model.

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

— 在修改权利要求的期限届满之前进行，在收到该修改后将重新公布(细则48.2(h))。

— 根据申请人的请求，在条约第21条(2)(a)所规定的期限届满之前进行。

(57) 摘要：一种基于文本编辑技术的文本生成模型的训练方法及装置（100），该方法包括：获取预设的源文本集（S110）；根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集（S120）；根据所述源文本集、所述目标文本集构建词汇表（S130）；根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列（S140）；将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列（S150）；根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整（S160）。通过上述方法对文本生成模型进行训练，不仅极大的提升了文本生成模型的训练效率，而且提高了文本生成模型生成高语义的文本准确率。

基于文本编辑技术的文本生成模型的训练方法及装置

本申请要求于2020年10月22日提交中国专利局、申请号为202011139506.2，发明名称为“基于文本编辑技术的文本生成模型的训练方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请属于人工智能中的机器学习技术领域，尤其涉及一种基于文本编辑技术的文本生成模型的训练方法及装置。

背景技术

文本生成是自然语言处理领域一项重要的任务，也是人工智能面临的一个重大挑战。虽然文本生成可以辅助专业人员进行专业写作，例如法律文书补全、自动生成新闻、生成文本摘要、文本复述等，但是发明人意识到现有技术中文本生成模型的训练需依赖于大量的数据，尤其在特定领域的高质量的文本数据却比较匮乏，造成文本生成模型生成的高语义文本的准确度不高。

发明内容

本申请实施例提供了一种基于文本编辑技术的文本生成模型的训练方法及装置，解决了现有技术中文本生成模型需要大量高质量的文本数据进行训练才能准确获取高语义文本的问题。

第一方面，本申请实施例提供了一种基于文本编辑技术的文本生成模型的训练方法，其包括：

获取预设的源文本集；

根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

根据所述源文本集、所述目标文本集构建词汇表；

根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

第二方面，本申请实施例提供了一种基于文本编辑技术的文本生成模型的训练装置，其包括：

第一获取单元，用于获取预设的源文本集；

编辑单元，用于根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

第一构建单元，用于根据所述源文本集、所述目标文本集构建词汇表；

处理单元，用于根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

输入单元，用于将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

第一调整单元，用于根据所述第一标签序列、所述第二标签序列对所述文本生成模型进

行配置参数的调整。

第三方面，本申请实施例又提供了一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其中，所述处理器执行所述计算机程序时执行以下步骤：

获取预设的源文本集；

根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

根据所述源文本集、所述目标文本集构建词汇表；

根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

第四方面，本申请实施例还提供了一种计算机可读存储介质，其中所述计算机可读存储介质存储有计算机程序，所述计算机程序当被处理器执行时使所述处理器执行以下步骤：

获取预设的源文本集；

根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

根据所述源文本集、所述目标文本集构建词汇表；

根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

本申请通过所述的基于文本编辑技术的文本生成模型的训练方法，使得当源文本集较少时，仍然可以完成文本模型的训练并极大的提高了文本生成模型的训练效率，而且提高了生成高语义文本的准确率。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例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。

图 1 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练方法的流程示意图；

图 2 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练方法的子流程示意图；

图 3 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练方法的另一子流程示意图；

图 4 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练方法的另一子流程示意图；

图 5 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练方法的另一子流程示意图；

图 6 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练方法的另一子流程示意图；

图 7 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练装置的示意性框图；

图 8 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练装置的子单元示意性框图；

图 9 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练装置的另一子单元示意性框图；

图 10 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练装置的另一子单元示意性框图；

图 11 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练装置的另一子单元示意性框图；

图 12 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练装置的另一子单元示意性框图；

图 13 为本申请实施例提供的计算机设备的示意性框图。

具体实施方式

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。

应当理解，当在本说明书和所附权利要求书中使用时，术语“包括”和 “包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

还应当理解，在此本申请说明书中所使用的术语仅仅是出于描述特定实施例的目的而并不意在限制本申请。如在本申请说明书和所附权利要求书中所使用的那样，除非上下文清楚地指明其它情况，否则单数形式的“一”、“一个”及“该”意在包括复数形式。

还应当进一步理解，在本申请说明书和所附权利要求书中使用的术语“和/ 或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。

请参阅图 1，图 1 为本申请实施例提供的基于文本编辑技术的文本生成模型的训练方法的流程示意图。所述基于文本编辑技术的文本生成模型的训练方法在服务器中进行搭建并运行，在服务器中对文本生成模型进行训练的过程时，通过获取对文本生成模型进行训练所需的源文本集后，将源文本集中的每一条源文本进行编辑以得到每一条源文本的目标文本，然后通过预设的词汇表以及目标文本对源文本进行处理以得到第一标签序列，同时将源文本输入到待训练的文本生成模型中以得到第二标签序列，通过计算第一标签序列与第二标签序列的相似度来对待训练的文本生成模型进行配置参数的调整，使得当源文本集较少时，仍然可以完成文本模型的训练并极大的提高了文本生成模型的训练效率，而且提高了生成高语义文本的准确率。

如图 1 所示，该方法包括步骤 S110~S150。

S110、获取预设的源文本集。

获取预设的源文本集。具体的，所述源文本集为需对文本生成模型进行训练的数据集，所述源文本集中的文本数量可根据用户需求进行配置，既可以为大量的文本数据，也可以为少量的文本数据。在本申请实施例中，采用数据量较少的源文本集队文本生成模型进行训练。

S120、根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集。

根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集。具体的，所述文本编辑器为可用于对所述源文本集中的每一条源文本进行编辑以得到具有高语义的目标文本的文本编辑工具，Window 旗下的记事本，Mac OS X 旗下的文本编辑，Linux 旗下的 vi、emacs、gedit 等均可用于对所述源文本集中的每一条源文本进行编辑。例如：当源文本为“小 明 出 生 于 1993 年 。小 明 生 在 上 海 ”时，使用文本编辑器编辑后的目标文本为“小 明 于 1993 年 初 出 生 在 上 海”。

S130、根据所述源文本集、所述目标文本集构建词汇表。

根据所述源文本集、所述目标文本集构建词汇表。具体的，将所述目标文本集中每一目标文本中不存在于该目标文本的源文本中的词语作为所述词汇表中的词语，在构建词汇表的过程中，通常为了减少后续使用词汇表时的计算量，需要对词汇表进行优化以使得词汇表尽可能小，而从源文本集中获取词汇表中的词语需根据该词语在目标文本集中出现的频率进行筛选，例如，将词汇表中在目标文本集中出现十次以下的词语进行剔除，便可得到优化后的词汇表。本申请实施例中的词汇表构建完成后存储于区块链中，保证了词汇表存储的安全性能。

在一实施例中，如图 2 所示，步骤 S130 包括子步骤 S131 和 S132。

S131、根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列。

根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列。具体的，通过使用最长公共子序列技术来获取每一源文本以及每一源文本的目标文本的最长公共子序列。最长公共子序列的定义为：一个序列 S，如果分别是两个或多个已知序列的子序列，且是所有符合此条件序列中最长的，则 S 称为已知序列的最长公共子序列。

在一实施例中，如图 3 所示，步骤 S131 包括子步骤 S1311 和 S1312。

S1311、获取所述每一源文本的子序列集合以及所述每一源文本的目标文本的子序列集合。

获取所述每一源文本的子序列集合以及所述每一源文本的目标文本的子序列集合。具体的，所述每一源文本的子序列集合中的子序列为在不改变每一源文本字符顺序的前提下将每一源文本进行拆分而得到子序列，每一源文本拆分后的子序列组合成所述每一源文本的子序列集合，同样所述每一源文本的目标文本的子序列集合参考每一源文本的子序列集合的获取方式得到。

S1312、将所述每一源文本的子序列集合中的每一子序列分别与所述目标文本的子序列集合中的每一子序列进行匹配以得到所述每一源文本与所述每一源文本的目标文本的公共子序列集合并将所述公共子序列集合中的最长公共子序列作为所述最长公共子序列。

将所述每一源文本的子序列集合中的每一子序列分别与所述目标文本的子序列集合中的每一子序列进行匹配以得到所述每一源文本与所述每一源文本的目标文本的公共子序列集合并将所述公共子序列集合中的最长公共子序列作为所述最长公共子序列。具体的，所述公共子序列集合中的每一个公共子序列均为所述每一源文本和所述每一源文本的目标文本的公共子序列，所述公共子序列集合中最长的序列便为所述每一源文本和所述每一源文本的目标文本的最长公共子序列。

S132、根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表。

根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表。具体的，所述最长公共子序列为所述每一源文本的目标文本与所述每一源文本的最长公共子序列，通过所述最长公共子序列从所述每一源文本的目标文本中获取不存在于所述最长公共子序列中的词语并将该词语作为所述词汇表中的词语，进而完成词汇表的构建。

在一实施例中，如图 4 所示，步骤 S132 包括子步骤 S1321 和 S1322。

S1321、将所述每一源文本的目标文本进行分词处理以得到所述每一源文本的目标文本的词语。

将所述每一源文本的目标文本进行分词处理以得到所述每一源文本的目标文本的词语。具体的，本实施例中采用基于字符串的分词方法中的逆向最大匹配法对每一源文本的目标文本进行分词处理，其分词过程为：设定预置的词典中最长词条所包含的汉字数量为 L，从目标文本的字符串末尾开始处理。在每一次循环开始时，都取所述字符串最后的 L 个字作为处理对象，查找所述词典。若所述词典中存在这样的一个 L 字词，则匹配成功，所述处理对象则被作为一个词被切分；若不成功，则去掉该处理对象的第一个汉字，剩下的字符串作为新的处理对象，再次进行匹配，直到切分成功为止，即完成一轮匹配，切分出一个词，类此循环直至目标文本中的词语全部被切分出来为止。

S1322、将所述每一源文本的目标文本的词语与所述最长公共子序列进行匹配以从所述每一源文本的目标文本的词语中获取构成所述词汇表的词语。

将所述每一源文本的目标文本的词语与所述最长公共子序列进行匹配以从所述每一源文本的目标文本的词语中获取构成所述词汇表的词语。具体的，将所述每一源文本的目标文本的词语是否存在于所述最长公共子序列作为目标文本的词语与最长公共子序列是否匹配成功的结果，如果匹配成功，则该词语不为词汇表中的词语；若匹配不成功，则将其作为词汇表中的词语。

S140、根据所述词汇表、所述每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列。

根据所述词汇表、所述每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列。具体的，所述每一源文本中不包含有所述词汇表中的词汇，通过标注所述每一源文

本中与所述每一源文本的目标文本的最长公共子序列以对所述每一源文本进行标注，然后将标注后的每一源文本进行拆分得到标注后的每一源文本的字符，将该字符与词汇表中的词语进行匹配以得到新的词语，然后将匹配后的词语进行拼接便可得到第一标签序列。

在一实施例中，如图 5 所示，步骤 S140 包括子步骤 S141、S142、S143 和 S144。

S141、根据所述最长公共子序列对所述每一源文本进行标注以得到标注后的每一源文本。

根据所述最长公共子序列对所述每一源文本进行标注以得到标注后的每一源文本。具体的，将所述每一源文本中属于最长公共子序列中的字符标注第一标签，记作符号“keep”；将不属于最长公共子序列中的字符标注为第二标签，记作符号“delete”，进而使得所述每一源文本标注有第一标签和第二标签的文本。例如，源文本为“小 明 出 生 于 1993 年 。小 明 生 在 上 海 ”时，标注有标签的源文本标注的标签的顺序为“keep keep delete delete keep keep keep delete delete delete keep keep keep keep”。

S142、将所述标注后的每一源文本进行分词处理以得到所述标注后的每一源文本的字符集合。

将所述标注后的每一源文本进行分词处理以得到所述标注后的每一源文本的字符集合。具体的，所述标注后的每一源文本的字符集合为标注有第一标签的字符的集合，分词的过程为：首先将每一源文本中标注有第一标签和第二标签相邻的两个词进行分词处理，然后单独将标注有第一标签的语句进行分词处理以得到所述标注后的每一源文本的字符集合。例如，源文本为“小 明 出 生 于 1993 年 。小 明 生 在 上 海 ”时，标注有标签的源文本标注的标签的顺序为“keep keep delete delete keep keep keep delete delete delete keep keep keep keep”，最终得到的标注后的每一源文本的字符集合为：[小、明、于、1993、年、生、在、上、海]，其中，每一个字符上均标注有第一标签。

S143、将所述词汇表中的词语分别与所述字符集合中的字符进行匹配以得到词语集合。

将所述词汇表中的词语分别与所述字符集合中的字符进行匹配以得到词语集合。具体的，将词汇表中的每一个词语均与字符集合中的每一个字符进行匹配以组成新的词语，然后在预设的词典中进行查找以得到该词典中是否存在新组成的词语，若不存在，则忽略该新组成的词语，最终筛选出的新的词语组成所述词语集合。

S144、将所述词语集合中的词语进行拼接以得到所述第一标签序列。

将所述词语集合中的词语进行拼接以得到所述第一标签序列。具体的，将所述词语集合中的词语以所述标注后的每一源文本中字符的排列顺序进行拼接以得到所述第一标签序列。在对词语集合中所有的词语进行拼接的过程中，需按照源文本字符组成的顺序进行拼接，拼接完成至少得到一条文本，然后将拼接完的得到的文本进行句法分析，筛选出最符合源文本的语句并将其作为所述第一标签序列，通过所述第一标签序列便可预测出源文本的目标文本。例如，源文本为“小 明 出 生 于 1993 年 。小 明 生 在 上 海 ”时，第一标签序列为“keep keep delete delete keep keep keep ^{初}|delete delete delete ^{出}|keep keep keep keep”，其中“^{初}|”和“^{出}|”均为根据词汇表中的词语进行标注的标签。

S150、将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列。

将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列。具体的，所述待训练的文本生成模型为 encoder-decoder 模型架构的模型，即所述文本生成模型包括编码器和解码器。源文本输入到待训练的文本生成模型中后，通过编码和解码便可得到第二标签序列，通过第二标签序列便可预测出该源文本的目标文本。在本申请实施例中，所述文本生成模型的编码器采用预训练的 RoBERTa 中文模型，即由 12 层 transformer 组成；解码器采用单层 transformer，可在保证精度的同时兼顾模型的推理速度。

S160、根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。具体的，通过所述第一标签序列和所述第二标签序列均可获得源文本的目标文本，所述第一标签序列在源文本的目标文本的基础上获得，而第二标签序列文本通过待训练的文本生成模型生成，即第一标签序列相对于第二标签序列而言，根据第一标签序列获得的目标文本更准确，通过将第一标签序列与第二标签序列进行计算以得到第一标签序列与第二标签序列的相似度，最后根据相似度对文本生成模型的配置参数进行相应的调整，以使得文本生成模型生成的第二标签序列更接近于第一标签序列，进而完成对文本生成模型的训练。

在一实施例中，如图 6 所示，步骤 S160 包括子步骤 S161 和 S162。

S161、获取所述第二标签序列与所述第一标签序列的相似度。

获取所述第二标签序列与所述第一标签序列的相似度。具体的，在计算所述第一标签序列和第二标签序列的相似度时，需将第一标签序列和第二标签序列进行向量化，然后进行距离计算，将计算得到的距离作为所述第二标签序列与所述第一标签序列的相似度，距离越长，相似度越低，距离越短，相似度越高。在本申请实施例中采用欧式距离计算方式得到所述相似度。所述欧式距离是一个通常采用的距离定义，指在 n 维空间中两个点之间的真实距离，或者向量的自然长度。所述第二标签序列与所述第一标签序列的欧式距离计算公式为：$d_{12}=\sqrt{\sum_{k=1}^{n}\left(x_{1k}-x_{2k}\right)^{2}}$，其中，n 表示向量的维度，$x_{1k}$为第一标签序列的向量，$x_{2k}$为第二标签序列的向量。

S162、若所述相似度低于预设的阈值，根据所述相似度对所述文本生成模型的配置参数进行调整。

若所述相似度低于预设的阈值，根据所述相似度对所述文本生成模型的配置参数进行调整。具体的，预设的阈值为是否对文本生成模型的参数进行调整的以使得文本生成模型能更加准确生成高语义文本。所述阈值可根据实际情况进行设定，在此不做限定。

本申请所述的基于文本编辑技术的文本生成模型的训练方法，通过获取预设的源文本集；根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；根据所述源文本集、所述目标文本集构建词汇表；根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；根据所述第一标签序列、所述第二标签序列对所

述文本生成模型进行配置参数的调整。本申请所述的基于文本编辑技术的文本生成模型的训练方法不仅极大的提升了文本生成模型的训练效率，而且提高了文本生成模型生成高语义的文本准确率。

本申请实施例还提供了一种基于文本编辑技术的文本生成模型的训练装置 100，该装置用于执行前述基于文本编辑技术的文本生成模型的训练方法的任一实施例。具体地，请参阅图 7，图 7 是本申请实施例提供的基于文本编辑技术的文本生成模型的训练装置 100 的示意性框图。

如图 7 所示，所述的基于文本编辑技术的文本生成模型的训练装置 100，该装置包括第一获取单元 110、编辑单元 120、第一构建单元 130、处理单元 140、输入单元 150 和第一调整单元 160。

第一获取单元 110，用于获取预设的源文本集。

编辑单元 120，用于根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集。

第一构建单元 130，用于根据所述源文本集、所述目标文本集构建词汇表。

在其他发明实施例中，如图 8 所示，所述第一构建单元 130 包括：第二构建单元 131 和第三构建单元 132。

第二构建单元 131，用于根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列。

在其他发明实施例中，如图 9 所示，所述第二构建单元 131 包括：第二获取单元 1311 和第一匹配单元 1312。

第二获取单元 1311，用于获取所述每一源文本的子序列集合以及所述每一源文本的目标文本的子序列集合。

第一匹配单元 1312，用于将所述每一源文本的子序列集合中的每一子序列分别与所述目标文本的子序列集合中的每一子序列进行匹配以得到所述每一源文本与所述每一源文本的目标文本的公共子序列集合并将所述公共子序列集合中的最长公共子序列作为所述最长公共子序列。

第三构建单元 132，用于根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表。

在其他发明实施例中，如图 10 所示，所述第三构建单元 132 包括：第一分词单元 1321 和第二匹配单元 1322。

第一分词单元 1321，用于将所述每一源文本的目标文本进行分词处理以得到所述每一源文本的目标文本的词语。

第二匹配单元 1322，用于将所述每一源文本的目标文本的词语与所述最长公共子序列进行匹配以从所述每一源文本的目标文本的词语中获取构成所述词汇表的词语。

处理单元 140，用于根据所述词汇表、所述源文本集中每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列。

在其他发明实施例中，如图 11 所示，所述处理单元 140 包括：标注单元 141、第二分词单元 142、第三匹配单元 143 和拼接单元 144。

标注单元 141，用根据所述最长公共子序列对所述每一源文本进行标注以得到标注后的每一源文本。

第二分词单元 142，用于将所述标注后的每一源文本进行分词处理以得到所述标注后的每一源文本的字符集合。

第三匹配单元 143，用于将所述词汇表中的词语分别与所述字符集合中的字符进行匹配以得到词语集合。

拼接单元 144，用于将所述词语集合中的词语进行拼接以得到所述第一标签序列。

输入单元 150，用于将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列。

第一调整单元 160，用于根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

在其他发明实施例中，如图 12 所示，所述第一调整单元 160 包括：第三获取单元 161 和第二调整单元 162。

第三获取单元 161，用于获取所述第二标签序列与所述第一标签序列的相似度。

第二调整单元 162，用于若所述相似度低于预设的阈值，根据所述相似度对所述文本生成模型的配置参数进行调整。

本申请实施例所提供的基于文本编辑技术的文本生成模型的训练装置 100 用于执行上述用于获取预设的源文本集；根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；根据所述源文本集、所述目标文本集构建词汇表；根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

请参阅图 13，图 13 是本申请实施例提供的计算机设备的示意性框图。

参阅图 13，该设备 500 包括通过系统总线 501 连接的处理器 502、存储器和网络接口 505，其中，存储器可以包括非易失性存储介质 503 和内存储器 504。

该非易失性存储介质 503 可存储操作系统 5031 和计算机程序 5032。该计算机程序 5032 被执行时，可使得处理器 502 执行基于文本编辑技术的文本生成模型的训练方法。该处理器 502 用于提供计算和控制能力，支撑整个设备 500 的运行。该内存储器 504 为非易失性存储介质 503 中的计算机程序 5032 的运行提供环境，该计算机程序 5032 被处理器 502 执行时，可使得处理器 502 执行基于文本编辑技术的文本生成模型的训练方法。该网络接口 505 用于进行网络通信，如提供数据信息的传输等。本领域技术人员可以理解，图 13 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的设备 500 的限定，具体的设备 500 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

其中，所述处理器 502 用于运行存储在存储器中的计算机程序 5032，以实现上述基于文本编辑技术的文本生成模型的训练方法的任一实施例。

本领域普通技术人员可以理解的是实现上述实施例的方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成。该计算机程序可存储于一存储介质中，该存储介质可以为计算机可读存储介质。该计算机程序被该计算机系统中的至少一个处理器执行，以实现上述方法的实施例的流程步骤。

因此，本申请还提供了一种计算机可读存储介质。该计算机可读存储介质可以是非易失性，也可以是易失性。该存储介质存储有计算机程序，该计算机程序当被处理器执行时实现上述基于文本编辑技术的文本生成模型的训练方法的任一实施例。

该计算机可读存储介质可以是 U 盘、移动硬盘、只读存储器（ROM, Read-Only Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

在本申请所提供的几个实施例中，应该理解到，所揭露的装置、设备和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，上述描述的装置、设备和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权利要求书

1.一种基于文本编辑技术的文本生成模型的训练方法，其中，包括以下步骤：

获取预设的源文本集；

根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

根据所述源文本集、所述目标文本集构建词汇表；

根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

2.根据权利要求 1 所述的基于文本编辑技术的文本生成模型的训练方法，其中，所述根据所述源文本集、所述目标文本集构建所述词汇表，包括：

根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列；

根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表。

3.根据权利要求 2 所述的基于文本编辑技术的文本生成模型的训练方法，其中，所述根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列，包括：

获取所述每一源文本的子序列集合以及所述每一源文本的目标文本的子序列集合；

将所述每一源文本的子序列集合中的每一子序列分别与所述目标文本的子序列集合中的每一子序列进行匹配以得到所述每一源文本与所述每一源文本的目标文本的公共子序列集合并将所述公共子序列集合中的最长公共子序列作为所述最长公共子序列。

4.根据权利要求 2 所述的基于文本编辑技术的文本生成模型的训练方法，其中，所述根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表，包括：

将所述每一源文本的目标文本进行分词处理以得到所述每一源文本的目标文本的词语；

将所述每一源文本的目标文本的词语与所述最长公共子序列进行匹配以从所述每一源文本的目标文本的词语中获取构成所述词汇表的词语。

5.根据权利要求 2 所述的基于文本编辑技术的文本生成模型的训练方法，其中，所述根据预设的词汇表、所述每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列，包括：

根据所述最长公共子序列对所述每一源文本进行标注以得到标注后的每一源文本；

将所述标注后的每一源文本进行分词处理以得到所述标注后的每一源文本的字符集合；

将所述词汇表中的词语分别与所述字符集合中的字符进行匹配以得到词语集合；

将所述词语集合中的词语进行拼接以得到所述第一标签序列。

6.根据权利要求 5 所述的基于文本编辑技术的文本生成模型的训练方法，其中，所述将所述词语集合中的词语进行拼接以得到所述第一标签序列，包括：

将所述词语集合中的词语以所述标注后的每一源文本中字符的排列顺序进行拼接以得到所述第一标签序列。

7.根据权利要求 1 所述的基于文本编辑技术的文本生成模型的训练方法，其中，所述根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整，包括：

获取所述第二标签序列与所述第一标签序列的相似度；

若所述相似度低于预设的阈值，根据所述相似度对所述文本生成模型的配置参数进行调整。

8.一种基于文本编辑技术的文本生成模型的训练装置，其中，包括：

第一获取单元，用于获取预设的源文本集；

编辑单元，用于根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

第一构建单元，用于根据所述源文本集、所述目标文本集构建词汇表；

处理单元，用于根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

输入单元，用于将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

第一调整单元，用于根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

9.一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其中，所述处理器执行所述计算机程序时执行以下步骤：

获取预设的源文本集；

根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

根据所述源文本集、所述目标文本集构建词汇表；

根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

10.根据权利要求 9 所述的计算机设备，其中，所述根据所述源文本集、所述目标文本集构建所述词汇表，包括：

根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列；

根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表。

11.根据权利要求 10 所述的计算机设备，其中，所述根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列，包括：

获取所述每一源文本的子序列集合以及所述每一源文本的目标文本的子序列集合；

将所述每一源文本的子序列集合中的每一子序列分别与所述目标文本的子序列集合中的每一子序列进行匹配以得到所述每一源文本与所述每一源文本的目标文本的公共子序列集合并将所述公共子序列集合中的最长公共子序列作为所述最长公共子序列。

12.根据权利要求 10 所述的计算机设备，其中，所述根据所述每一源文本的目标文本、

所述最长公共子序列构建所述词汇表，包括：

将所述每一源文本的目标文本进行分词处理以得到所述每一源文本的目标文本的词语；

将所述每一源文本的目标文本的词语与所述最长公共子序列进行匹配以从所述每一源文本的目标文本的词语中获取构成所述词汇表的词语。

13.根据权利要求 10 所述的计算机设备，其中，所述根据预设的词汇表、所述每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列，包括：

根据所述最长公共子序列对所述每一源文本进行标注以得到标注后的每一源文本；

将所述标注后的每一源文本进行分词处理以得到所述标注后的每一源文本的字符集合；

将所述词汇表中的词语分别与所述字符集合中的字符进行匹配以得到词语集合；

将所述词语集合中的词语进行拼接以得到所述第一标签序列。

14.根据权利要求 13 所述的计算机设备，其中，所述将所述词语集合中的词语进行拼接以得到所述第一标签序列，包括：

将所述词语集合中的词语以所述标注后的每一源文本中字符的排列顺序进行拼接以得到所述第一标签序列。

15.根据权利要求 9 所述的计算机设备，其中，所述根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整，包括：

获取所述第二标签序列与所述第一标签序列的相似度；

若所述相似度低于预设的阈值，根据所述相似度对所述文本生成模型的配置参数进行调整。

16.一种计算机可读存储介质，其中，所述计算机可读存储介质存储有计算机程序，所述计算机程序当被处理器执行以下步骤：

获取预设的源文本集；

根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集；

根据所述源文本集、所述目标文本集构建词汇表；

根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列；

将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列；

根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整。

17.根据权利要求 16 所述的计算机可读存储介质，其中，所述根据所述源文本集、所述目标文本集构建所述词汇表，包括：

根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列；

根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表。

18.根据权利要求 17 所述的计算机可读存储介质，其中，所述根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列，包括：

获取所述每一源文本的子序列集合以及所述每一源文本的目标文本的子序列集合；

将所述每一源文本的子序列集合中的每一子序列分别与所述目标文本的子序列集合中的每一子序列进行匹配以得到所述每一源文本与所述每一源文本的目标文本的公共子序列集合并将所述公共子序列集合中的最长公共子序列作为所述最长公共子序列。

19.根据权利要求 17 所述的计算机可读存储介质，其中，所述根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表，包括：

将所述每一源文本的目标文本进行分词处理以得到所述每一源文本的目标文本的词语；

将所述每一源文本的目标文本的词语与所述最长公共子序列进行匹配以从所述每一源文本的目标文本的词语中获取构成所述词汇表的词语。

20.根据权利要求 17 所述的计算机可读存储介质，其中，所述根据预设的词汇表、所述每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列，包括：

根据所述最长公共子序列对所述每一源文本进行标注以得到标注后的每一源文本；

将所述标注后的每一源文本进行分词处理以得到所述标注后的每一源文本的字符集合；

将所述词汇表中的词语分别与所述字符集合中的字符进行匹配以得到词语集合；

将所述词语集合中的词语进行拼接以得到所述第一标签序列。

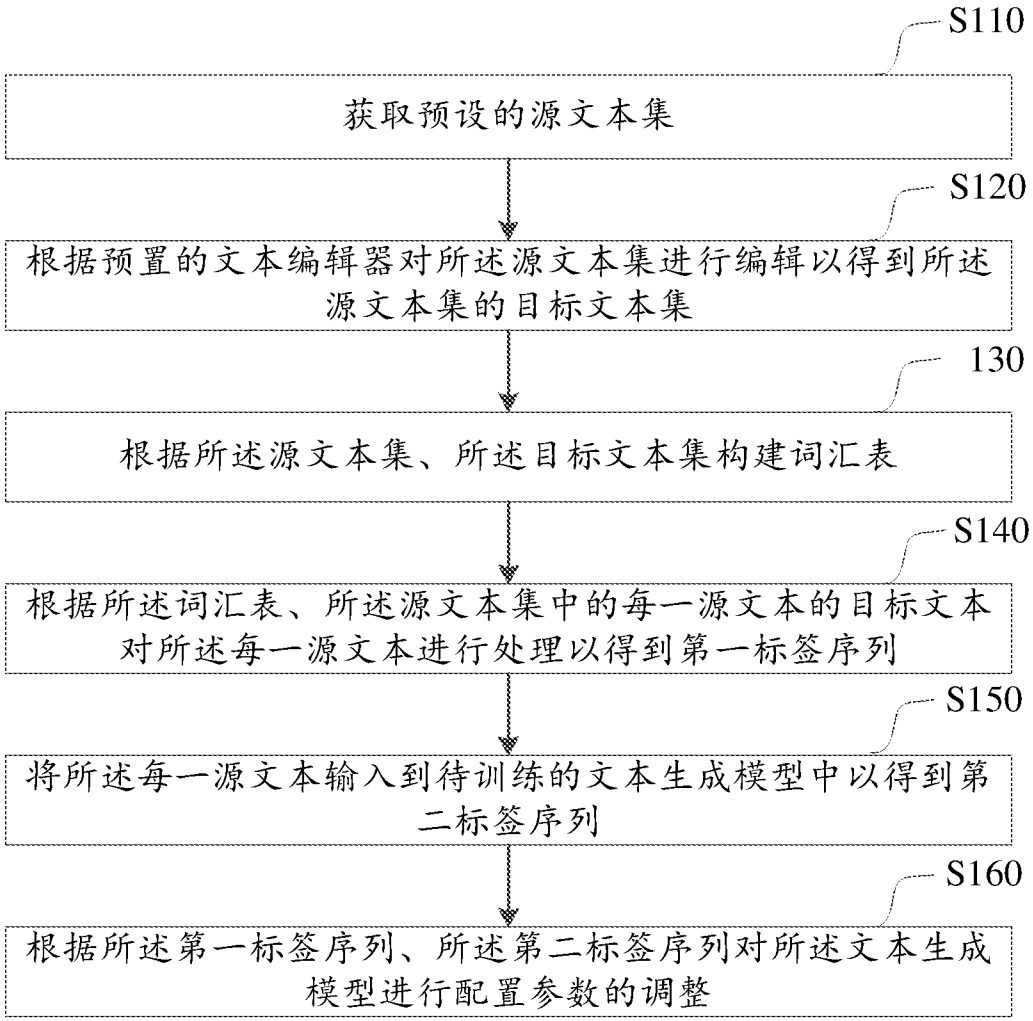

S110
获取预设的源文本集
S120
根据预置的文本编辑器对所述源文本集进行编辑以得到所述源文本集的目标文本集
130
根据所述源文本集、所述目标文本集构建词汇表
S140
根据所述词汇表、所述源文本集中的每一源文本的目标文本对所述每一源文本进行处理以得到第一标签序列
S150
将所述每一源文本输入到待训练的文本生成模型中以得到第二标签序列
S160
根据所述第一标签序列、所述第二标签序列对所述文本生成模型进行配置参数的调整

图 1

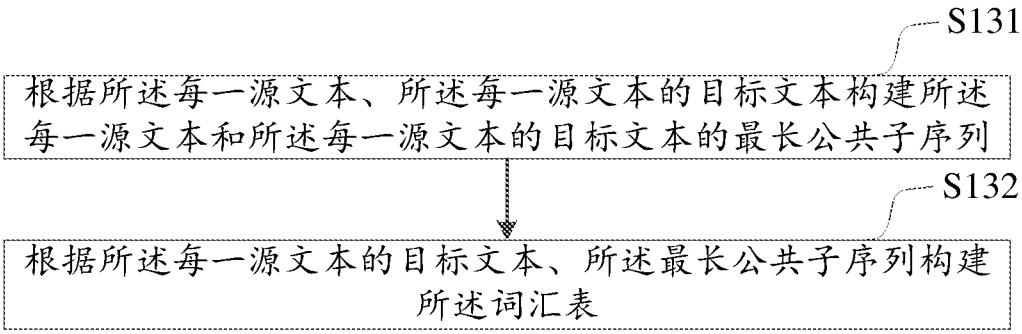

S131
根据所述每一源文本、所述每一源文本的目标文本构建所述每一源文本和所述每一源文本的目标文本的最长公共子序列
S132
根据所述每一源文本的目标文本、所述最长公共子序列构建所述词汇表

图 2

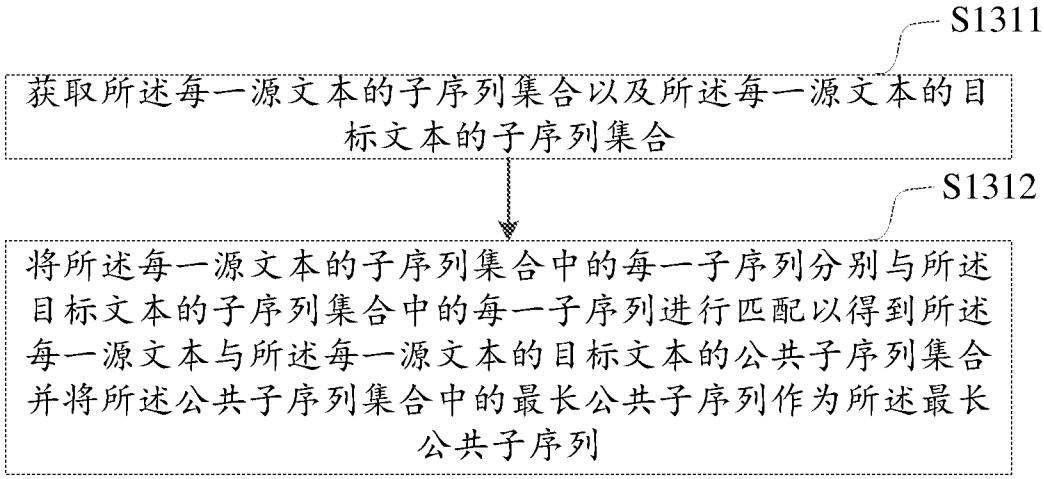

S1311
获取所述每一源文本的子序列集合以及所述每一源文本的目标文本的子序列集合
S1312
将所述每一源文本的子序列集合中的每一子序列分别与所述目标文本的子序列集合中的每一子序列进行匹配以得到所述每一源文本与所述每一源文本的目标文本的公共子序列集合并将所述公共子序列集合中的最长公共子序列作为所述最长公共子序列

图 3

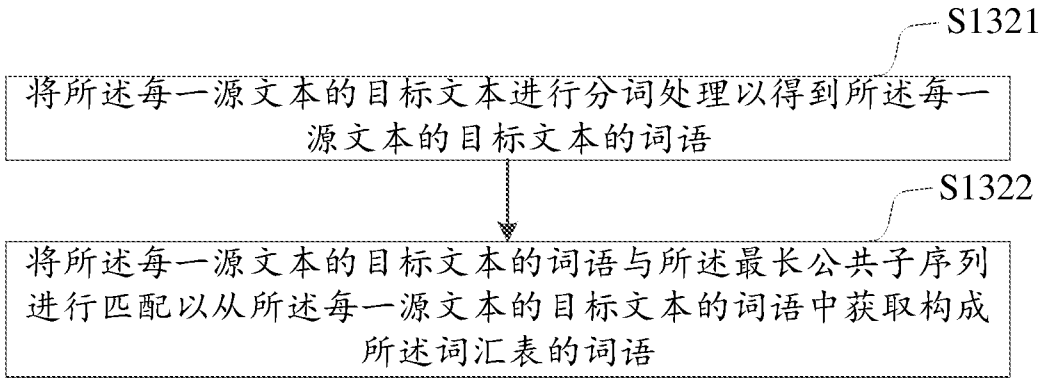

S1321
将所述每一源文本的目标文本进行分词处理以得到所述每一源文本的目标文本的词语
S1322
将所述每一源文本的目标文本的词语与所述最长公共子序列进行匹配以从所述每一源文本的目标文本的词语中获取构成所述词汇表的词语

图 4

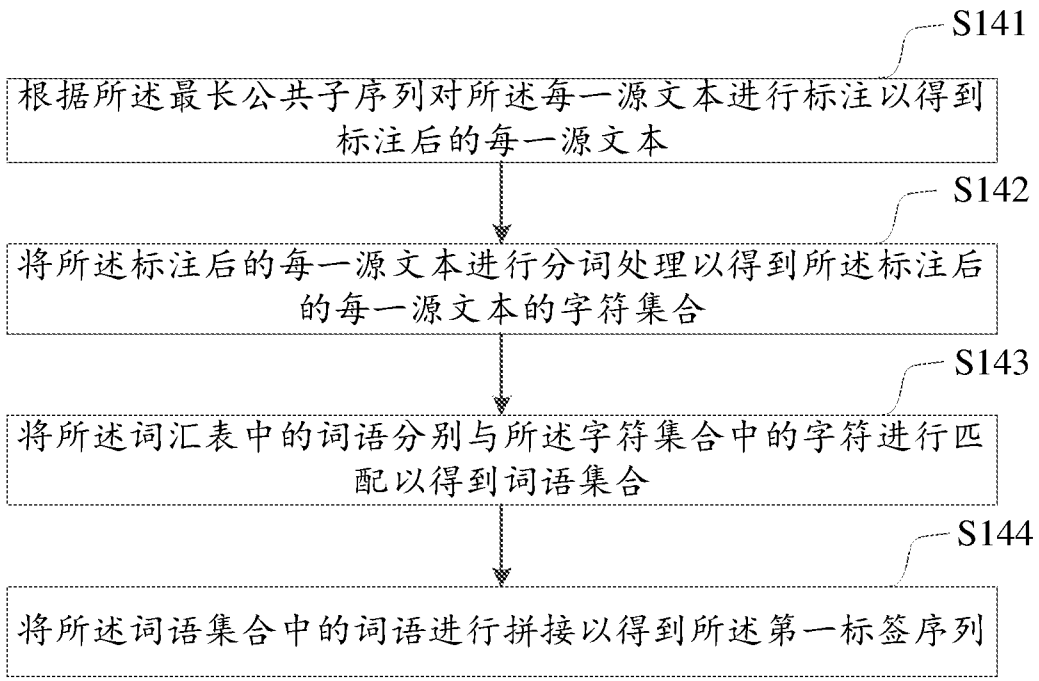

S141
根据所述最长公共子序列对所述每一源文本进行标注以得到标注后的每一源文本
S142
将所述标注后的每一源文本进行分词处理以得到所述标注后的每一源文本的字符集合
S143
将所述词汇表中的词语分别与所述字符集合中的字符进行匹配以得到词语集合
S144
将所述词语集合中的词语进行拼接以得到所述第一标签序列

图 5

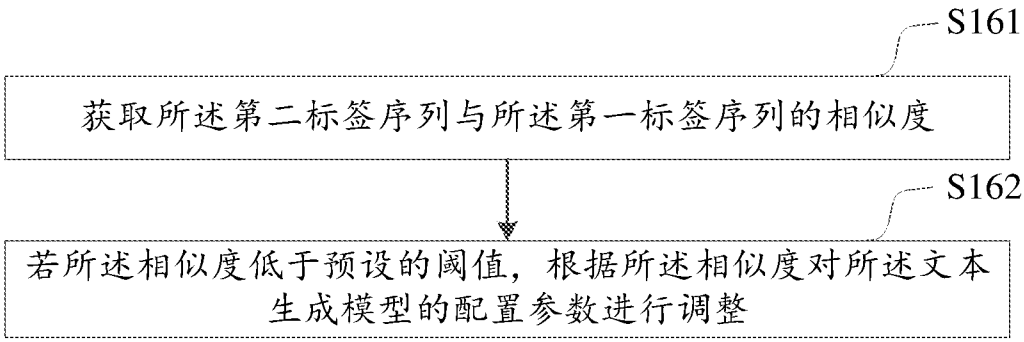

S161
获取所述第二标签序列与所述第一标签序列的相似度
S162
若所述相似度低于预设的阈值，根据所述相似度对所述文本生成模型的配置参数进行调整

图 6

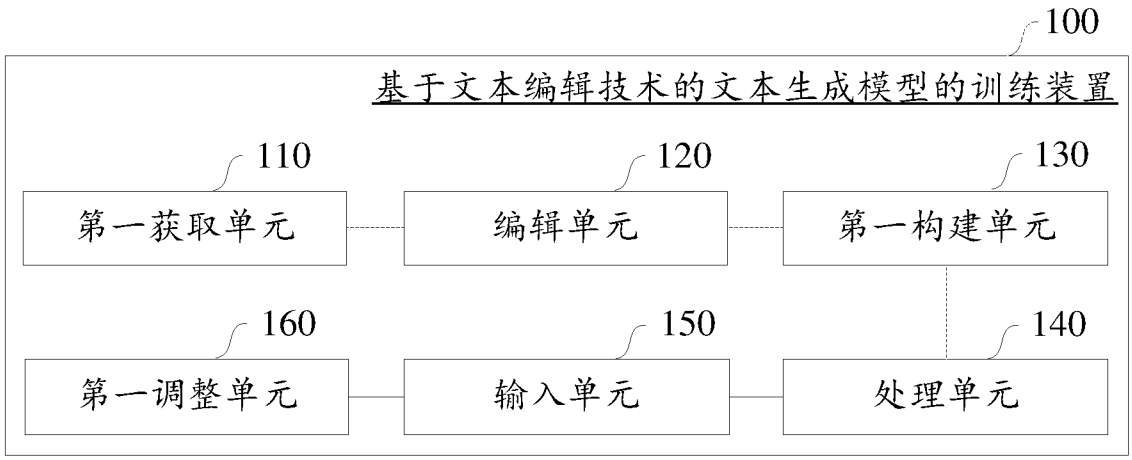

100
基于文本编辑技术的文本生成模型的训练装置
110
第一获取单元
120
编辑单元
130
第一构建单元
160
第一调整单元
150
输入单元
140
处理单元

图 7

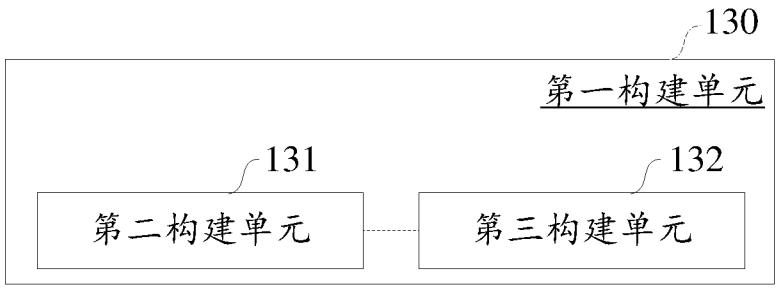

130
第一构建单元
131
第二构建单元
132
第三构建单元

图 8

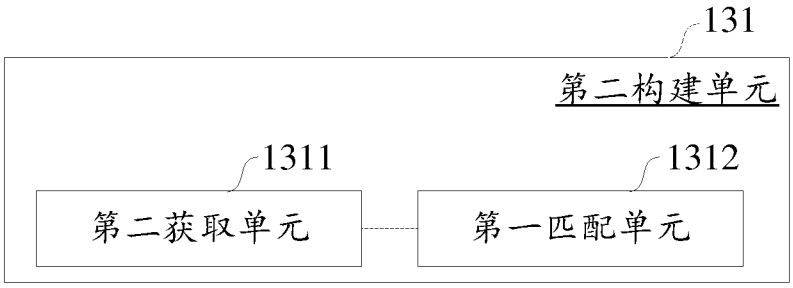

131
第二构建单元
1311
第二获取单元
1312
第一匹配单元

图 9

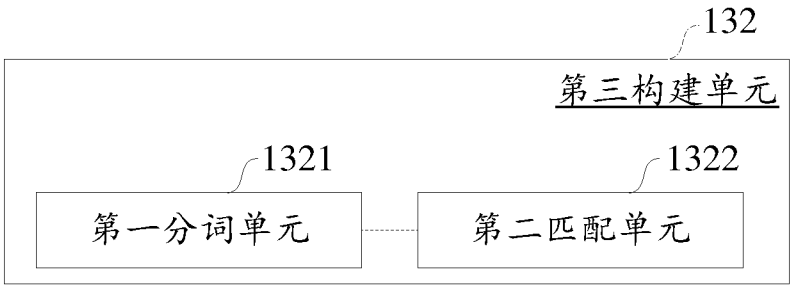

132
第三构建单元
1321
1322
第一分词单元
第二匹配单元

图 10

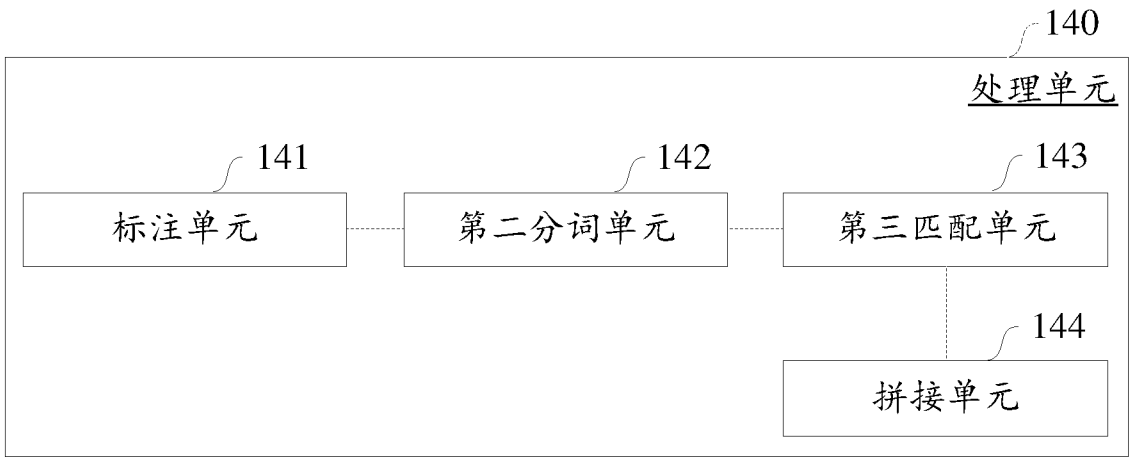

140
处理单元
141
142
143
标注单元
第二分词单元
第三匹配单元
144
拼接单元

图 11

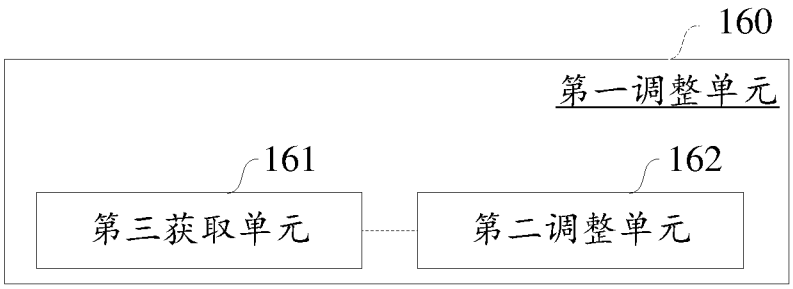

160
第一调整单元
161
162
第三获取单元
第二调整单元

图 12

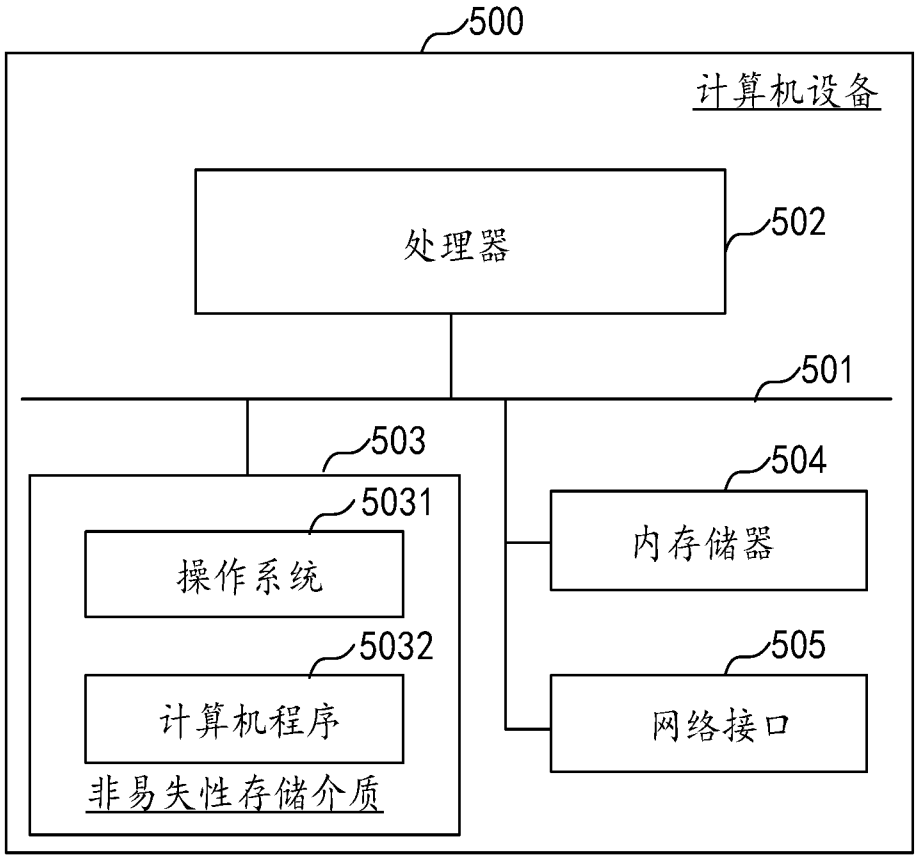

500
计算机设备
处理器
502
501
503
504
5031
操作系统
内存储器
5032
505
计算机程序
网络接口
非易失性存储介质

图 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/131757

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/30(2020.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT; CNKI; WPI; EPODOC: 文本, 生成, 模型, 训练, 编辑, 词汇表, 标签, 序列, 参数, text, generate, model, training, edit, vocabulary, label, sequence, parameter

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110263350 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD. et al.) 20 September 2019 (2019-09-20) description, paragraphs [0040]-[0076]	1-20
A	CN 110097085 A (ALIBABA GROUP HOLDING LIMITED) 06 August 2019 (2019-08-06) entire document	1-20
A	CN 109933662 A (BEIJING QIYI CENTURY SCIENCE & TECHNOLOGY CO., LTD.) 25 June 2019 (2019-06-25) entire document	1-20
A	CN 109657051 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 19 April 2019 (2019-04-19) entire document	1-20
A	US 2018329886 A1 (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 15 November 2018 (2018-11-15) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 July 2021	**21 July 2021**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/131757

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 110263350 A	20 September 2019	None	
CN 110097085 A	06 August 2019	None	
CN 109933662 A	25 June 2019	None	
CN 109657051 A	19 April 2019	WO 2020107878 A1	04 June 2020
US 2018329886 A1	15 November 2018	CN 107168952 A	15 September 2017

国际检索报告

国际申请号
PCT/CN2020/131757

A. 主题的分类

G06F 40/30(2020.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT;CNKI;WPI;EPODOC:文本，生成，模型，训练，编辑，词汇表，标签，序列，参数，text，generate，model，training，edit，vocabulary，label，sequence，parameter

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 110263350 A (腾讯科技深圳有限公司 等) 2019年 9月 20日 (2019-09-20) 说明书第[0040]-[0076]段	1-20
A	CN 110097085 A (阿里巴巴集团控股有限公司) 2019年 8月 6日 (2019-08-06) 全文	1-20
A	CN 109933662 A (北京奇艺世纪科技有限公司) 2019年 6月 25日 (2019-06-25) 全文	1-20
A	CN 109657051 A (平安科技深圳有限公司) 2019年 4月 19日 (2019-04-19) 全文	1-20
A	US 2018329886 A1 (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 2018年 11月 15日 (2018-11-15) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2021年 7月 10日	2021年 7月 21日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	焦月
传真号 (86-10)62019451	电话号码 86-(10)-53961306

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/131757

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 110263350 A	2019年 9月 20日	无	
CN 110097085 A	2019年 8月 6日	无	
CN 109933662 A	2019年 6月 25日	无	
CN 109657051 A	2019年 4月 19日	WO 2020107878 A1	2020年 6月 4日
US 2018329886 A1	2018年 11月 15日	CN 107168952 A	2017年 9月 15日