



(12) 发明专利申请

(10) 申请公布号 CN 117649913 A

(43) 申请公布日 2024. 03. 05

(21) 申请号 202311701514.5

G06N 20/00 (2019.01)

(22) 申请日 2023.12.12

(71) 申请人 郑州新益华信息科技有限公司

地址 450000 河南省郑州市市辖区郑东新区平安大道197号龙子湖中央广场1单元19层31-32号

(72) 发明人 张贺伟

(74) 专利代理机构 郑州中鼎万策专利代理事务所(普通合伙) 41179

专利代理师 徐文婷

(51) Int. Cl.

G16H 10/60 (2018.01)

G06F 40/12 (2020.01)

G06F 16/901 (2019.01)

G06F 16/903 (2019.01)

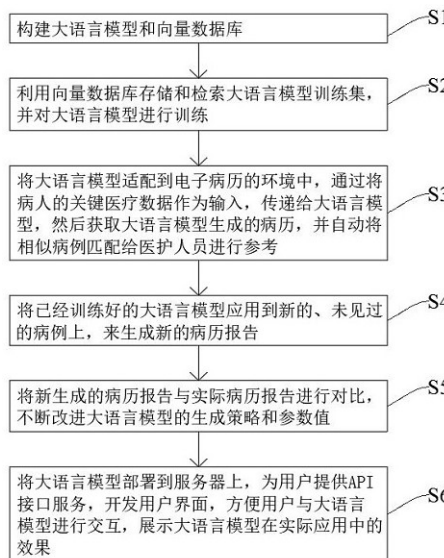
权利要求书2页 说明书4页 附图3页

(54) 发明名称

一种基于大型语言模型的电子病历生成方法

(57) 摘要

本发明提供一种基于大型语言模型的电子病历生成方法,涉及病历生成技术领域,该基于大型语言模型的电子病历生成方法,包括以下步骤:步骤S1,构建大语言模型和向量数据库;步骤S2,利用向量数据库存储和检索大语言模型训练集;步骤S3,将大语言模型适配到电子病历的环境中;步骤S4,将已经训练好的大语言模型应用到新的、未见过的病例上;通过构建向量数据库和大语言模型,以使其能够准确理解和生成医学领域的文本,从而使其可以根据医生输入的关键词和诊断要点,自动生成符合要求的电子病历文本,避免了传统填写方式中的大量输入的繁琐和重复劳动,还能够自动快速检索和匹配相关历史病历,以提供给医护人员更全面的医疗信息和参考依据。



1. 一种基于大型语言模型的电子病历生成方法,其特征是,包括以下步骤:

步骤S1,构建大语言模型和向量数据库;

步骤S2,利用向量数据库存储和检索大语言模型训练集,并对大语言模型进行训练;

步骤S3,将大语言模型适配到电子病历的环境中,通过将病人的关键医疗数据作为输入,传递给大语言模型,然后获取大语言模型生成的病历,并自动将相似病例匹配给医护人员进行参考;

步骤S4,将已经训练好的大语言模型应用到新的、未见过的病例上,来生成新的病历报告。

2. 根据权利要求1所述的一种基于大型语言模型的电子病历生成方法,其特征是,所述步骤S4之后还包括:

步骤S5,将新生成的病历报告与实际病历报告进行对比,不断改进大语言模型的生成策略和参数值;

步骤S6,将大语言模型部署到服务器上,为用户提供API接口服务,开发用户界面,方便用户与大语言模型进行交互,展示大语言模型在实际应用中的效果。

3. 根据权利要求1所述的一种基于大型语言模型的电子病历生成方法,其特征是,所述步骤S1中向量数据库的构建,包括以下步骤:

步骤S11,数据获取,采集医学文本数据到数据库并进行整理;

步骤S12,数据清洗,对收集到数据库的数据进行数据合并、离群值检测、缺失值处理、去重以及文本处理;

步骤S13,数据标注,将文本数据映射到低维向量空间,其中每个词汇被映射到一个低维的连续向量空间,使大语言模型具备深度理解文本和进行高级文本分析的能力;

步骤S14,数据划分,将清洗、标注后的数据划分为训练集、验证集和测试集。

4. 根据权利要求3所述的一种基于大型语言模型的电子病历生成方法,其特征是,在所述步骤S11中,医学文本数据包括病历系统、教科书和临床试验数据等。

5. 根据权利要求3所述的一种基于大型语言模型的电子病历生成方法,其特征是,在所述步骤S12中,文本处理包括分词、去停用词和词干提取等。

6. 根据权利要求1所述的一种基于大型语言模型的电子病历生成方法,其特征是,所述步骤S2中,大语言模型训练集包括医学术语和病历格式。

7. 根据权利要求1所述的一种基于大型语言模型的电子病历生成方法,其特征是,所述步骤S2中大语言模型的训练,包括以下步骤:

步骤S21,构建词汇表,将词汇表中的每个单词映射到高维向量空间,添加位置信息以捕获单词顺序,以便大语言模型能够理解单词的顺序;

步骤S22,计算单词之间的依赖关系,捕捉长距离依赖关系,为大语言模型提供深度理解文本的能力;

步骤S23,将大语言模型的输出转换为所需的标签或下一个单词的概率分布,以构建大语言模型的基础架构;

步骤S24,大语言模型的训练和优化,计算大语言模型预测输出与实际输出之间的损失,根据损失函数的梯度,更新大语言模型的参数,验证和调整学习率、批次大小和隐藏层大小参数;

步骤S25,通过计算大语言模型的预测结果和真实结果之间的均方误差,测试预测大语言模型误差情况。

8.根据权利要求7所述的一种基于大型语言模型的电子病历生成方法,其特征是,在所述步骤S21中,词汇表包含医学文本的所有不同的单词,可以通过词嵌入、位置编码和自注意力机制等技术将词汇表中的每个单词映射到高维向量空间。

9.根据权利要求7所述的一种基于大型语言模型的电子病历生成方法,其特征是,在所述步骤S24中,当验证集上的性能开始下降时,停止训练。

10.根据权利要求7所述的一种基于大型语言模型的电子病历生成方法,其特征是,在所述步骤S25中,通过比较大语言模型在不同时间、不同数据上的预测结果以评估大语言模型的稳定性。

一种基于大型语言模型的电子病历生成方法

技术领域

[0001] 本发明涉及病历生成技术领域,具体是一种基于大型语言模型的电子病历生成方法。

背景技术

[0002] 病历是医务人员对患者疾病的发生、发展、转归,以及对患者进行检查、诊断、治疗等医疗活动过程的记录。也是对采集到的资料加以归纳、整理、综合分析,按规定的格式和要求书写的患者医疗健康档案。对医疗、预防、教学、科研、医院管理等都有重要的作用。

[0003] 随着电子信息技术的发展,电子病历越来越多的应用到医疗活动过程中。然而,目前的电子病历只提供针对医疗活动过程中某个医疗环节的电子病历模板,由各个医疗环节的医务人员在电子病历模板中输入医疗数据生成最终的针对该医疗环节的电子病历。

[0004] 但是,一般技术中的电子病历生成方法具有以下缺点:传统的电子病历需要医务人员输入大量的医疗数据生成最终的针对该医疗环节的电子病历,这种填写方式较为繁琐,且具有大量的重复性劳动,工作效率较低,而且不能患者的基本信息和病历特征自动匹配相关的历史病历以供医务人员进行参考,不便于医务人员进行使用。

发明内容

[0005] 本发明的目的在于提供一种基于大型语言模型的电子病历生成方法,旨在解决一般技术中的电子病历生成方法工作效率较低且不便于使用的问题。

[0006] 为实现上述目的,本发明采用如下技术方案:所述基于大型语言模型的电子病历生成方法,包括以下步骤:

步骤S1,构建大语言模型和向量数据库;

步骤S2,利用向量数据库存储和检索大语言模型训练集,并对大语言模型进行训练;

步骤S3,将大语言模型适配到电子病历的环境中,通过将病人的关键医疗数据作为输入,传递给大语言模型,然后获取大语言模型生成的病历,并自动将相似病例匹配给医护人员进行参考;

步骤S4,将已经训练好的大语言模型应用到新的、未见过的病例上,来生成新的病历报告。

[0007] 有益效果是:通过构建向量数据库和大语言模型,并对大语言模型进行训练,以使其能够准确理解和生成医学领域的文本,从而使其可以根据医生输入的关键词和诊断要点,自动生成符合要求的电子病历文本,避免了传统填写方式中的大量输入的繁琐和重复劳动,大大提高了医生的工作效率,还能够根据患者的基本信息和病历特征,自动快速检索和匹配相关历史病历,以提供给医护人员更全面的医疗信息和参考依据。

[0008] 本发明的进一步的技术方案为,所述步骤S4之后还包括:

步骤S5,将新生成的病历报告与实际病历报告进行对比,不断改进大语言模型的

生成策略和参数值;

步骤S6,将大语言模型部署到服务器上,为用户提供API接口服务,开发用户界面,方便用户与大语言模型进行交互,展示大语言模型在实际应用中的效果。

[0009] 本发明的进一步的技术方案为,所述步骤S1中向量数据库的构建,包括以下步骤:

步骤S11,数据获取,采集医学文本数据到数据库并进行整理;

步骤S12,数据清洗,对收集到数据库的数据进行数据合并、离群值检测、缺失值处理、去重以及文本处理;

步骤S13,数据标注,将文本数据映射到低维向量空间,其中每个词汇被映射到一个低维的连续向量空间,使大语言模型具备深度理解文本和进行高级文本分析的能力;

步骤S14,数据划分,将清洗、标注后的数据划分为训练集、验证集和测试集。

[0010] 有益效果是:通过对文本数据进行标注,可以使大语言模型能够深度理解文本和进行高级文本分析,可以使其不仅限于数据的表面理解,还能深入解析病症、症状、药物和剂量,实现高度精准的文本生成,为医学从业者在生成病历报告和协助医学决策等方面提供可靠的文本支持。

[0011] 本发明的进一步的技术方案为,在所述步骤S11中,医学文本数据包括病历系统、教科书和临床试验数据等。

[0012] 本发明的进一步的技术方案为,在所述步骤S12中,文本处理包括分词、去停用词和词干提取等。

[0013] 本发明的进一步的技术方案为,所述步骤S2中,大语言模型训练集包括医学术语和病历格式。

[0014] 本发明的进一步的技术方案为,所述步骤S2中大语言模型的训练,包括以下步骤:

步骤S21,构建词汇表,将词汇表中的每个单词映射到高维向量空间,添加位置信息以捕获单词顺序,以便大语言模型能够理解单词的顺序;

步骤S22,计算单词之间的依赖关系,捕捉长距离依赖关系,为大语言模型提供深度理解文本的能力;

步骤S23,将大语言模型的输出转换为所需的标签或下一个单词的概率分布,以构建大语言模型的基础架构;

步骤S24,大语言模型的训练和优化,计算大语言模型预测输出与实际输出之间的损失,根据损失函数的梯度,更新大语言模型的参数,验证和调整学习率、批次大小和隐藏层大小参数;

步骤S25,通过计算大语言模型的预测结果和真实结果之间的均方误差,测试预测大语言模型误差情况。

附图说明

[0015] 图1是本发明的流程图。

[0016] 图2是本发明的具体实施例中数据库构建的流程图。

[0017] 图3是本发明的具体实施例中大语言模型训练的流程图。

具体实施方式

[0018] 下面结合附图对本发明的具体实施方式作进一步的说明。

[0019] 如图1所示,一种基于大型语言模型的电子病历生成方法,包括以下步骤:

步骤S1,构建大语言模型和向量数据库;

步骤S2,利用向量数据库存储和检索大语言模型训练集,并对大语言模型进行训练,还可在大语言模型生成病历文本时进行参考和调用,以弥补大语言模型记忆性不足的问题,其中大语言模型训练集包括医学术语和病历格式;

步骤S3,将大语言模型适配到电子病历的环境中,通过将病人的关键医疗数据作为输入,传递给大语言模型,然后获取大语言模型生成的病历,并自动将相似病例匹配给医护人员进行参考;

步骤S4,将已经训练好的大语言模型应用到新的、未见过的病例上,来生成新的病历报告;

步骤S5,将新生成的病历报告与实际病历报告,进行对比,不断改进大语言模型的生成策略、参数值,从而提高生成病历报告的准确性;

步骤S6,将大语言模型部署到服务器上,为用户提供API接口服务;开发用户界面,方便用户与大语言模型进行交互,展示大语言模型在实际应用中的效果。

[0020] 在本实施例中,通过大语言模型的自然语言处理能力,可以根据医生输入的关键词和诊断要点,自动生成符合要求的电子病历文本,避免了传统填写方式中的繁琐和重复劳动,同时还能够提高文本的质量和准确性,使生成的电子病历文本更加规范和专业。

[0021] 如图2所示,其中,步骤S1中向量数据库的构建,包括以下步骤:

步骤S11,数据获取,采集医学文本数据到数据库并进行整理,医学文本数据包括病历系统、教科书和临床试验数据等,综合性的数据采集方式将医学领域的知识汇聚在一个数据库中,为大语言模型提供了全面而深入的学习材料,这一综合性的数据收集工作构建了训练集的基础,为后续大语言模型的训练提供了丰富的语料库;

步骤S12,数据清洗,对收集到数据库的数据进行数据合并、离群值检测、缺失值处理、去重以及文本处理,其中文本处理包括分词、去停用词和词干提取等,以确保训练数据的高质量和可用性;

步骤S13,数据标注,将文本数据映射到低维向量空间,其中每个词汇被映射到一个低维的连续向量空间,以便计算机可以更好地通过向量空间的距离计算理解 and 处理文本,使大语言模型具备深度理解文本和进行高级文本分析的能力;

步骤S14,数据划分,将清洗、标注后的数据划分为训练集、验证集和测试集,在本实施例中,训练集、验证集和测试集的比例为4:1:1。

[0022] 在步骤S13中,使大语言模型能够深度理解文本和进行高级文本分析,可以使其不仅限于数据的表面理解,还能深入解析病症、症状、药物和剂量,实现高度精准的文本生成,为医学从业者在生成病历报告和协助医学决策等方面提供可靠的文本支持。

[0023] 如图3所示,其中,步骤S2中大语言模型的训练,包括以下步骤:

步骤S21,构建词汇表,词汇表包含医学文本的所有不同的单词,通过词嵌入、位置编码和自注意力机制等技术将词汇表中的每个单词映射到高维向量空间,添加位置信息以捕获单词顺序,以便大语言模型能够理解单词的顺序;

步骤S22,计算单词之间的依赖关系,捕捉长距离依赖关系,为大语言模型提供深度理解文本的能力;

步骤S23,将大语言模型的输出转换为所需的标签或下一个单词的概率分布,以构建大语言模型的基础架构;

步骤S24,大语言模型的训练和优化,随机初始化模型词嵌入矩阵、位置编码矩阵、自注意力权重矩阵、FFN权重矩阵的参数;计算大语言模型预测输出与实际输出之间的损失;根据损失函数的梯度,更新大语言模型的参数;验证和调整学习率、批次大小、隐藏层大小参数;当验证集上的性能开始下降时,停止训练;

步骤S25,通过计算大语言模型的预测结果和真实结果之间的均方误差,测试预测大语言模型误差情况;通过比较大语言模型的预测结果和真实结果的差异来评估大语言模型拟合程度;通过比较大语言模型在不同时间、不同数据上的预测结果来评估大语言模型的稳定性;测量大语言模型处理一批数据所需的时间,然后除以这批数据的大小,评估大语言模型的处理速度;测量大语言模型运行时使用的CPU、内存、GPU资源的用量,然后除以大语言模型的运行时间,得到大语言模型的计算资源耗用量。

[0024] 在本具体实施例中,通过构建向量数据库和大语言模型,并对大语言模型进行训练,以使其能够准确理解和生成医学领域的文本,从而使其可以根据医生输入的关键词和诊断要点,自动生成符合要求的电子病历文本,避免了传统填写方式中的大量输入的繁琐和重复劳动,大大提高了医生的工作效率,还能够根据患者的基本信息和病历特征,自动快速检索和匹配相关历史病历,以提供给医护人员更全面的医疗信息和参考依据。

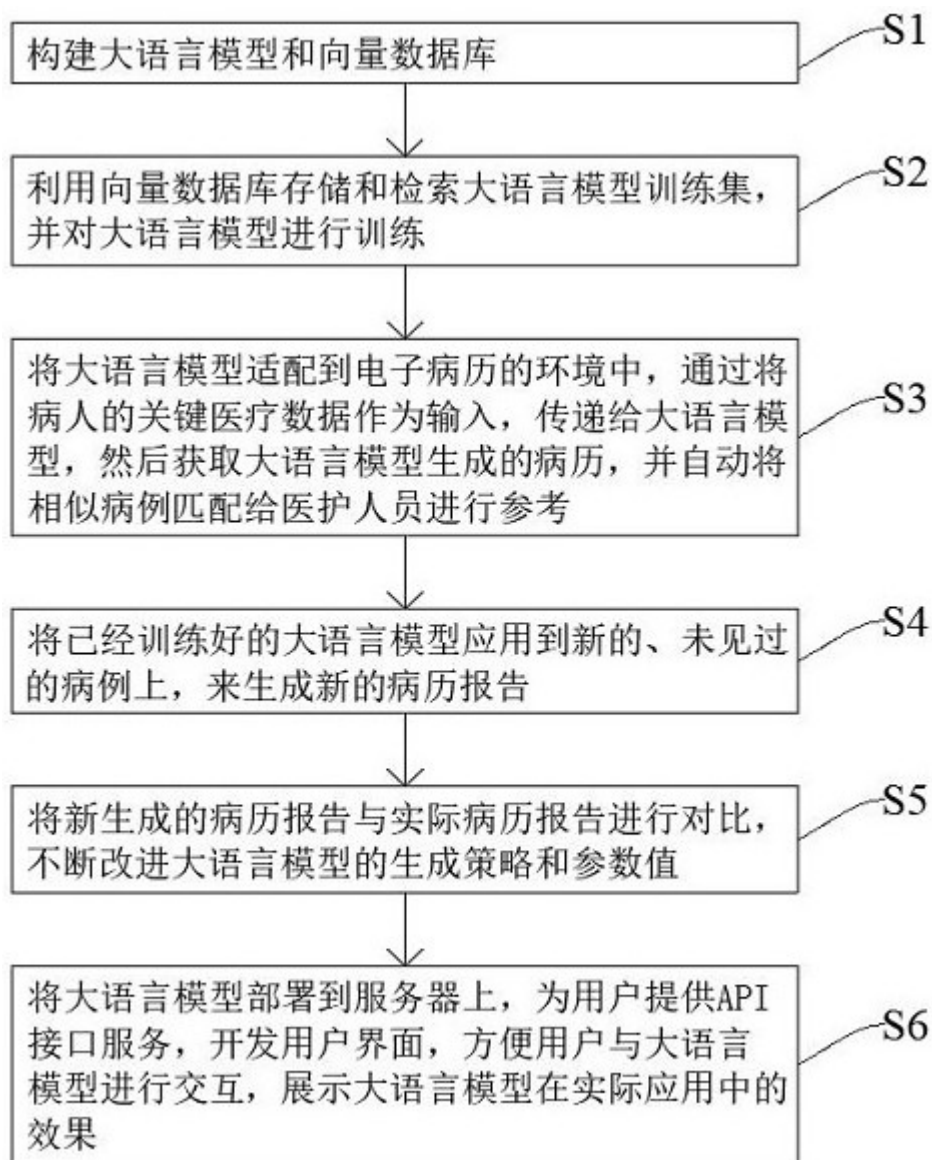


图 1

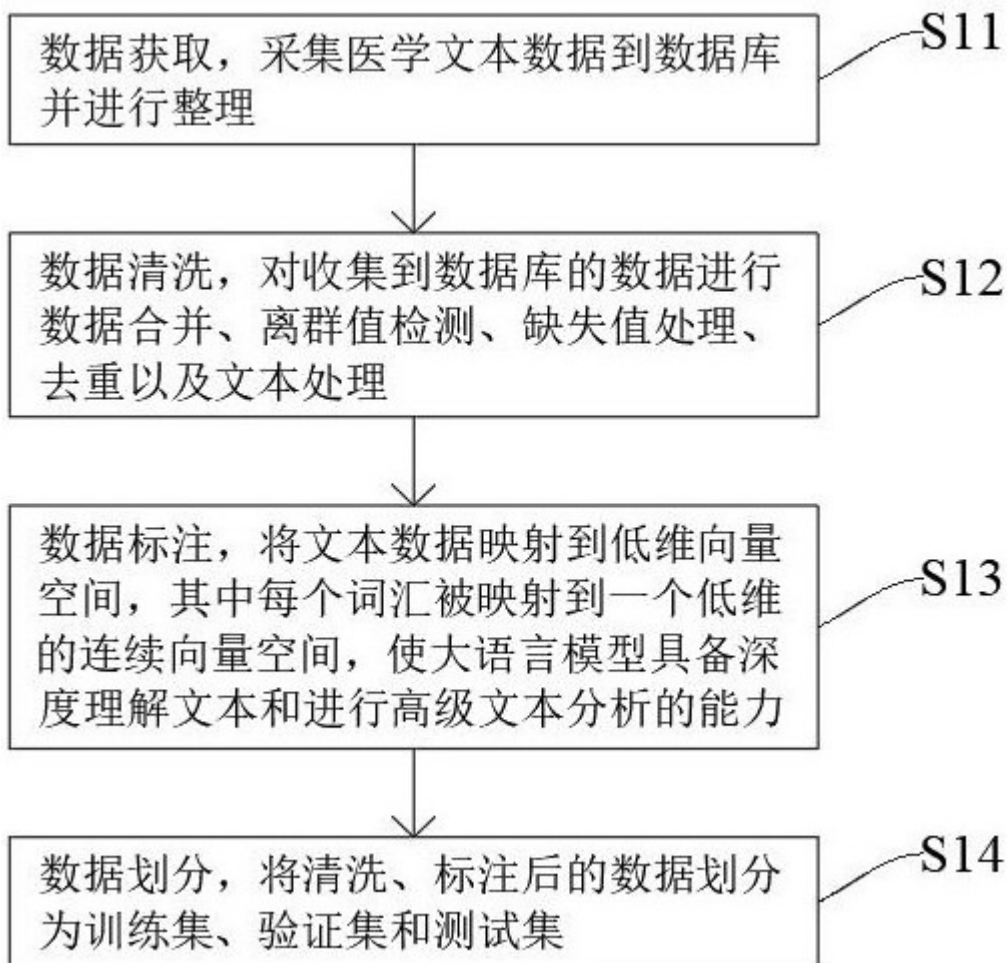


图 2

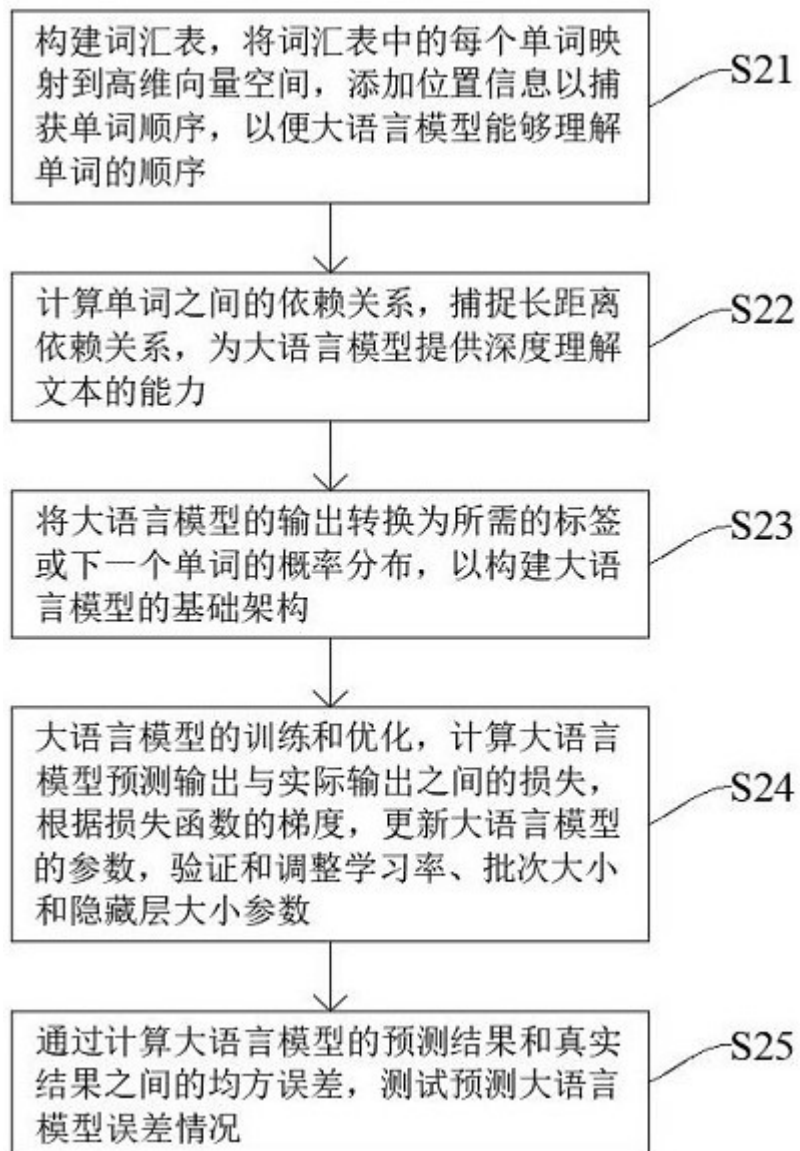


图 3