(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 10 月 21 日 (21.10.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/208612 A1

(51) 国际专利分类号:
G06F 16/33 (2019.01)

(21) 国际申请号: PCT/CN2021/078390

(22) 国际申请日: 2021 年 3 月 1 日 (01.03.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010286915.9 2020年4月13日 (13.04.2020) CN

(71) 申请人: 华为技术有限公司 (**HUAWEI TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 廖亿 (**LIAO, Yi**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 李博文 (**LI, Bowen**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 郑豪 (**ZHENG, Hao**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 蒋欣 (**JIANG, Xin**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 刘群 (**LIU, Qun**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 北京龙双利达知识产权代理有限公司 (**LONGSUN LEAD IP LTD.**); 中国北京市海淀区北清路68号院3号楼101, Beijing 100094 (CN)。

(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,

(54) **Title:** DATA PROCESSING METHOD AND DEVICE

(54) **发明名称**: 数据处理的方法与装置

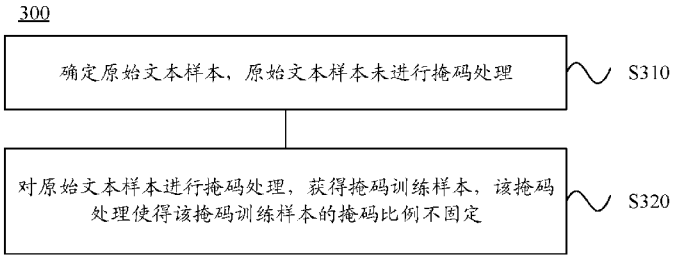

图 3

S310 Determine original text samples, the original text samples having not masked
S320 Mask the original text samples to obtain masked training samples, wherein the masking makes the masking ratios of the masked training samples not fixed

(57) **Abstract:** The present application relates to the field of artificial intelligence, specifically, the field of natural language processing, and provides a data processing method and device. The method comprises: determining original text samples, the original text samples having not masked; and masking the original text samples to obtain masked training samples, wherein the masking makes the masking ratios of the masked training samples not fixed, and the masked training samples are used for training a pretrained language model (PLM). Using masked training samples having the masking ratios not fixed to train a PLM can make the training samples for the PLM more diverse in modes, and thus, features learned by the PLM can be more diverse, the generalization ability of the PLM can be improved, and the natural language understanding ability of the trained PLM can be improved.

(57) **摘要**: 本申请提供一种数据处理的方法与装置。涉及人工智能领域，具体涉及自然语言处理领域。该方法包括：确定原始文本样本，原始文本样本未进行掩码处理；对原始文本样本进行掩码处理，获得掩码训练样本，该掩码处理使得掩码训练样本的掩码比例不固定，掩码训练样本用于训练预训练语言模型PLM。使用掩码比例不固定的掩码训练样本训练PLM，可以增强PLM的训练样本的模式多样性，从而可以使得PLM学习到的特征也较为多样，可以提高PLM的泛化能力，可以提高训练得到的PLM的自然语言理解能力。

WO 2021/208612 A1

[见续页]

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))。

数据处理的方法与装置

本申请要求于2020年4月13日提交中国专利局、申请号为202010286915.9、申请名称为“数据处理的方法与装置”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

本申请涉及人工智能领域，具体涉及一种数据处理的方法与装置。

背景技术

自然语言处理（natural language processing，NLP）是让计算机理解并处理人类自然语言的技术，是实现人工智能的重要技术手段。预训练语言模型(pertrained language model，PLM)是近年来兴起的NLP领域的一个重要的通用模型。PLM的训练方案是本领域的研究热点，PLM的训练方案具有两个改进方向：第一，提高PLM的自然语言理解能力；第二，加快模型训练速度（即加快模型收敛速度）。PLM 常用的训练方案叫做掩码语言模型(masked language model，MLM)。

MLM的训练原理是，使得PLM学习到捕捉文字上下文信息的能力。在MLM训练方案中，PLM的训练样本是被掩码处理后的文本，即部分文字被替换成特殊的标记符号（例如，[MASK]）的句子，例如，原文本是“今天是晴朗的周六”，被掩码处理后的文本为“今[MASK]是晴[MASK]的周六”；被掩码处理后的文本输入到PLM，PLM需要预测出被掩码的字分别是“天”和“朗”。PLM的训练样本可以称为掩码训练样本。

在当前的MLM训练方案中，按照固定掩码比例使用随机策略选择每个文本中的字进行掩码处理，获得掩码训练样本。这样获得的掩码训练样本会存在模式单一的问题，因此，使用这样的掩码训练样本训练PLM，会给PLM带来自然语言理解能力上的瓶颈。

发明内容

本申请提供一种数据处理的方法与装置，可以提高PLM的自然语言理解能力。

第一方面，提供一种数据处理的方法，所述方法包括：确定原始文本样本，所述原始文本样本未进行掩码处理；对所述原始文本样本进行掩码处理，获得掩码训练样本，所述掩码处理使得所述掩码训练样本的掩码比例不固定，所述掩码训练样本用于训练预训练语言模型PLM。

掩码训练样本的掩码比例包括文本级别掩码比例，和/或字级别掩码比例。

文本级别掩码比例用于表示，一个文本中被掩码处理的字占该文本中所有字的比例。

文本级别掩码比例也可以称为，句子级别的掩码比例或者文本级别掩码比例。

字级别掩码比例用于表示，一个字被掩码处理的概率。在一个文本中，每个字都具有一个字级别掩码比例。

字级别掩码比例也可称为字的掩码概率。

其中，所述掩码训练样本的掩码比例不固定包括:

所述掩码训练样本中不同样本的文本级别掩码比例不完全相同；和/或

所述掩码训练样本中任一个样本中每个字的字级别掩码比例不完全相同。

应理解，使用掩码比例不固定的掩码训练样本训练 PLM，可以增强 PLM 的训练样本的模式多样性，从而可以使得 PLM 学习到的特征也较为多样，可以提高 PLM 的泛化能力，因此，可以提高训练得到的 PLM 的自然语言理解能力。

可以采用多种实现方式，对原始文本样本进行掩码处理，获得掩码训练样本。

结合第一方面，在第一方面的一种可能的实现方式中，所述对所述原始文本样本进行掩码处理，获得掩码训练样本，包括：使用先验概率分布模型，生成所述原始文本样本中每个样本的文本级别掩码比例，所述先验概率分布模型使得所述原始文本样本中不同样本的文本级别掩码比例不完全相同；按照所述原始文本样本中每个样本的文本级别掩码比例，对相应样本进行掩码处理，获得所述掩码训练样本。

可选地，所述先验概率分布模型的概率值区间长度不小于 40%。

应理解，使用文本级别的掩码比例不固定的掩码训练样本训练 PLM，可以增强 PLM 的训练样本的模式多样性，从而可以使得 PLM 学习到的特征也较为多样，可以提高训练得到的 PLM 的自然语言理解能力。

结合第一方面，在第一方面的一种可能的实现方式中，所述对所述原始文本样本进行掩码处理，获得掩码训练样本，包括：获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，所述第一文本样本中不同字的字级别掩码比例不完全相同；根据所述第一文本样本中各个字的字级别掩码比例，对所述第一文本样本中的部分字进行掩码处理，获得所述掩码训练样本中的第一训练样本。

可许地，所述根据所述第一文本样本中各个字的字级别掩码比例，对所述第一文本样本中的部分字进行掩码处理，获得所述掩码训练样本中的第一训练样本，包括：按照字级别掩码比例从高到低的顺序，对所述第一文本样本中前 S 个字或者位于前 G%的字进行掩码处理，获得所述第一训练样本，S 为取值小于所述第一文本样本中字的总数量的正整数，G 为大于 0 且小于 100 的整数。

应理解，通过使得原始文本样本的每个样本中的字具有不完全相同的掩码比例，并在原始文本样本的掩码处理过程中，是根据每个字的字级别掩码比例来确定掩码策略，而非按照随机策略确定，这样可以减少或者避免掩码训练样本存在重复特征，从而可在一定程度上避免 PLM 在训练过程中重复性地学习相同的样本，可以实现模型快速收敛。

可以采用多种实施方式，获取原始文本样本中的第一文本样本中每个字的字级别掩码比例，以使得第一文本样本中不同字的字级别掩码比例不完全相同。

结合第一方面，在第一方面的一种可能的实现方式中，所述获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，包括：使用先验概率分布模型，生成所述第一文本样本中每个字的字级别掩码比例，所述先验概率分布模型使得所述第一文本样本中不同字的字级别掩码比例不完全相同。

结合第一方面，在第一方面的一种可能的实现方式中，所述获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，包括：将所述第一文本样本输入神经网络模

型，从所述神经网络模型的输出获得所述第一文本样本中每个字的字级别掩码比例，所述神经网络模型的输出为输入的文本中各个字的字级别掩码比例，其中，所述神经网络模型通过如下步骤进行优化学习得到，其中，i 的初始取值为 1。

1），将所述原始文本样本中第 i 个样本输入所述神经网络模型，从所述神经网络模型的输出获得所述第 i 个样本中每个字的字级别掩码比例；

2），根据所述第 i 个样本中各个字的字级别掩码比例，对所述第 i 个样本中的部分字进行掩码处理，获得所述第 i 个样本对应的训练样本；

3），将所述第 i 个样本对应的训练样本输入所述 PLM，获得所述 PLM 针对被掩码处理的字输出的损失值；

4），根据所述 PLM 针对被掩码处理的字输出的损失值，以及所述神经网络模型针对所述被掩码处理的字的输出信号，更新优化所述神经网络网络；

5），判断所述神经网络网络是否满足收敛条件，若是，转到步骤 6），若否，将 i 的取值加 1，转到步骤 1）；

6），将所述步骤 4）得到的神经网络模型作为优化学习到的所述神经网络模型。

可选地，所述步骤 3）包括：利用所述第 i 个样本对应的训练样本对所述 PLM 进行一次训练更新；将所述第 i 个样本对应的训练样本输入经过所述训练更新的所述 PLM，获得经过所述训练更新的所述 PLM 针对所述被掩码处理的字的损失值；其中，所述步骤 4）包括：根据经过所述训练更新的所述 PLM 针对所述被掩码处理的字的损失值，以及所述神经网络模型针对所述被掩码处理的字的输出信号，更新优化所述神经网络网络。

应理解，通过使用优化学习到的神经网络模型生成原始文本样本中每个样本中每个字的字级别掩码比例，相当于对掩码策略进行了优化学习，从而可以生成更优的掩码训练样本，因此，使用这样的掩码训练样本训练 PLM，可以实现 PLM 的模型快速收敛，以及 PLM 的自然语言理解能力的提升。

通过在优化学习用于生成字的字级别掩码比例时同时训练更新 PLM，可以更进一步生成更优的掩码训练样本，使用这样的掩码训练样本训练 PLM，可以实现 PLM 的模型快速收敛，以及 PLM 的自然语言理解能力的提升。

第二方面，提供一种数据处理的方法，所述方法包括：通过第一方面提供的方法获得掩码训练样本；使用所述掩码训练样本训练预训练语言模型 PLM，所述 PLM 用于预测被掩码处理的文字。

通过使用掩码比例不固定的掩码训练样本训练 PLM，可以增强 PLM 的训练样本的模式多样性，从而可以使得 PLM 学习到的特征也较为多样，可以提高 PLM 的泛化能力，因此，可以提高训练得到的 PLM 的自然语言理解能力。

第三方面，提供一种数据处理的方法，所述方法包括：确定待预测的目标文本，所述目标文本包括缺少部分文字的语句；将所述目标文本输入预训练语言模型 PLM，从所述 PLM 的输出预测所述目标文本中缺少的文字，其中，所述 PLM 通过第二方面提供的方法训练得到。

第四方面，提供一种数据处理的装置，所述装置包括第一处理单元与第二处理单元。所述第一处理单元，用于确定原始文本样本，所述原始文本样本未进行掩码处理。所述第二处理单元，用于对所述原始文本样本进行掩码处理，获得掩码训练样本，所述掩码处理

使得所述掩码训练样本的掩码比例不固定，所述掩码训练样本用于训练预训练语言模型PLM。

掩码训练样本的掩码比例包括文本级别掩码比例，和/或字级别掩码比例。详见上文描述，这里不再赘述。

结合第四方面，在第四方面的一种可能的实现方式中，所述第二处理单元用于：使用先验概率分布模型，生成所述原始文本样本中每个样本的文本级别掩码比例，所述先验概率分布模型使得所述原始文本样本中不同样本的文本级别掩码比例不完全相同；按照所述原始文本样本中每个样本的文本级别掩码比例，对相应样本进行掩码处理，获得所述掩码训练样本。

可选地，所述先验概率分布模型的概率值区间长度不小于40%。

结合第四方面，在第四方面的一种可能的实现方式中，所述第二处理单元用于，获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，所述第一文本样本中不同字的字级别掩码比例不完全相同；根据所述第一文本样本中各个字的字级别掩码比例，对所述第一文本样本中的部分字进行掩码处理，获得所述掩码训练样本中的第一训练样本。

结合第四方面，在第四方面的一种可能的实现方式中，所述第二处理单元用于，使用先验概率分布模型，生成所述第一文本样本中每个字的字级别掩码比例，所述先验概率分布模型使得所述第一文本样本中不同字的字级别掩码比例不完全相同。

结合第四方面，在第四方面的一种可能的实现方式中，所述第二处理单元用于，将所述第一文本样本输入神经网络模型，从所述神经网络模型的输出获得所述第一文本样本中每个字的字级别掩码比例，所述神经网络模型的输出为输入的文本中各个字的字级别掩码比例，其中，所述神经网络模型通过前文描述的步骤1）至步骤6）进行优化学习得到，其中，i的初始取值为1。详见前文，这里不再赘述。

结合第四方面，在第四方面的一种可能的实现方式中，所述第二处理单元用于，按照字的字级别掩码比例从高到低的顺序，对所述第一文本样本中前S个字或者位于前G%的字进行掩码处理，获得所述第一训练样本，S为取值小于所述第一文本样本中字的总数量的正整数，G为大于0且小于100的整数。

第五方面，提供一种数据处理的装置，所述装置包括：第一处理单元，用于通过第一方面提供的方法获得掩码训练样本；第二处理单元，用于使用所述掩码训练样本训练预训练语言模型PLM，所述PLM用于预测被掩码处理的文字。

第六方面，提供一种数据处理的装置，所述装置包括：第一处理单元，用于确定待预测的目标文本，所述目标文本包括缺少部分文字的语句；第二处理单元，用于将所述目标文本输入预训练语言模型PLM，从所述PLM的输出预测所述目标文本中缺少的文字，其中，所述PLM通过第二方面提供的方法训练得到。

第七方面，提供一种数据处理的装置，该装置包括：存储器，用于存储程序；处理器，用于执行存储器存储的程序，当存储器存储的程序被执行时，处理器用于执行上述第一方面、第二方面或第三方面中的方法。

第八方面，提供一种计算机可读介质，该计算机可读介质存储用于设备执行的程序代码，该程序代码包括用于执行上述第一方面、第二方面或第三方面中的方法。

第九方面，提供一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述第一方面、第二方面或第三方面中的方法。

第十方面，提供一种芯片，所述芯片包括处理器与数据接口，所述处理器通过所述数据接口读取存储器上存储的指令，执行上述第一方面、第二方面或第三方面中的方法。

可选地，作为一种实现方式，所述芯片还可以包括存储器，所述存储器中存储有指令，所述处理器用于执行所述存储器上存储的指令，当所述指令被执行时，所述处理器用于执行上述第一方面、第二方面或第三方面中的方法。

第十一方面，提供一种电子设备，该电子设备包括上述第四方面、第五方面、第六方面或第七方面提供的装置。

在本申请提供的方案中，使用掩码比例不固定的掩码训练样本训练 PLM，可以增强 PLM 的训练样本的模式多样性，从而可以使得 PLM 学习到的特征也较为多样，可以提高 PLM 的泛化能力，因此，可以提高训练得到的 PLM 的自然语言理解能力。

附图说明

图 1 是预训练语言模型(PLM)的训练原理示意图。

图 2 是本申请实施例可应用的系统架构示意图。

图 3 是本申请实施例提供的获取掩码训练样本的方法的示意性流程图。

图 4 是本申请实施例提供的获取掩码训练样本的方法的另一示意性流程图。

图 5 是本申请实施例提供的获取掩码训练样本的方法的再一示意性流程图。

图 6 是本申请实施例中原始文本样本中字的字级别掩码比例的示意图。

图 7 是本申请实施例中的用于生成字的字级别掩码比例的神经网络模型的优化学习的示意性流程图。

图 8 是本申请实施例中的用于生成字的字级别掩码比例的神经网络模型的优化学习的另一示意性流程图。

图 9 是本申请另一实施例提供的数据处理的方法的示意性流程图。

图 10 是本申请又一实施例提供的数据处理的方法的示意性流程图。

图 11 是本申请实施例提供的数据处理的装置的示意性框图。

图 12 是图 11 所示的装置的应用示意图。

图 13 是本申请实施例提供的数据处理的装置的另一示意性框图。

图 14 是本申请实施例提供的数据处理的装置的又一示意性框图。

图 15 是本申请实施例提供的一种芯片硬件结构示意图。

具体实施方式

下面将结合附图，对本申请中的技术方案进行描述。

自然语言处理（natural language processing，NLP）是让计算机理解并处理人类自然语言的技术，是实现人工智能（artificial intelligence，AI）的重要技术手段。例如，NLP 可以涵盖如下多种下游任务：情感分析、词性分析、意图分析、命名实体识别、阅读理解、逻辑推理、机器翻译或对话机器人等。预训练语言模型(pertrained language model，PLM)是近年来兴起的 NLP 领域的一个重要的通用模型。PLM 在大部分 NLP 领域的下游任务上

都有较好的效果。

PLM 常用的训练方案叫做掩码语言模型(masked language model，MLM)。MLM 的训练原理是，使得 PLM 学习到捕捉文字上下文信息的能力。

如图 1 所示，在 MLM 训练方案中，PLM 的训练样本是被掩码处理后的文本，即部分文字被替换成特殊的标记符号（例如，[MASK]）的句子，例如，原文本是“今天是晴朗的周六”，被掩码处理后的文本为“今[MASK]是晴[MASK]的周六”；被掩码处理后的文本输入到 PLM，PLM 需要预测出被掩码的字分别是“天”和“朗”。PLM 的训练样本可以称为掩码训练样本。在一个文本（例如，句子）中，对于被掩码处理的字，未被掩码处理的字是它的上下文信息，PLM 通过预测被掩码处理的字，学习到了捕捉文字上下文信息的能力。因此按照 MLM 训练方案训练完成的 PLM 具有理解自然语言深度语义的能力，可用于一系列 NLP 相关的下游任务。

在当前的 MLM 训练方案中，按照固定掩码比例使用随机策略选择每个文本中的字进行掩码处理，获得掩码训练样本。

如前文描述，PLM 的训练方案具有两个改进的方向：第一，提高 PLM 的自然语言理解能力；第二，加快模型训练速度（即加快模型收敛速度）。使用现有的 MLM 训练方案获得的掩码训练样本训练 PLM，会给 PLM 带来自然语言理解能力上的瓶颈。原因如下。

在当前的 MLM 训练方案中，按照固定掩码比例使用随机策略选择每个文本中的字进行掩码处理，获得掩码训练样本。例如，将固定掩码比例记为 r，对于每个文本，随机地选取 r*N 个字进行掩码处理，N 表示文本包含的字的数量（若将文本视为句子，则 N 表示该句子的长度）。例如，假设某个句子的长度 N=100，在掩码比例 r=15%的情况下，随机地选择该句子中 100*15%=15 个字替换成[MASK]。

在当前的 MLM 训练方案中，掩码训练样本是按照固定掩码比例使用随机策略得到的，这会导致 PLM 的训练样本的模式较为单一，从而使得 PLM 学习到的特征也较为固定，导致 PLM 在泛化能力上有所欠缺，因此，给训练得到的 PLM 带来自然语言理解能力上的瓶颈。

针对上述问题，本申请实施例提出一种生成 PLM 的掩码训练样本的方案，可以提高训练得到的 PLM 的自然语言理解能力。换言之，采用本申请实施例获得的掩码训练样本训练 PLM，可以克服现有技术存在的 PLM 在自然语言理解能力上的瓶颈。

图 2 为本申请实施例可应用的系统架构的示意图。该系统可以包括数据收集设备 21、服务器设备 22 与客户端设备 23。数据收集设备 21、服务器设备 22 与客户端设备 23 通过通信网络连接。

数据收集设备 21 用于，获取原始文本样本（例如，大量的句子），并将原始文本样本传输至服务器设备 22。

数据收集设备 21 可以通过多种途径，获取原始文本样本。例如，通过人工输入和/或网络查找等方式获取。

服务器设备 22 用于，使用本申请实施例提供的方案获得掩码训练数据，进而获得训练后的 PLM，可以将 PLM 输出给客户端设备 23。

客户端设备 23 用于，使用服务器设备 22 训练得到的 PLM 进行自然语言理解与处理，例如，进行下列 NLP 下游任务中的任一种或多种：情感分析、词性分析、意图分析、命

名实体识别、阅读理解、逻辑推理、机器翻译或对话机器人等。

需要说明的是，图 2 仅为示例而非限定。

例如，数据收集设备 21 是可选的。例如，数据收集设备 21 的操作可以在服务器设备 22 上执行。

又如，客户端设备 23 是可选的。例如，客户端设备 23 的操作可以在服务器设备 22 上执行。

为了便于理解与描述，对本文中的术语做如下解释。

1、原始文本样本

原始文本样本表示，待进行掩码处理的文本的集合。原始文本样本中的每个样本表示一个文本（或称为文本语句）。例如，原始文本样本是多个文本句子的集合。

2、掩码训练样本

掩码训练样本表示，被掩码处理后的文本的集合。掩码训练样本中的每个样本表示一个经过掩码处理后的文本。

3、掩码比例

本申请实施例中涉及的掩码比例包括文本级别掩码比例与字级别掩码比例。

文本级别掩码比例用于表示，一个文本中被掩码处理的字占该文本中所有字的比例。

文本级别掩码比例也可以称为，句子级别的掩码比例或者样本级别掩码比例。

字级别掩码比例用于表示，一个字被掩码处理的概率。在一个文本中，每个字都具有一个字级别掩码比例。

本文中涉及的表述“掩码训练样本的掩码比例”、“原始文本样本的掩码比例”，包括，文本级别掩码比例，和/或，字级别掩码比例。

字级别掩码比例也可称为字的掩码概率。

图 3 为本申请实施例提供的数据处理的方法 300 的示意性流程图。例如，该方法 300 可以由图 2 中的服务器设备 22 执行。该方法 300 包括步骤 S310 与步骤 S320。

S310，确定原始文本样本，该原始文本样本未进行掩码处理。

作为示例，原始文本样本中的一个样本为“今天是晴朗的周六”。

可以通过多种途径，获取原始文本样本。例如，通过人工输入和/或网络查找等方式获取。

S320，对原始文本样本进行掩码处理，获得掩码训练样本，该掩码处理使得该掩码训练样本的掩码比例不固定。掩码训练样本用于训练 PLM，PLM 用于预测被掩码处理的文字。

作为示例，原始文本样本中的一个样本为“今天是晴朗的周六”，该样本被掩码处理后得到对应的训练样本为“今[MASK]是晴[MASK]的周六”。

掩码训练样本的掩码比例包括文本级别掩码比例，和/或字级别掩码比例。

例如，掩码训练样本的掩码比例为文本级别掩码比例。

在本例中，掩码训练样本的掩码比例不固定指的是，掩码训练样本中不同样本的文本级别掩码比例不完全相同。至于每个样本中不同字的字级别掩码比例，可以相同，或不同，或不完全相同。

作为示例，在掩码训练样本中，包括第一样本与第二样本，第一样本的文本级别掩码

比例为 15%，第二样本的文本级别掩码比例为 20%，假设第一样本包含的字的总数量与第二样本包含的字的总数量均为 100，则第一样本中 15 个字被掩码处理了，第二样本中 20 个字被掩码处理了。

又例如，掩码训练样本的掩码比例为字级别掩码比例。

在本例中，掩码训练样本的掩码比例不固定指的是，掩码训练样本中每个样本中的不同字的字级别掩码比例不完全相同。至于掩码训练样本中不同样本的文本级别掩码比例，可以相同，或不同，或不完全相同。

再例如，掩码训练样本的掩码比例包括文本级别掩码比例与字级别掩码比例。

在本例中，掩码训练样本的掩码比例不固定指的是，掩码训练样本中不同样本的文本级别掩码比例不完全相同，并且掩码训练样本中每个样本中的不同字的字级别掩码比例不完全相同。

例如，步骤 S320 包括：获取掩码策略，该掩码策略能够使得原始文本样本的掩码比例不固定；根据该掩码策略，判断原始文本样本中每个样本中每个字是否需要进行掩码处理，若是，将其替换为标记符号（例如[MASK]），若否，不作处理，最终获得掩码训练样本。

其中，可以采用多种方式获取该掩码策略，下文将描述，这里暂不详述。

应理解，使用掩码比例不固定的掩码训练样本训练 PLM，可以增强 PLM 的训练样本的模式多样性，从而可以使得 PLM 学习到的特征也较为多样，可以提高 PLM 的泛化能力，因此，可以提高训练得到的 PLM 的自然语言理解能力。

需要说明的是，步骤 S310 是可选的。例如，在实际应用中，在原始文本样本为已知或现成的情况下，可以直接对原始文本样本进行掩码处理获得掩码训练样本，即直接执行步骤 S310，而无需执行步骤 S310。

在步骤 S320 中，可以采用多种实现方式，对原始文本样本进行掩码处理，获得掩码训练样本。换言之，可以采用多种方式获取原始文本样本的掩码策略。

可选地，作为一种实现方式，如图 4 所示，步骤 S320 包括步骤 S321 与步骤 S322。

S321，使用先验概率分布模型，生成原始文本样本中每个样本的文本级别掩码比例，先验概率分布模型使得原始文本样本中不同样本的文本级别掩码比例不完全相同。

换句话说，使用先验概率分布模型，为原始文本样本中每个样本，生成掩码比例。

作为示例，针对原始文本样本中第 i 个样本，使用先验概率分布模型生成一个概率，并将该概率作为第 i 个样本的文本级别掩码比例，i 为 1，…，M，M 表示原始文本样本的样本数量。

先验概率分布模型生成的概率服从某种概率分布，因此，使用先验概率分布模型生成的掩码比例是动态变化的，而非固定不变的。也就是说，使用先验概率分布模型生成的原始文本样本中每个样本的文本级别掩码比例不是完全相同的，例如，所有样本的文本级别掩码比例不同，或者，至少一部分样本中不同样本的文本级别掩码比例不同。

作为示例，将先验概率分布模型记为 P(r)，r 表示概率，r 的取值区间可以为 0%到 100%之间，则使用 P(r)生成的掩码比例的取值区间为 0%到 100%之间。

先验概率分布模型服从的概率分布可以是任意连续或离散的概率分布。例如，先验概率分布模型服从的概率分布为均匀分布或高斯分布等。高斯分布也可以称为正态分布

（normal distribution）。

可选地，先验概率分布模型服从的概率分布为截断高斯分布（也称为截断正态分布（truncated normal distribution））。

可以根据应用需求，设置截断高斯分布的变量限制范围。

S322，按照原始文本样本中每个样本的文本级别掩码比例，对相应样本进行掩码处理，获得掩码训练样本。

作为示例，假设在步骤 S321 中分别为原始文本样本中的文本样本 1 与文本样本 2 生成的掩码比例为 r1 与 r2，则在步骤 S322 中，按照掩码比例 r1 对文本样本 1 进行掩码处理，获得文本样本 1 对应的训练样本（记为训练样本 1），按照掩码比例 r2 对文本样本 2 进行掩码处理，获得文本样本 2 对应的训练样本（记为训练样本 2）。可以理解到，若 r1 与 r2 不同，则训练样本 1 与训练样本 2 的掩码比例不同。

在如上示例中，假设文本样本 1 包含的字的总数量为 N1，获得文本样本 1 对应的训练样本的一种实现方式为：按照掩码比例 r1，使用随机策略选择文本样本 1 中的 r1*N1 个子进行掩码处理，获得文本样本 1 对应的训练样本。或者，还可以使用其它可行的策略选择文本样本 1 中的 r1*N1 个子进行掩码处理，获得文本样本 1 对应的训练样本。本申请实施例对此不作限定。

应理解，在图 4 所示实施例中，掩码训练样本的文本级别掩码比例不完全相同，或者说，文本级别掩码比例不固定。

在本申请实施例中，使用文本级别掩码比例不固定的掩码训练样本训练 PLM，可以增强 PLM 的训练样本的模式多样性，从而可以使得 PLM 学习到的特征也较为多样，可以提高训练得到的 PLM 的自然语言理解能力。

可选地，在图 4 所示的实施例中，先验概率分布模型的概率值区间长度不低于 40%。

例如，先验概率分布模型的概率值区间为 0%~40%。

通过仿真实验表明，使用本实施例获得的掩码训练样本训练得到的 PLM 具有按照随机顺序生成自然语言的能力。

作为示例，使用本实施例获得的掩码训练样本训练得到的 PLM 可以按照如表 1 所示的随机顺序生成方式生成自然语言。

通常意义上，自然语言文字生成的顺序是从左往右依次生成，而使用本实施例获得的掩码训练样本训练得到的 PLM 可以每次指定下一个生成的文字的坐标，在顺序随机的情况下，依然可以生成流畅的文本。

表 1

步骤	预测索引	序列的状态							
0	n/a	_	_	_	_	_	_	_	_
1	3	_	_	a	_	_	_	_	_
2	7	_	_	a	_	_	_	random	_
3	1	This	_	a	_	_	_	random	_
4	2	This	is	a	_	_	_	random	_
5	4	This	is	a	sentence	_	_	random	_
6	6	This	is	a	sentence	_	in	random	_
7	5	This	is	a	sentence	generated	in	random	_
8	8	This	is	a	sentence	generated	in	random	order

在表 1 中，序列“This is a sentence generated in random order”中每个字的生成顺序为 3→7→1→2→4→6→5→8。

可选地，作为另一种实现方式，如图 5 所示，步骤 S320 包括：步骤 S323 与步骤 S324。通过步骤 S323 与步骤 S324，可以获得原始文本样本中第一文本样本对应的第一训练样本。

为了便于描述而非限定，在如图 5 所示的实施例中，以第一文本样本为例说明原始文本样本中每个样本。也就是说，下文中对第一文本样本的描述适用于原始文本样本中的每个样本。

S323，获取原始文本样本中的第一文本样本中每个字的字级别掩码比例，第一文本样本中不同字的字级别掩码比例不完全相同。

第一文本样本中不同字的字级别掩码比例不完全相同，表示，第一文本样本中至少有两个字的字级别掩码比例不同。

可选地，第一文本样本中不同字的字级别掩码比例均不同。

可选地，第一文本样本中有部分字的字级别掩码比例不同，有部分字的字级别掩码比例相同。

作为示例，假设第一文本样本为“今天是晴朗的周六”，在步骤 S323 中，获取的“今天是晴朗的周六”中每个字的字级别掩码比例的分布示意图如图 6 所示。在图 6 的示例中，第一文本样本中所有字的字级别掩码比例都不同。

字的字级别掩码比例表示，这个字被掩码处理的概率。

S324，根据第一文本样本中各个字的字级别掩码比例，对第一文本样本中的部分字进行掩码处理，获得掩码训练样本中的第一训练样本。

对第一文本样本中的部分字进行掩码处理，指的是，对第一文本样本中掩码比例较大的字进行掩码处理。

基于第一文本样本中各个字的字级别掩码比例，可以采用多种方式，对第一文本样本进行掩码处理，获得所述第一训练样本。

可选地，作为一种方式，步骤 S324 包括：按照字级别掩码比例从高到低的顺序，对第一文本样本中前 S 个字进行掩码处理，获得第一训练样本，S 为取值小于第一文本样本中字的总数量的正整数。

作为示例，还以图 6 为例，第一文本样本为“今天是晴朗的周六”，第一文本样本中各个字的字级别掩码比例如图 6 所示，假设 S 的取值为 2，则对第一文本样本中掩码比例最

大的 2 个字“朗”与“天”进行掩码处理，获得第一训练样本“今[MASK]是晴[MASK]的周六”。

可选地，作为另一种方式，步骤 S324 包括：按照字级别掩码比例从高到低的顺序，对第一文本样本中位于前 G%的字进行掩码处理，获得第一训练样本，G 为大于 0 且小于 100 的整数。

作为示例，还以图 6 为例，第一文本样本为“今天是晴朗的周六”，第一文本样本中各个字的字级别掩码比例如图 6 所示，假设 G 的取值为 25，则按照字级别掩码比例从高到低的顺序，对第一文本样本中位于前 25%的字，即“朗”与“天”进行掩码处理，获得第一训练样本“今[MASK]是晴[MASK]的周六”。

可选地，作为又一种方式，步骤 S324 包括：对第一文本样本中掩码比例达到 D 的字进行掩码处理，获得所述第一训练样本，D 为大于 0 且小于 1 的小数，并且 D 小于第一文本样本中字级别掩码比例最小的字的字级别掩码比例。

字的字级别掩码比例达到 D，表示，字的字级别掩码比例大于或等于 D。

作为示例，还以图 6 为例，第一文本样本为“今天是晴朗的周六”，第一文本样本中各个字的字级别掩码比例如图 6 所示，假设只有“朗”与“天”的掩码比例达到 D，则对“朗”与“天”进行掩码处理，获得第一训练样本“今[MASK]是晴[MASK]的周六”。

应理解，在图 5 所示实施例中，掩码训练样本的字级别掩码比例不完全相同，或者说，字的字级别掩码比例不固定。

如前文描述，现有技术按照固定掩码比例使用随机策略选择每个文本中的字进行掩码处理，获得掩码训练样本，而随机产生的掩码训练样本可能具有重复的特征，使用这样的掩码训练样本训练 PLM 会导致 PLM 在训练过程中重复性地学习同样的训练样本，从而无法保证模型快速收敛。

在本申请实施例中，使得原始文本样本的每个样本中的字具有不完全相同的掩码比例，并在原始文本样本的掩码处理过程中，是根据每个字的字级别掩码比例来确定掩码策略，而非按照随机策略确定，这样可以减少或者避免掩码训练样本存在重复特征，从而可在一定程度上避免 PLM 在训练过程中重复性地学习相同的样本，可以实现模型快速收敛。

在步骤 S323，可以采用多种实施方式，获取原始文本样本中的第一文本样本中每个字的字级别掩码比例，以使得第一文本样本中不同字的字级别掩码比例不完全相同。

可选地，作为一种实施方式，步骤 S323 包括：使用先验概率分布模型，生成第一文本样本中每个字的字级别掩码比例，先验概率分布模型使得第一文本样本中不同字的字级别掩码比例不完全相同。

换句话说，使用先验概率分布模型，为第一文本样本中的每个字，生成掩码比例。例如，针对第一文本样本中的第 j 个字，使用先验概率分布模型生成一个概率，并将该概率作为第 j 个字的字级别掩码比例，j 为 1，…，N1，N 表示第一文本样本中包含的字的总数量。

作为示例，假设第一文本样本为“今天是晴朗的周六”，在步骤 S323 中，使用先验概率分布模型获取的“今天是晴朗的周六”中每个字的字级别掩码比例的分布示意图如图 6 所示。

应理解，先验概率分布模型生成的概率服从某种概率分布，例如，先验概率分布模型的概率取值区间为 0%到 100%之间，因此，使用先验概率分布模型生成的掩码比例是动态

变化的，而非固定不变的。

作为示例，将先验概率分布模型记为 P(r)，r 表示概率，r 的取值区间可以为 0%到 100%之间，则使用 P(r)生成的掩码比例的取值区间为 0%到 100%之间。

这里提及的先验概率分布模型与前文在步骤 S321 中提及的先验概率分布模型相同，关于先验概率分布模型的说明详见前文描述，这里不再赘述。

可选地，作为另一种实施方式，步骤 S323 包括：将第一文本样本输入神经网络模型，从该神经网络模型的输出获得第一文本样本中每个字的字级别掩码比例，该神经网络模型的输出为输入的文本中各个字的字级别掩码比例。该神经网络模型是优化学习得到的，该神经网络模型的学习优化过程如图 7 所示。在图 7 中，i 的初始取值为 1。

1），将原始文本样本中第 i 个样本输入神经网络模型，从神经网络模型的输出获得第 i 个样本中每个字的字级别掩码比例。

从神经网络模型的输出获得第 i 个样本中每个字的字级别掩码比例，表示，可以根据神经网络模型针对第 i 个样本中每个字的输出信号，获得每个字的字级别掩码比例。

例如，该神经网络模型可以针对每个字输出一个损失值（loss），该损失值可以映射到一个掩码比例。如 8 所示，神经网络模型针对样本“今天是晴朗的周六”中每个字可以输出一个损失值，例如 loss_0 表示神经网络模型针对样本中的“今”输出的损失值，loss_1 至 loss_7 的含义类似，这里不再赘述。

在本例中，获取一个字的概率掩码的方法为，根据神经网络模型针对该字输出的损失值，以及损失值与掩码比例之间的映射关系，获得该字的字级别掩码比例。

在本例中，神经网络模型输出的损失值与掩码比例之间的映射关系可以根据应用需求设计，本申请对此不作限定。

又例如，该神经网络模型可以针对每个字直接输出该字的字级别掩码比例。

在本例中，可以直接根据神经网络模型针对第 i 个样本中每个字的输出信号，获得每个字的字级别掩码比例。

2），根据第 i 个样本中各个字的字级别掩码比例，对第 i 个样本中的部分字进行掩码处理，获得第 i 个样本对应的训练样本。

步骤 2）可以对应步骤 S324，关于根据第 i 个样本中各个字的字级别掩码比例对第 i 个样本进行掩码处理的实现方式详见前文，这里不再赘述。

3），将第 i 个样本对应的训练样本输入 PLM，获得 PLM 针对被掩码处理的字输出的损失值。

因为步骤 3）获取的 PLM 针对被掩码处理的字输出的损失值是作为神经网络模型的反馈信号，因此，可以将步骤 3）获取的 PLM 针对被掩码处理的字输出的损失值称为反馈信号。

PLM 针对输入的掩码训练数据可以预测被掩码处理的字，还可以输出针对被掩码处理的字的损失值（loss’）。

可选地，PLM 可以是参数固定的模型。下文将描述，这里暂不详述。

4），根据 PLM 针对被掩码处理的字输出的损失值，以及该神经网络模型针对被掩码处理的字的输出信号，更新优化神经网络网络。

根据 PLM 针对被掩码处理的字输出的损失值获得第一信号，根据该神经网络模型针

对被掩码处理的字的输出信号获得第二信号，第一信号与第二信号是含义相同的信号（即能够进行比较的信号）；通过第一信号与第二信号之间的差值，对该神经网络模型进行优化更新。

可选地，PLM 针对被掩码处理的字输出的损失值（记为输出信号 1），与该神经网络模型针对被掩码处理的字的输出信号（记为输出信号 2）是相同含义的信号，即可直接进行比较，则可以直接根据输出信号 1 与输出信号 2 之间的差值，对该神经网络模型进行优化更新。

作为示例，如图 8 所示，PLM 针对被掩码处理的字“天”与“朗”分别输出损失值 loss_1’与 loss_4’，该神经网络模型针对被掩码处理的字的输出信号也为损失值（如图 8 中所示的 loss_1 与 loss_4），则可以通过比较 PLM 输出的损失值与神经网络模型输出的损失值，对该神经网络模型进行优化更新。

可选地，PLM 针对被掩码处理的字输出的损失值（记为输出信号 1），与该神经网络模型针对被掩码处理的字的输出信号（记为输出信号 2）不是相同含义的信号，即无法进行比较，这种情况下，可以将输出信号 1 与输出信号 2 中的一方处理为与另一方相同含义的信号，然后进行比较。

作为一个示例，该神经网络模型针对被掩码处理的字的输出信号为掩码比例，PLM 针对被掩码处理的字输出的损失值与掩码比例具有映射关系，这种情况下，可以先根据该映射关系将 PLM 针对被掩码处理的字输出的损失值换算为掩码比例，然后将其与该神经网络模型针对被掩码处理的字输出的掩码比例进行比较，从而对该神经网络模型进行优化更新。

关于损失值与掩码比例的映射关系的建立方法，本申请实施例对此不作限定。例如，分别用每个损失值除以同一个较大的数值（大于所有损失值），求得的比值作为各个损失值映射的掩码比例。

5），判断神经网络网络是否满足收敛条件，若是，转到步骤 6），若否，将 i 的取值加 1，转到步骤 1）。

6），将步骤 4）得到的神经网络模型作为优化学习到的神经网络模型。

作为一个示例，假设原始文本样本中的一个样本为“今天是晴朗的周六”，使用这个样本对该神经网络模型进行一次优化学习（一次迭代过程）的流程示意图如图 8 所示。

S810，将样本“今天是晴朗的周六”输入神经网络模型，从神经网络模型的输出获得样本“今天是晴朗的周六”中每个字的字级别掩码比例。S810 对应图 7 中的步骤 1）。

从神经网络模型的输出获得样本“今天是晴朗的周六”中每个字的字级别掩码比例，表示，可以根据神经网络模型针对样本中每个字的输出信号，获得每个字的字级别掩码比例。

如 8 所示，神经网络模型针对样本“今天是晴朗的周六”中每个字输出一个损失值，例如 loss_0 表示神经网络模型针对样本中的“今”输出的损失值，loss_1 至 loss_7 的含义类似，这里不再赘述。神经网络模型针对每个字输出的损失值与掩码比例具有映射关系，如图 8 中所示的 loss_0 至 loss_7 分别映射一个掩码比例，则可以根据损失值与该映射关系，获得每个字的字级别掩码比例，当然可以获得被掩码处理的字的字级别掩码比例。

S820，对样本“今天是晴朗的周六”中掩码比例达到条件的字“天”与“朗”进行掩码处理，获得样本“今天是晴朗的周六”对应的训练样本“今[MASK]是晴[MASK]的周六”。S820 对应

图 7 中的步骤 2）。

S830，将掩码训练样本“今[MASK]是晴[MASK]的周六”输入 PLM，从 PLM 的输出获得该掩码训练样本中被掩码处理的字的预测结果，还可以获得 PLM 针对被掩码处理的字（即“天”与“朗”）的输出信号。S830 对应图 7 中的步骤 3）。

S840，根据神经网络模型针对被掩码处理的字（即“天”与“朗”）的输出信号，以及 PLM 针对被掩码处理的字输出的损失值，更新优化该神经网络网络。S840 对应图 7 中的步骤 4）。

可选地，在图 7 所示实施例中，PLM 是基于掩码训练样本进行参数实时更新的模型。

例如，步骤 3）包括：利用第 i 个样本对应的训练样本对 PLM 进行一次训练更新；将第 i 个样本对应的训练样本输入经过训练更新的 PLM，获得经过训练更新的 PLM 针对被掩码处理的字输出的损失值。

利用第 i 个样本对应的训练样本对 PLM 进行一次训练更新，表示，利用第 i 个样本对应的训练样本对 PLM 进行一次训练，使得 PLM 的参数发生更新。

其中，步骤 4）包括：根据经过训练更新的 PLM 针对被掩码处理的字输出的损失值，以及神经网络模型针对被掩码处理的字的输出信号，更新优化神经网络网络。

通过在优化学习用于生成字的字级别掩码比例时同时训练更新 PLM，可以更进一步生成更优的掩码训练样本，使用这样的掩码训练样本训练 PLM，可以实现 PLM 的模型快速收敛，以及 PLM 的自然语言理解能力的提升。

应理解，在图 7 或图 8 所示实施例中，实现神经网络模型的优化学习，相当于实现了掩码策略的优化学习。

在本实施例中，使用优化学习到的神经网络模型生成原始文本样本中每个样本中每个字的字级别掩码比例，相当于对掩码策略进行了优化学习，从而可以生成更优的掩码训练样本，因此，使用这样的掩码训练样本训练 PLM，可以实现 PLM 的模型快速收敛，以及 PLM 的自然语言理解能力的提升。

相比于图 7 或图 8 所示实施例中通过可优化学习的神经网络模型获取掩码策略，前文描述根据概率分布模型获取掩码策略的实施例，可以视为，根据预设模型（或根据经验）获取掩码策略。

在本申请实施例中，可以通过一定的方式控制掩码训练样本的生成。例如，通过控制文本（即句子）的掩码比例（即文本级别掩码比例）来控制掩码训练样本的生成；又例如，通过控制文本中字的字级别掩码比例来控制掩码训练样本的生成。因此，在本申请实施例中，可以通过控制的方式生成掩码训练样本，而非随机生成掩码训练样本，从而可以通过对掩码训练样本的控制来实现 PLM 的自然语言理解能力的提高以及 PLM 收敛速度的提升。

本申请实施例提供的获得 PLM 的掩码训练样本的方案可应用于所有基于 MLM 的 PLM。

如图 9 所示，本申请实施例还提供一种训练 PLM 的方法 900。例如，该方法 900 可以由图 2 中的服务器设备 22 执行。该方法 900 包括步骤 S910 与步骤 S920。

S910，通过上文实施例中的方法 300 获得掩码训练样本。

S920，使用掩码训练样本训练预训练语言模型 PLM，PLM 用于预测被掩码处理的文字。

在本申请实施例中，使用掩码比例不固定的掩码训练样本训练 PLM，可以增强 PLM 的训练样本的模式多样性，从而可以使得 PLM 学习到的特征也较为多样，可以提高 PLM 的泛化能力，因此，可以提高训练得到的 PLM 的自然语言理解能力。

进一步地，通过使得原始文本样本的每个样本中的字具有不完全相同的掩码比例，并在原始文本样本的掩码处理过程中，是根据每个字的字级别掩码比例来确定掩码策略，而非按照随机策略确定，这样可以减少或者避免掩码训练样本存在重复特征，从而可在一定程度上避免 PLM 在训练过程中重复性地学习相同的样本，可以实现模型快速收敛。

再进一步，在原始文本样本的掩码处理过程中，通过使用优化学习到的神经网络模型生成原始文本样本中每个样本中每个字的字级别掩码比例，然后根据每个字的字级别掩码比例来确定掩码策略，可以生成更优的掩码训练样本，因此，使用这样的掩码训练样本训练 PLM，可以实现 PLM 的模型快速收敛，以及 PLM 的自然语言理解能力的提升。

本申请实施例提供的训练 PLM 的方案可应用于所有基于 MLM 的 PLM。

如图 10 所示，本申请实施例还提供一种数据处理的方法 1000。例如，该方法 1000 可以由图 2 中的服务器设备 22 或客户端设备 23 执行。该方法 1000 包括步骤 S1010 与步骤 S1020。

S1010，确定待预测的目标文本，目标文本包括缺少部分文字的语句。

目标文本包括缺少部分文字的语句，也可以表述为目标文本包括缺少上下文信息的语句。

S1020，将目标文本输入 PLM，从该 PLM 的输出预测得到目标文本中缺少的文字，该 PLM 为通过上文实施例提供的方法 900 训练得到 PLM。

实验仿真表明，使用本申请实施例提供的方案训练得到的 PLM，在自然语言理解的相关下游任务上具有显著提升的效果。

作为一个示例，表 2 示出采用本申请实施例提供的方案训练得到的 PLM（记为 u-PMLM-A）相较于现有的 BERT(A)模型在分数上具有提升。

表2

Model	COLA	SST2	MRPC	STSB	QQP	MNLI-m/mm	QNLI	RTE	AX	AVG.
BERT(A)	52.1	93.5	88.9/84.8	87.1/85.8	71.2/89.2	84.6/83.4	90.5	66.4	34.2	78.3
u-PMLM-A	56.5	94.3	88.8/84.4	87.0/85.9	71.4/89.2	84.5/83.5	91.8	66.1	37.0	79.0

表 2 的横向表头中的 COLA、SST2、MRPC、STSB、QQP、MNL1-m/mm、QNLI、RTE、AX 分别表示自然语言处理任务集合（GLUE）中的一个子任务的名称，AVG.表示这些子任务的平均分。

本文中描述的各个实施例可以为独立的方案，也可以根据内在逻辑进行组合，这些方案都落入本申请的保护范围中。

上文描述了本申请提供的方法实施例，下文将描述本申请提供的装置实施例。应理解，装置实施例的描述与方法实施例的描述相互对应，因此，未详细描述的内容可以参见上文方法实施例，为了简洁，这里不再赘述。

图 11 为本申请实施例提供的数据处理的装置 1100 的示意性框图。装置 1100 包括掩码生成模块 1110 与 PLM 模型训练模块 1120。

掩码生成模块 1110，用于通过本申请实施例提供的方法 300，对原始文本样本进行掩

码处理，获得掩码训练数据。

例如，在一些实施例中，掩码生成模块 1110 包括如图 8 所示的神经网络模型。

PLM 模型训练模块 1120 用于，利用掩码生成模块 1110 获得的掩码训练数据进行 PLM 模型训练。

作为示例，如图 12 所示，将文本样本“今天是晴朗的周六”输入掩码生成模块 1110，掩码生成模块 1110 输出训练样本“今[MASK]是晴[MASK]的周六”；将训练样本“今[MASK]是晴[MASK]的周六”输入 PLM 模型训练模块 1120，PLM 模型训练模块 1120 输出被掩码处理的字的预测结果“天”与“朗”。

如图 13 所示，本申请实施例还提供一种数据处理的装置 1300，装置 1300 用于执行上文方法实施例。装置 1300 包括第一处理单元 1310 与第二处理单元 1320。

可选地，作为第一种设计，装置 1300 用于执行上文方法实施例中的方法 300。第一处理单元 1310，用于确定原始文本样本，原始文本样本未进行掩码处理；第二处理单元 1320，用于对原始文本样本进行掩码处理，获得掩码训练样本，掩码处理使得掩码训练样本的掩码比例不固定，掩码训练样本用于训练 PLM。

掩码训练样本的掩码比例包括文本级别掩码比例，和/或字级别掩码比例。详见上文描述，这里不再赘述。

可选地，第二处理单元 1320 用于：使用先验概率分布模型，生成原始文本样本中每个样本的文本级别掩码比例，先验概率分布模型使得原始文本样本中不同样本的文本级别掩码比例不完全相同；按照原始文本样本中每个样本的文本级别掩码比例，对相应样本进行掩码处理，获得掩码训练样本。

可选地，先验概率分布模型的概率值区间长度不小于 40%。

可选地，第二处理单元 1320 用于，获取原始文本样本中的第一文本样本中每个字的字级别掩码比例，第一文本样本中不同字的字级别掩码比例不完全相同；根据第一文本样本中各个字的字级别掩码比例，对第一文本样本中的部分字进行掩码处理，获得掩码训练样本中的第一训练样本。

可选地，第二处理单元 1320 用于，使用先验概率分布模型，生成第一文本样本中每个字的字级别掩码比例，先验概率分布模型使得第一文本样本中不同字的字级别掩码比例不完全相同。

可选地，第二处理单元 1320 用于，将第一文本样本输入神经网络模型，从神经网络模型的输出获得第一文本样本中每个字的字级别掩码比例，神经网络模型的输出为输入的文本中各个字的字级别掩码比例。其中，神经网络模型通过如图 7 所示的步骤进行优化学习得到，其中，i 的初始取值为 1。详见前文描述，这里不再赘述。

可选地，第二处理单元 1320 用于，按照字级别掩码比例从高到低的顺序，对第一文本样本中前 S 个字或者位于前 G%的字进行掩码处理，获得第一训练样本，S 为取值小于第一文本样本中字的总数量的正整数，G 为大于 0 且小于 100 的整数。

第一种设计下的装置 1300 可以被设置在装置 1100 中的掩码生成模块 1110 中。

在一些实施例中，装置 1100 中的掩码生成模块 1110 包括第一种设计下的装置 1300。

可选地，作为第二种设计，装置 1300 用于执行上文方法实施例中的方法 900。第一处理单元 1310，用于通过上文方法实施例中的方法 300 获得掩码训练样本；第二处理单

元 1320，用于使用掩码训练样本训练预训练语言模型 PLM，PLM 用于预测被掩码处理的文字。

第二种设计下的装置 1300 可以被设置在装置 1100 中的 PLM 模型训练模块 1120 中。

在一些实施例中，装置 1100 中的 PLM 模型训练模块 1120 包括第二种设计下的装置 1300。

可选地，作为第三种设计，装置 1300 用于执行上文方法实施例中的方法 1000。第一处理单元 1310，用于确定待预测的目标文本，目标文本包括缺少部分文字的语句；第二处理单元 1320，用于将目标文本输入预训练语言模型 PLM，从 PLM 的输出预测目标文本中缺少的文字，其中，PLM 通过上文方法实施例中的方法 900 训练得到。

应理解，第一处理单元 1310 与第二处理单元 1320 可以通过处理器实现。

如图 14 所示，本申请实施例还提供一种数据处理的装置 1400。该装置 1400 包括处理器 1410，处理器 1410 与存储器 1420 耦合，存储器 1420 用于存储计算机程序或指令，处理器 1410 用于执行存储器 1420 存储的计算机程序或指令，使得上文方法实施例中的方法被执行。

可选地，如图 14 所示，该装置 1400 还可以包括存储器 1420。

可选地，如图 14 所示，该装置 1400 还可以包括数据接口 1430，数据接口 1430 用于与外界进行数据的传输。

可选地，作为一种方案，该装置 1400 用于实现上文实施例中的方法 300。

可选地，作为另一种方案，该装置 1400 用于实现上文实施例中的方法 900。

可选地，作为又一种方案，该装置 1400 用于实现上文实施例中的方法 1000。

本申请实施例还提供一种计算机可读介质，该计算机可读介质存储用于设备执行的程序代码，该程序代码包括用于执行上述实施例的方法。

本申请实施例还提供一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述实施例的方法。

本申请实施例还提供一种芯片，该芯片包括处理器与数据接口，处理器通过数据接口读取存储器上存储的指令，执行上述实施例的方法。

可选地，作为一种实现方式，该芯片还可以包括存储器，存储器中存储有指令，处理器用于执行存储器上存储的指令，当指令被执行时，处理器用于执行上述实施例中的方法。

本申请实施例还提供一种电子设备，该电子设备包括上述实施例中的装置 1100。

本申请实施例还提供一种电子设备，该电子设备包括第一种设计下的装置 1300，或第二种设计下的装置 1300，或第三种设计下的装置 1300。

本申请实施例还提供一种电子设备，该电子设备包括第一种设计下的装置 1300 与第二种设计下的装置 1300。

图 15 为本申请实施例提供的一种芯片硬件结构，该芯片上包括神经网络处理器 1500。该芯片可以被设置在如下任一种或多种装置中：

如图 13 所示的装置 1100、如图 13 所示的装置 1300、如图 14 中所示的装置 1400。

上文方法实施例中的方法 300、900 或 1000 均可在如图 15 所示的芯片中得以实现。

神经网络处理器 1500 作为协处理器挂载到主处理器（Host CPU）上，由主 CPU 分配任务。神经网络处理器 1500 的核心部分为运算电路 1503，控制器 1504 控制运算电路 1503

获取存储器（权重存储器 1502 或输入存储器 1501）中的数据并进行运算。

在一些实现中，运算电路 1503 内部包括多个处理单元（process engine，PE）。在一些实现中，运算电路 1503 是二维脉动阵列。运算电路 1503 还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中，运算电路 1503 是通用的矩阵处理器。

举例来说，假设有输入矩阵 A，权重矩阵 B，输出矩阵 C。运算电路 1503 从权重存储器 1502 中取矩阵 B 相应的数据，并缓存在运算电路 1503 中每一个 PE 上。运算电路 1503 从输入存储器 1501 中取矩阵 A 数据与矩阵 B 进行矩阵运算，得到的矩阵的部分结果或最终结果，保存在累加器（accumulator）1508 中。

向量计算单元 1507 可以对运算电路 1503 的输出做进一步处理，如向量乘，向量加，指数运算，对数运算，大小比较等等。例如，向量计算单元 1507 可以用于神经网络中非卷积/非 FC 层的网络计算，如池化（pooling），批归一化（batch normalization），局部响应归一化（local response normalization）等。

在一些实现种，向量计算单元能 1507 将经处理的输出的向量存储到统一存储器（也可称为统一缓存器）1506。例如，向量计算单元 1507 可以将非线性函数应用到运算电路 1503 的输出，例如累加值的向量，用以生成激活值。在一些实现中，向量计算单元 1507 生成归一化的值、合并值，或二者均有。在一些实现中，处理过的输出的向量能够用作到运算电路 1503 的激活输入，例如用于在神经网络中的后续层中的使用。

上文方法实施例中的方法 300、500 或 600 可以由 1503 或 1507 执行。

统一存储器 1506 用于存放输入数据以及输出数据。

权重数据直接通过存储单元访问控制器 1505 （direct memory access controller，DMAC）将外部存储器中的输入数据搬运到输入存储器 1501 和/或统一存储器 1506、将外部存储器中的权重数据存入权重存储器 1502，以及将统一存储器 1506 中的数据存入外部存储器。

总线接口单元（bus interface unit，BIU）1510，用于通过总线实现主 CPU、DMAC 和取指存储器 1509 之间进行交互。

与控制器 1504 连接的取指存储器（instruction fetch buffer）1509，用于存储控制器 1504 使用的指令；

控制器 1504，用于调用指存储器 1509 中缓存的指令，实现控制该运算加速器的工作过程。

在图 15 所示的芯片用于执行上文方法实施例中的方法 300 的情况下，这里的数据可以是原始文本样本。

在图 15 所示的芯片用于执行上文方法实施例中的方法 900 的情况下，这里的数据可以是掩码训练样本。

在图 15 所示的芯片用于执行上文方法实施例中的方法 1000 的情况下，这里的数据可以是待预测的目标文本。

一般地，统一存储器 1506，输入存储器 1501，权重存储器 1502 以及取指存储器 1509 均为片上（On-Chip） 存储器，外部存储器为该 NPU 外部的存储器，该外部存储器可以为双倍数据率同步动态随机存储器（double data rate synchronous dynamic random access

memory，DDR SDRAM）、高带宽存储器（high bandwidth memory，HBM）或其他可读可写的存储器。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请。

需要说明的是，本文中涉及的第一、第二、第三或第四等各种数字编号仅为描述方便进行的区分，并不用来限制本申请实施例的范围。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置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。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，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，例如，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，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：通用串行总线闪存盘（USB flash disk，UFD）（UFD也可以简称为U盘或者优盘）、移动硬盘、只读存储器（read-only memory，ROM）、随机存取存储器（random access memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

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，可轻易想到变化或替换，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所述权利要求的保护范围为准。

权 利 要 求 书

1. 一种数据处理的方法，其特征在于，包括:

确定原始文本样本，所述原始文本样本未进行掩码处理;

对所述原始文本样本进行掩码处理，获得掩码训练样本，所述掩码处理使得所述掩码训练样本的掩码比例不固定，所述掩码训练样本用于训练预训练语言模型 PLM。

2. 根据权利要求 1 所述的方法，其特征在于，所述掩码训练样本的文本级别掩码比例包括:

文本级别掩码比例，用于表示一个样本中被掩码处理的字占所述样本中所有字的比例; 和/或

字级别掩码比例，用于表示一个字被掩码处理的概率;

其中，所述掩码训练样本的掩码比例不固定包括:

所述掩码训练样本中不同样本的文本级别掩码比例不完全相同; 和/或

所述掩码训练样本中任一个样本中每个字的字级别掩码比例不完全相同。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述对所述原始文本样本进行掩码处理，获得掩码训练样本，包括:

使用先验概率分布模型，生成所述原始文本样本中每个样本的文本级别掩码比例，所述先验概率分布模型使得所述原始文本样本中不同样本的文本级别掩码比例不完全相同;

按照所述原始文本样本中每个样本的文本级别掩码比例，对相应样本进行掩码处理，获得所述掩码训练样本。

4. 根据权利要求 3 所述的方法，其特征在于，所述先验概率分布模型的概率值区间长度不小于 40%。

5. 根据权利要求 1 或 2 所述的方法，其特征在于，所述对所述原始文本样本进行掩码处理，获得掩码训练样本，包括:

获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，所述第一文本样本中不同字的字级别掩码比例不完全相同;

根据所述第一文本样本中各个字的字级别掩码比例，对所述第一文本样本中的部分字进行掩码处理，获得所述掩码训练样本中的第一训练样本。

6. 根据权利要求 5 所述的方法，其特征在于，所述获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，包括:

使用先验概率分布模型，生成所述第一文本样本中每个字的字级别掩码比例，所述先验概率分布模型使得所述第一文本样本中不同字的字级别掩码比例不完全相同。

7. 根据权利要求 5 所述的方法，其特征在于，所述获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，包括:

将所述第一文本样本输入神经网络模型，从所述神经网络模型的输出获得所述第一文本样本中每个字的字级别掩码比例，其中，所述神经网络模型通过如下步骤进行优化学习得到，其中，i 的初始取值为 1:

1），将所述原始文本样本中第 i 个样本输入所述神经网络模型，从所述神经网络模

型的输出获得所述第 i 个样本中每个字的字级别掩码比例;

2），根据所述第 i 个样本中各个字的字级别掩码比例，对所述第 i 个样本中的部分字进行掩码处理，获得所述第 i 个样本对应的训练样本;

3），将所述第 i 个样本对应的训练样本输入所述 PLM，获得所述 PLM 针对被掩码处理的字的损失值;

4），根据所述 PLM 针对被掩码处理的字输出的损失值，以及所述神经网络模型针对所述被掩码处理的字的输出信号，更新优化所述神经网络网络;

5），判断所述神经网络网络是否满足收敛条件，若是，转到步骤 6），若否，将 i 的取值加 1，转到步骤 1）;

6），将所述步骤 4）得到的神经网络模型作为优化学习到的所述神经网络模型。

8. 根据权利要求 7 所述的方法，其特征在于，所述步骤 3）包括:

利用所述第 i 个样本对应的训练样本对所述 PLM 进行一次训练更新;

将所述第 i 个样本对应的训练样本输入经过所述训练更新的所述 PLM，获得经过所述训练更新的所述 PLM 针对所述被掩码处理的字输出的损失值;

其中，所述步骤 4）包括：根据经过所述训练更新的所述 PLM 针对所述被掩码处理的字输出的损失值，以及所述神经网络模型针对所述被掩码处理的字的输出信号，更新优化所述神经网络网络。

9. 根据权利要求 5-8 中任一项所述的方法，其特征在于，所述根据所述第一文本样本中各个字的字级别掩码比例，对所述第一文本样本中的部分字进行掩码处理，获得所述掩码训练样本中的第一训练样本，包括:

按照字级别掩码比例从高到低的顺序，对所述第一文本样本中前 S 个字或者位于前 G%的字进行掩码处理，获得所述第一训练样本，S 为取值小于所述第一文本样本中字的总数量的正整数，G 为大于 0 且小于 100 的整数。

10. 一种数据处理的方法，其特征在于，包括:

通过如权利要求 1-9 中任一项所述的方法获得掩码训练样本;

使用所述掩码训练样本训练预训练语言模型 PLM，所述 PLM 用于预测被掩码处理的文字。

11. 一种数据处理的方法，其特征在于，包括:

确定待预测的目标文本，所述目标文本包括缺少部分文字的语句;

将所述目标文本输入预训练语言模型 PLM，从所述 PLM 的输出预测所述目标文本中缺少的文字，

其中，所述 PLM 通过权利要求 10 所述的方法训练得到。

12. 一种数据处理的装置，其特征在于，包括:

第一处理单元，用于确定原始文本样本，所述原始文本样本未进行掩码处理;

第二处理单元，用于对所述原始文本样本进行掩码处理，获得掩码训练样本，所述掩码处理使得所述掩码训练样本的掩码比例不固定，所述掩码训练样本用于训练预训练语言模型 PLM。

13. 根据权利要求 11 所述的装置，其特征在于，所述掩码训练样本的掩码比例包括:

文本级别掩码比例，用于表示一个样本中被掩码处理的字占所述样本中所有字的比

例；和/或

字级别掩码比例，用于表示一个字被掩码处理的概率；

其中，所述掩码训练样本的掩码比例不固定包括：

所述掩码训练样本中不同样本的文本级别掩码比例不完全相同；和/或

所述掩码训练样本中任一个样本中每个字的字级别掩码比例不完全相同。

14. 根据权利要求 12 或 13 所述的装置，其特征在于，所述第二处理单元用于：

使用先验概率分布模型，生成所述原始文本样本中每个样本的文本级别掩码比例，所述先验概率分布模型使得所述原始文本样本中不同样本的文本级别掩码比例不完全相同；

按照所述原始文本样本中每个样本的文本级别掩码比例，对相应样本进行掩码处理，获得所述掩码训练样本。

15. 根据权利要求 14 所述的装置，其特征在于，所述先验概率分布模型的概率值区间长度不小于 40%。

16. 根据权利要求 12 或 13 所述的装置，其特征在于，所述第二处理单元用于：

获取所述原始文本样本中的第一文本样本中每个字的字级别掩码比例，所述第一文本样本中不同字的字级别掩码比例不完全相同；

根据所述第一文本样本中各个字的字级别掩码比例，对所述第一文本样本中的部分字进行掩码处理，获得所述掩码训练样本中的第一训练样本。

17. 根据权利要求 16 所述的装置，其特征在于，所述第二处理单元用于，使用先验概率分布模型，生成所述第一文本样本中每个字的字级别掩码比例，所述先验概率分布模型使得所述第一文本样本中不同字的字级别掩码比例不完全相同。

18. 根据权利要求 16 所述的装置，其特征在于，所述第二处理单元用于，将所述第一文本样本输入神经网络模型，从所述神经网络模型的输出获得所述第一文本样本中每个字的字级别掩码比例，

其中，所述神经网络模型通过如下步骤进行优化学习得到，其中，i 的初始取值为 1：

1），将所述原始文本样本中第 i 个样本输入所述神经网络模型，从所述神经网络模型的输出获得所述第 i 个样本中每个字的字级别掩码比例；

2），根据所述第 i 个样本中各个字的字级别掩码比例，对所述第 i 个样本中的部分字进行掩码处理，获得所述第 i 个样本对应的训练样本；

3），将所述第 i 个样本对应的训练样本输入所述 PLM，获得所述 PLM 针对被掩码处理的字输出的损失值；

4），根据所述 PLM 针对被掩码处理的字输出的损失值，以及所述神经网络模型针对所述被掩码处理的字的输出信号，更新优化所述神经网络网络；

5），判断所述神经网络网络是否满足收敛条件，若是，转到步骤 6），若否，将 i 的取值加 1，转到步骤 1）；

6），将所述步骤 4）得到的神经网络模型作为优化学习到的所述神经网络模型。

19. 根据权利要求 18 所述的装置，其特征在于，所述步骤 3）包括：

利用所述第 i 个样本对应的训练样本对所述 PLM 进行一次训练更新；

将所述第 i 个样本对应的训练样本输入经过所述训练更新的所述 PLM，获得经过所述训练更新的所述 PLM 针对所述被掩码处理的字的损失值；

其中，所述步骤 4）包括：根据经过所述训练更新的所述 PLM 针对所述被掩码处理的字的损失值，以及所述神经网络模型针对所述被掩码处理的字的输出信号，更新优化所述神经网络网络。

20. 根据权利要求 16-19 中任一项所述的装置，其特征在于，所述第二处理单元用于，按照字级别掩码比例从高到低的顺序，对所述第一文本样本中前 S 个字或者位于前 G%的字进行掩码处理，获得所述第一训练样本，S 为取值小于所述第一文本样本中字的总数量的正整数，G 为大于 0 且小于 100 的整数。

21. 一种数据处理的装置，其特征在于，包括：

第一处理单元，用于通过如权利要求 1-9 中任一项所述的方法获得掩码训练样本；

第二处理单元，用于使用所述掩码训练样本训练预训练语言模型 PLM，所述 PLM 用于预测被掩码处理的文字。

22. 一种数据处理的装置，其特征在于，包括：

第一处理单元，用于确定待预测的目标文本，所述目标文本包括缺少部分文字的语句；

第二处理单元，用于将所述目标文本输入预训练语言模型 PLM，从所述 PLM 的输出预测所述目标文本中缺少的文字，其中，所述 PLM 通过权利要求 10 所述的方法训练得到。

23. 一种数据处理的装置，其特征在于，包括：

存储器，用于存储程序；

处理器，用于执行所述存储器中存储的程序，当所述存储器中存储的程序被执行时，所述处理器用于执行权利要求 1 至 11 中任一项所述的方法。

24. 一种计算机可读存储介质，其特征在于，所述计算机可读介质存储用于设备执行的程序代码，所述程序代码被执行时，所述所示设备执行权利要求 1 至 11 中任一项所述的方法。

25. 一种芯片，其特征在于，包括至少一个处理器和数据接口；

所述至少一个所述处理器用于，通过所述数据接口调用并运行存储在存储器上的计算机程序，以使所述芯片执行权利要求 1 至 11 中任一项所述的方法。

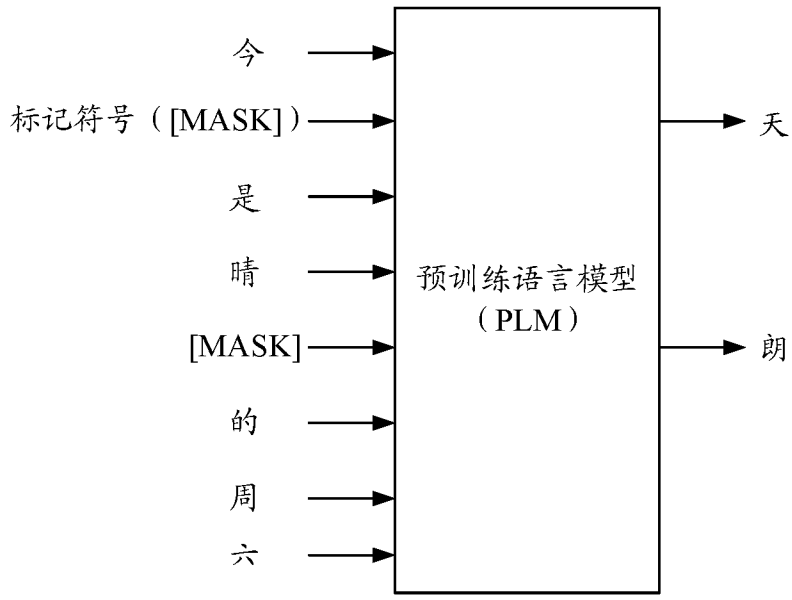

图 1

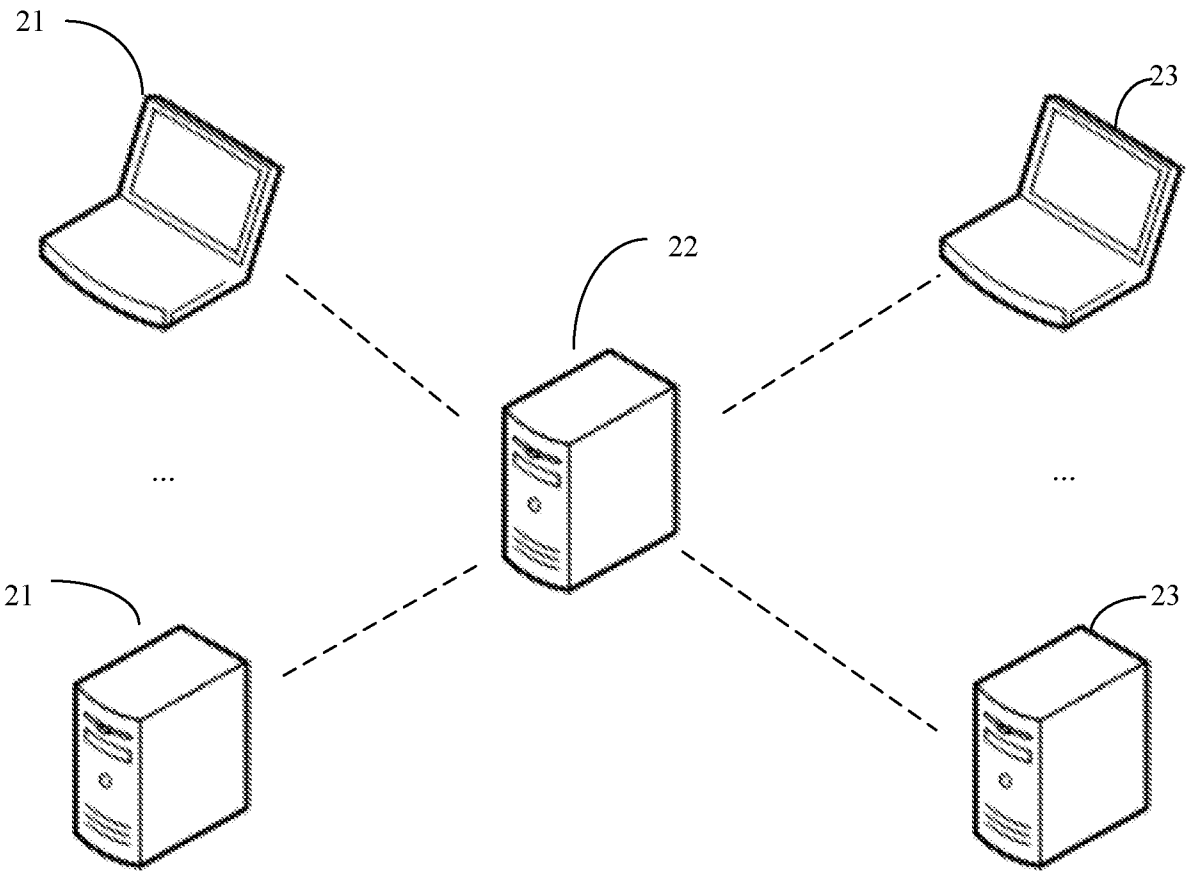

图 2

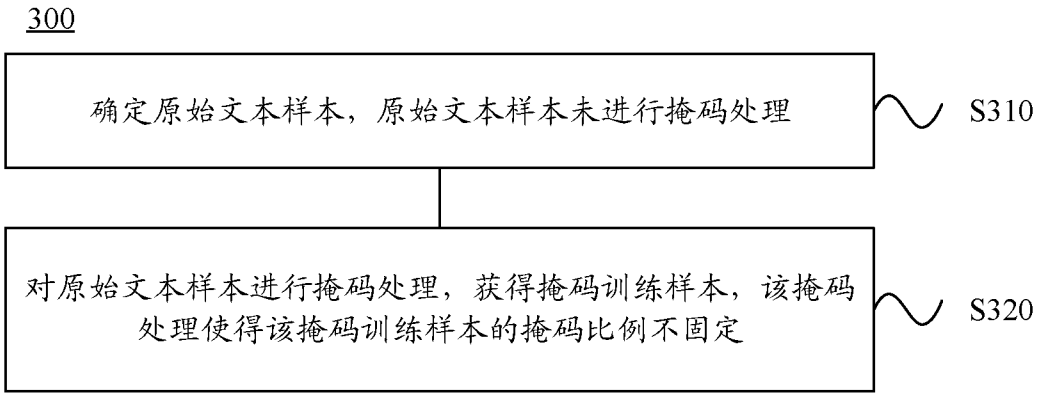
300
确定原始文本样本，原始文本样本未进行掩码处理
S310
对原始文本样本进行掩码处理，获得掩码训练样本，该掩码处理使得该掩码训练样本的掩码比例不固定
S320

图 3

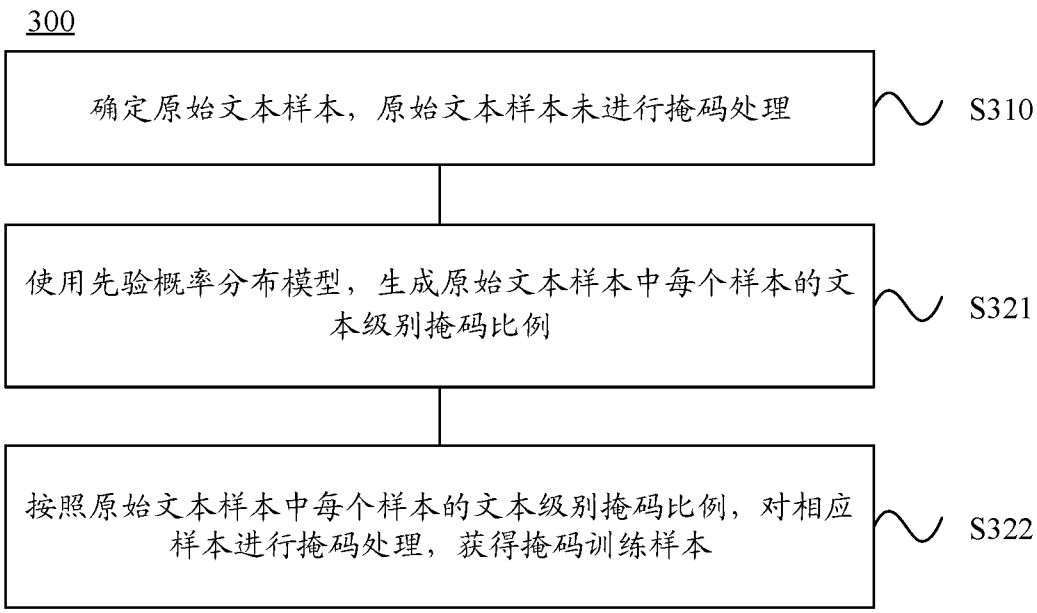
300
确定原始文本样本，原始文本样本未进行掩码处理
S310
使用先验概率分布模型，生成原始文本样本中每个样本的文本级别掩码比例
S321
按照原始文本样本中每个样本的文本级别掩码比例，对相应样本进行掩码处理，获得掩码训练样本
S322

图 4

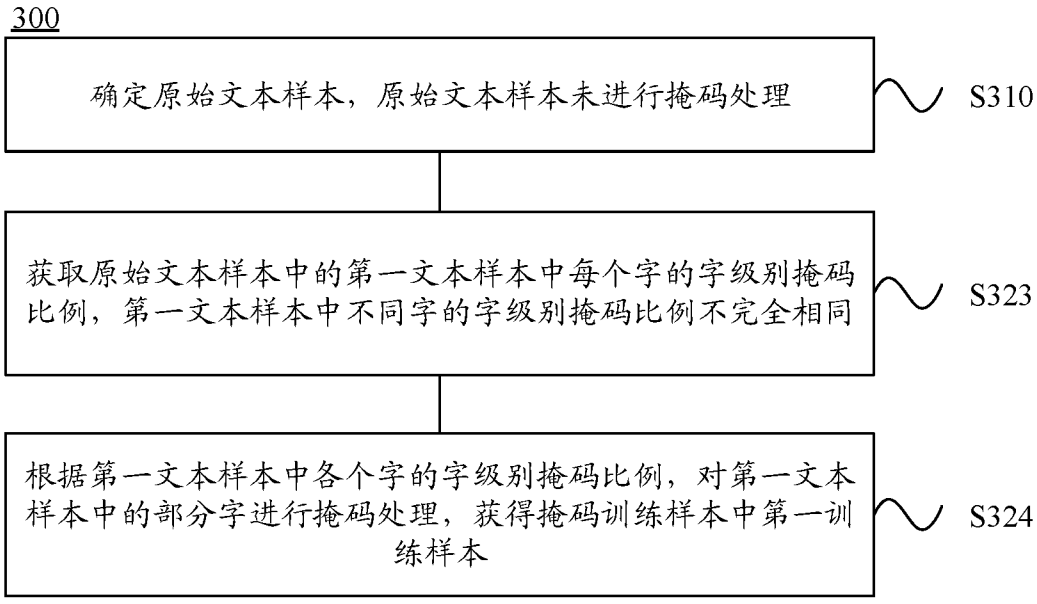
300
确定原始文本样本，原始文本样本未进行掩码处理
S310
获取原始文本样本中的第一文本样本中每个字的字级别掩码比例，第一文本样本中不同字的字级别掩码比例不完全相同
S323
根据第一文本样本中各个字的字级别掩码比例，对第一文本样本中的部分字进行掩码处理，获得掩码训练样本中第一训练样本
S324

图 5

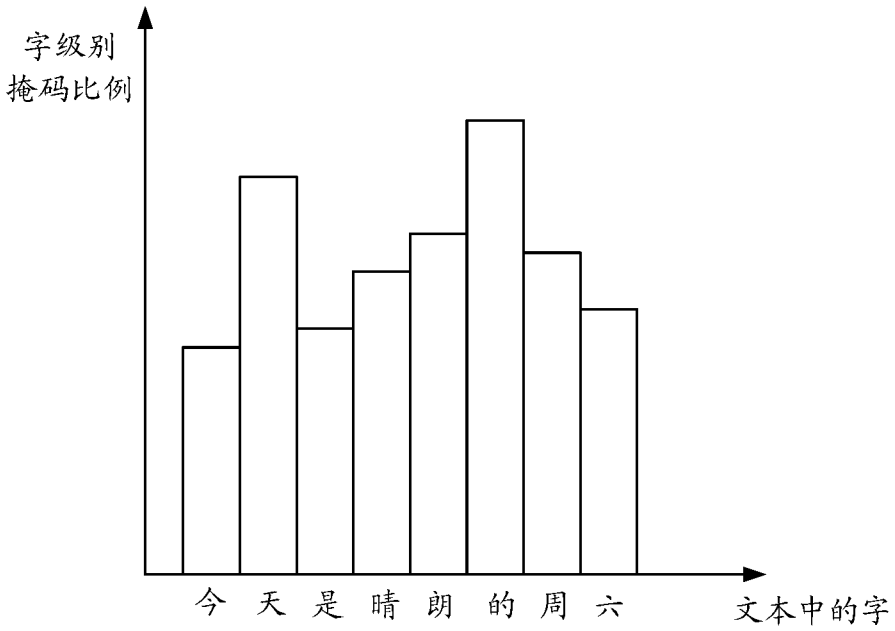

图 6

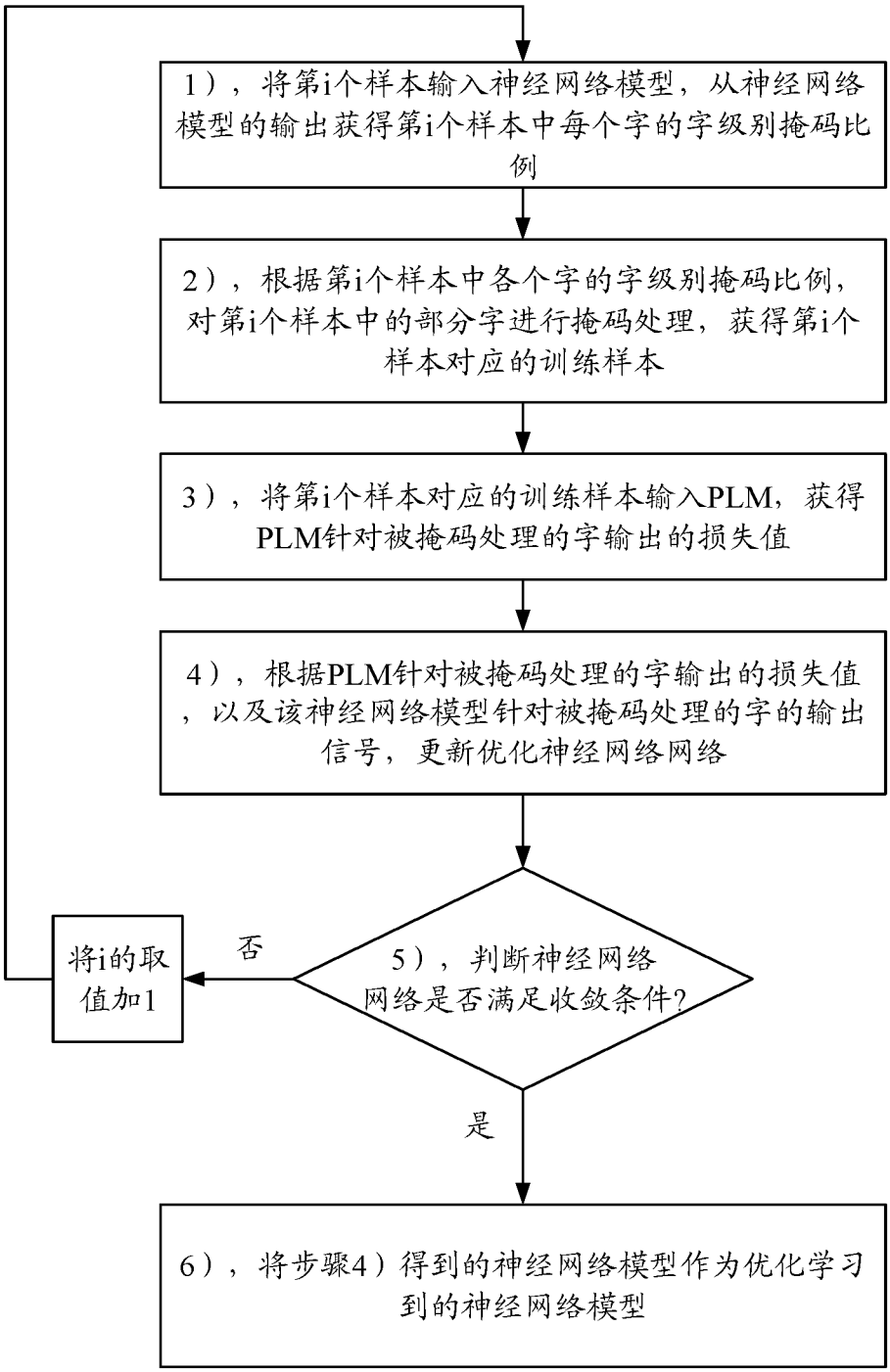
1），将第i个样本输入神经网络模型，从神经网络模型的输出获得第i个样本中每个字的字级别掩码比例
2），根据第i个样本中各个字的字级别掩码比例，对第i个样本中的部分字进行掩码处理，获得第i个样本对应的训练样本
3），将第i个样本对应的训练样本输入PLM，获得PLM针对被掩码处理的字输出的损失值
4），根据PLM针对被掩码处理的字输出的损失值，以及该神经网络模型针对被掩码处理的字的输出信号，更新优化神经网络网络
5），判断神经网络网络是否满足收敛条件?
否
将i的取值加1
是
6），将步骤4）得到的神经网络模型作为优化学习到的神经网络模型

图 7

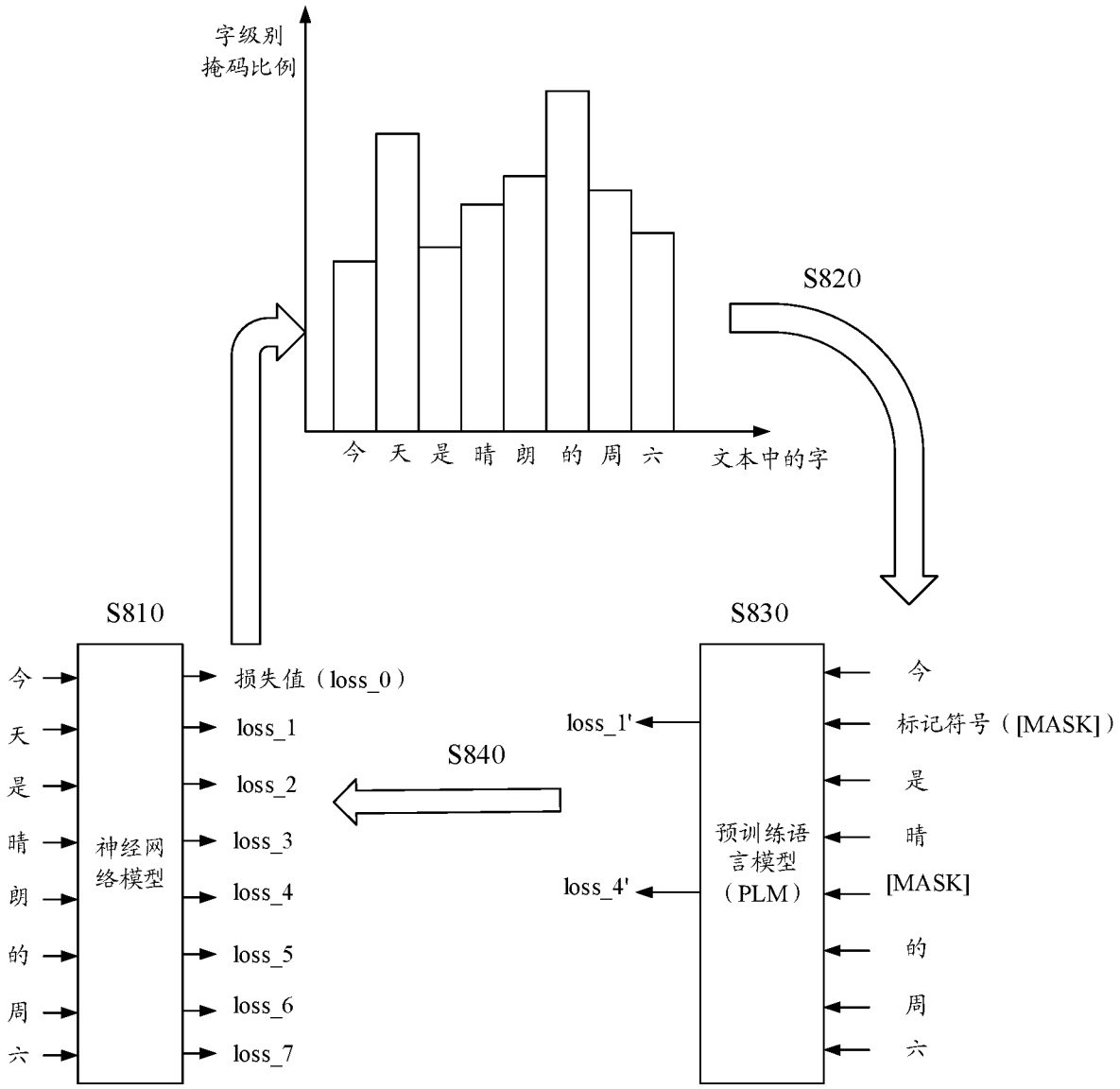
字级别
掩码比例
今 天 是 晴 朗 的 周 六
文本中的字
S820
S810
S830
S840
今
天
是
晴
朗
的
周
六
神经网络模型
损失值（loss_0）
loss_1
loss_2
loss_3
loss_4
loss_5
loss_6
loss_7
loss_1'
loss_4'
预训练语言模型（PLM）
今
标记符号（[MASK]）
是
晴
[MASK]
的
周
六

图 8

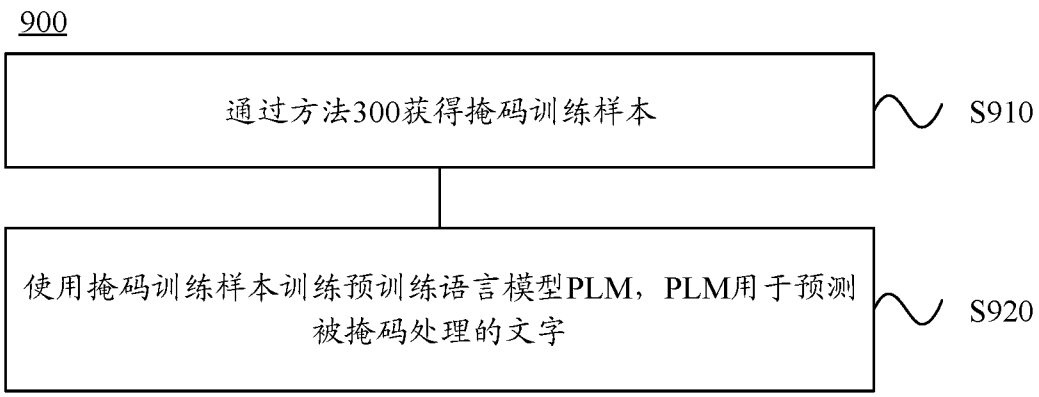
900
通过方法300获得掩码训练样本
S910
使用掩码训练样本训练预训练语言模型PLM，PLM用于预测被掩码处理的文字
S920

图 9

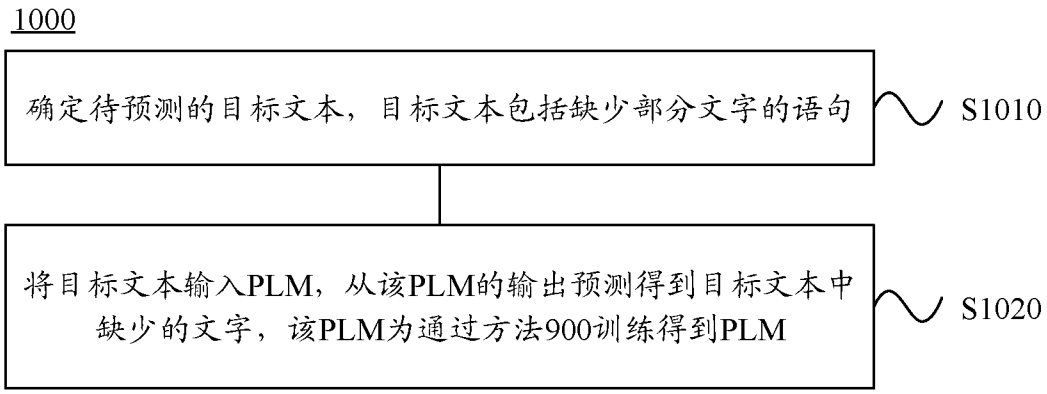
1000
确定待预测的目标文本，目标文本包括缺少部分文字的语句
S1010
将目标文本输入PLM，从该PLM的输出预测得到目标文本中缺少的文字，该PLM为通过方法900训练得到PLM
S1020

图 10

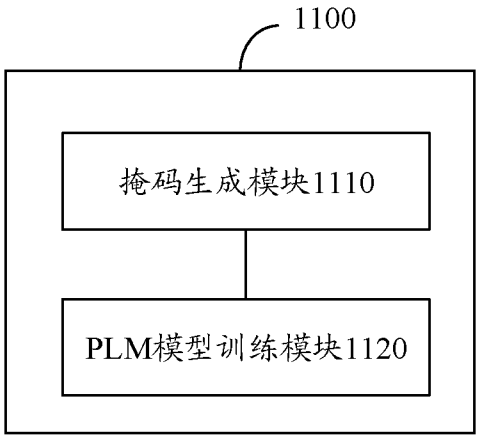
1100
掩码生成模块1110
PLM模型训练模块1120

图 11

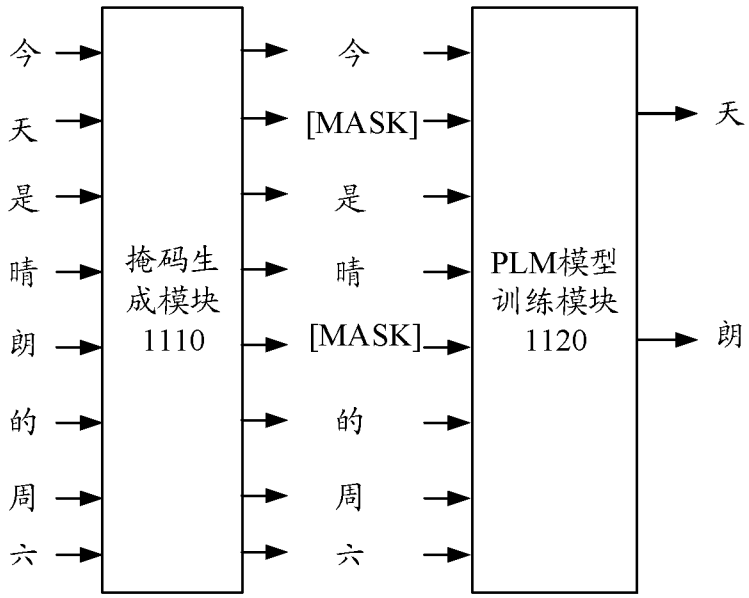
今
天
是
晴
朗
的
周
六
掩码生成模块1110
今
[MASK]
是
晴
[MASK]
的
周
六
PLM模型训练模块1120
天
朗

图 12

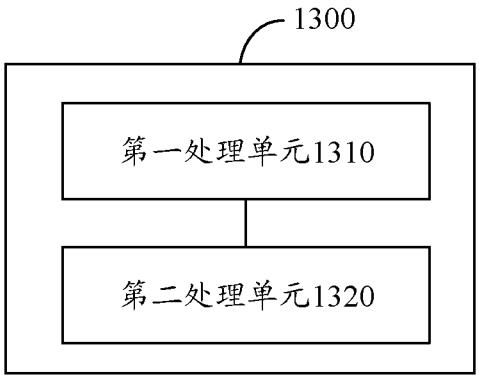
1300
第一处理单元1310
第二处理单元1320

图 13

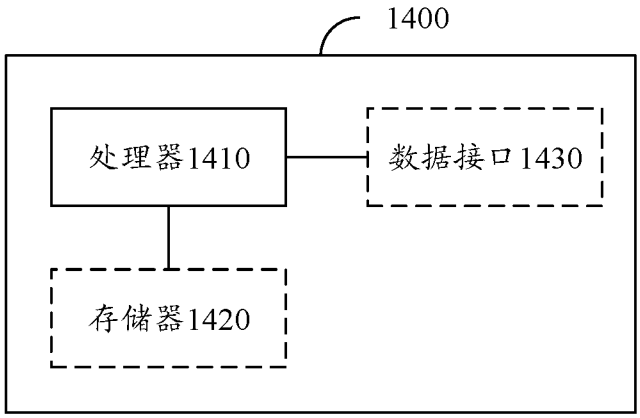
1400
处理器1410
数据接口1430
存储器1420

图 14

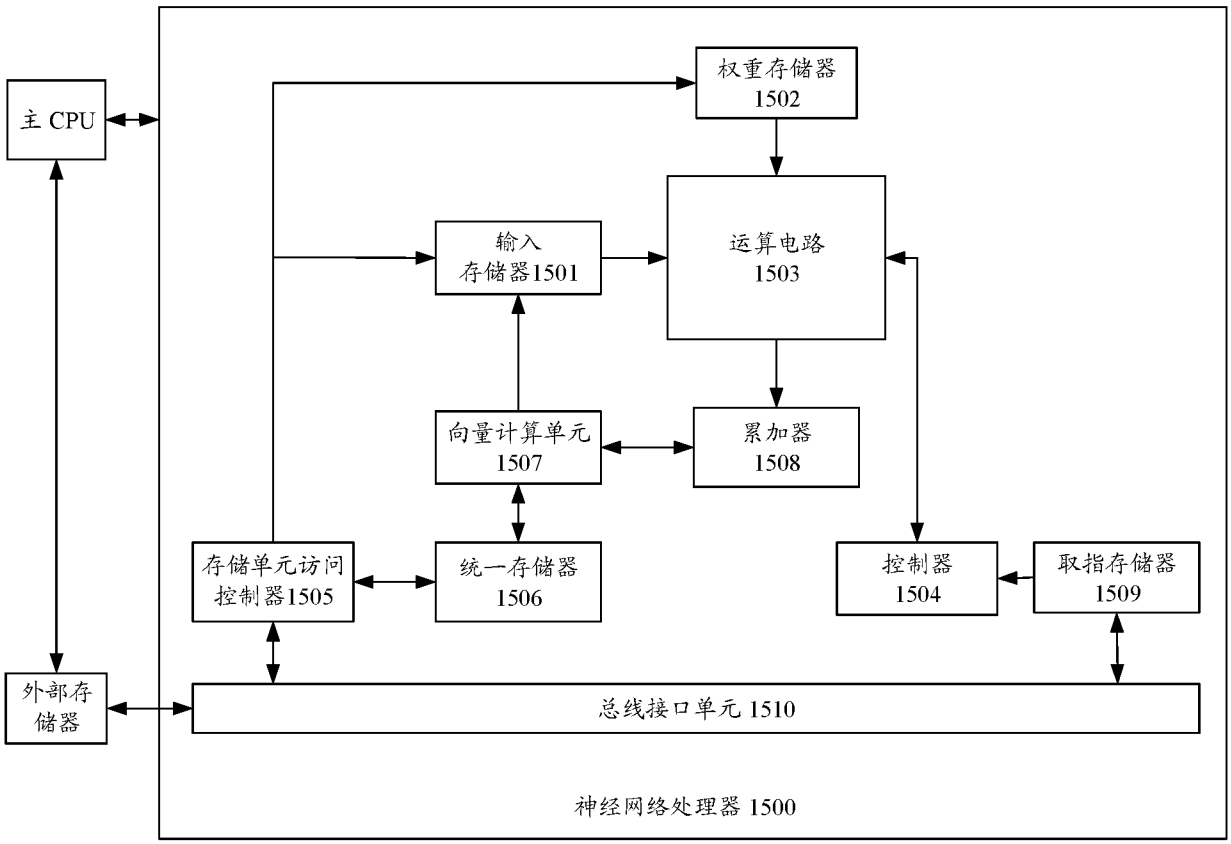
主 CPU
权重存储器
1502
输入
存储器1501
运算电路
1503
向量计算单元
1507
累加器
1508
存储单元访问
控制器1505
统一存储器
1506
控制器
1504
取指存储器
1509
外部存
储器
总线接口单元 1510
神经网络处理器 1500

图 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/078390

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/33(2019.01)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)

CNABS, CNKI, CNTXT, DWPI, SIPOABS; 语言模型, 训练, 样本, 掩码, 遮蔽, 遮盖; language model, train, sample, mask

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 111611790 A (HUAWEI TECHNOLOGIES CO., LTD.) 01 September 2020 (2020-09-01) description, pages 5-285, and figures 1-15	1-25
X	CN 110196894 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 03 September 2019 (2019-09-03) description, pages 3-280, and figures 1-12	1-25
A	CN 110377744 A (BEIJING XIANGNONG HUIYU TECHNOLOGY CO., LTD.) 25 October 2019 (2019-10-25) entire document	1-25
A	US 2011137653 A1 (AT&T INTELLECTUAL PROPERTY I, L.P.) 09 June 2011 (2011-06-09) entire document	1-25

☐ 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
20 May 2021	**07 June 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/CN2021/078390

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 111611790 A	01 September 2020	None	
CN 110196894 A	03 September 2019	None	
CN 110377744 A	25 October 2019	None	
US 2011137653 A1	09 June 2011	US 8589163 B2	19 November 2013

国际检索报告

国际申请号
PCT/CN2021/078390

A. 主题的分类

G06F 16/33(2019.01)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））

CNABS, CNKI, CNTXT, DWPI, SIPOABS; 语言模型，训练，样本，掩码，遮蔽，遮盖; language model, train, sample, mask

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
PX	CN 111611790 A (华为技术有限公司) 2020年 9月 1日 (2020-09-01) 说明书第5-285页、图1-15	1-25
X	CN 110196894 A (北京百度网讯科技有限公司) 2019年 9月 3日 (2019-09-03) 说明书第3-280页、图1-12	1-25
A	CN 110377744 A (北京香侬慧语科技有限责任公司) 2019年 10月 25日 (2019-10-25) 全文	1-25
A	US 2011137653 A1 (AT&T INTELLECTUAL PROPERTY I LP) 2011年 6月 9日 (2011-06-09) 全文	1-25

☐ 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 5月 20日	2021年 6月 7日
ISA/CN的名称和邮寄地址 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 传真号 (86-10)62019451	受权官员 孟田革 电话号码 86-10-62411653

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/078390

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 111611790 A	2020年 9月 1日	无	
CN 110196894 A	2019年 9月 3日	无	
CN 110377744 A	2019年 10月 25日	无	
US 2011137653 A1	2011年 6月 9日	US 8589163 B2	2013年 11月 19日