

语言转义方法、装置、设备及存储介质

本申请要求于 2022 年 07 月 22 日提交中国专利局、申请号为
5 202210873471.8、申请名称为“语言转义方法、装置、设备及存储介质”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

本申请涉及计算机技术领域，尤其涉及自然语言转义方法、装置、设
10 备及存储介质。

背景技术

数据库作为很多数据服务的核心基础技术，为各种数据服务需求提供
支持。在数据库的管理和使用当中，往往需要专业的人员参与数据库的使
15 用、交互工作。

在现有技术中，虽然有一些技术实现了通过人类自然语言到数据库中
进行数据查询。比如，可以是用户输入自然语言给神经网络，进而由神经
网络输出对应的查询语句（比如，结构化查询语句（Structured Query
Language, SQL））。然而，这些查询语句的可执行效果不佳。
20

发明内容

为解决或改善现有技术中存在的问题，本申请各实施例提供了语言转
义方法、装置、设备及存储介质。

第一方面，在本申请的一个实施例中，提供了一种语言转义方法。该
25 方法包括：

- 对获取到的目标语句进行拆分，得到至少一个目标子语句；
- 确定所述目标子语句中包含的关键词类型和关键词子句；
- 根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所
述目标子语句中的关键词子句进行修正；
- 30 对修正后的目标子语句进行组合，生成可被执行的机器语言。

第二方面，在本申请的一个实施例中，提供了一种电子设备，包括存
储器及处理器；其中，
所述存储器，用于存储程序；

所述处理器，与所述存储器耦合，用于执行所述存储器中存储的所述程序，以用于实现第一方面所述的一种语言转义方法。

第三方面，在本申请的一个实施例中，提供了一种非暂时性机器可读存储介质，所述非暂时性机器可读存储介质上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行如第一方面所述的一种语言转义方法。

第四方面，在本申请的一个实施例中，提供了一种数据库系统，所述系统包括：

编码器，用于获取到的自然语言进行编码处理得到的语义特征向量；
10 解码器，用于所述编码器输出的所述语义特征向量进行解码处理，生成目标语句；

拆分模块，用于对获取到的所述目标语句进行拆分，得到至少一个目标子语句；

修正模块，用于根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中的关键词子句进行修正；
15

组合模块，用于对修正后的目标子语句进行组合，生成可被执行的机器语言；

数据库，用于基于所述可被执行的机器语言执行数据查询任务。

本申请实施例提供的技术方案中，当用户提供查询作用的自然语言后，
20 可以由编码器对自然语言进行编码处理从而得到该自然语言的语义特征向量。利用解码器对语言特征向量进行进一步处理，得到自然语言对应的可被数据库识别的目标语句。对目标语句进行拆分处理，得到包含至少一个目标子语句的序列。在每个目标子语句中都包含有关键词，在对目标子语句进行修正的时候，基于目标子语句中包含的关键词类型和对应的语义约束规则，对目标子语句进行修正，并对修正后的目标子语句进行组合，得到能够被数据库识别的机器语言。通过上述方式，能够有效提高自然语言转义准确性以及用户使用数据库查询的工作效率，同时降低对使用数据库查询工作人员的技术能力要求。

30 附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，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。

35 图 1 为本申请实施例提供的语言转义方法的流程示意图；

图 2 为本申请实施例提供的语言转义系统的结构示意图；

- 图 3 为本申请实施例提供的对目标语句进行拆分方法的流程视图；
图 4 为本申请实施例举例说明的数据表结构示意图；
图 5 为本申请实施例举例说明的有向图的示意图；
图 6 为本申请实施例提供的一种语言转义装置的结构示意图；
5 图 7 为本申请实施例提供的一种电子设备的结构示意图；
图 8 为本申请实施例提供的一种数据库系统的示意图。

具体实施方式

为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请
10 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。

在本申请的说明书、权利要求书及上述附图中描述的一些流程中，包
含了按照特定顺序出现的多个操作，这些操作可以不按照其在本文中出现的
顺序来执行或并行执行。操作的序号如 101、102 等，仅仅是用于区分各
个不同的操作，序号本身不代表任何的执行顺序。另外，这些流程可以包
15 括更多或更少的操作，并且这些操作可以按顺序执行或并行执行。需要说
明的是，本文中的“第一”、“第二”等描述，是用于区分不同的消息、设备、
模块等，不代表先后顺序，也不限定“第一”和“第二”是不同的类型。此外，
下文描述的实施例仅仅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，本领域技术人员在没有做出创造性劳动前提下所获
20 得的所有其他实施例，都属于本申请保护的范围。

数据库是信息产业绝大多数数据服务的核心基础设施，数据处理过程
中时刻需要数据库的支撑。现阶段终端用户和数据库运维人员仍然主要通
过数据库查询机器语言（如 SQL、Gremlin 等）与数据库进行交互，这需
要使用者有较强的数据库专业知识、运维经验、以及对相关数据库实例有
25 较深的了解。因此，在数据库的管理、使用、运维过程中需要大量专业工
作人员参与到数据库交互工作中，技术门槛要求高，人工成本也高。随着
人工智能技术的发展，可以利用机器学习技术，将工作人员提供的自然语
言转换成为机器语言，以便计算机能够理解工作人员真实意图，并执行相
应任务。然而，在实际应用中，利用自然语言处理、机器学习算法等技术
30 实现自然语言到数据库查询语言的转义（NL2SQL）所得到的机器语言往
往存在一些语法错误、语义错误等问题，导致数据库没法很好的执行转义
得到的机器语言，还需要具有一定技术基础的工作人员对得到的机器语言
进行调整后才能够被顺利执行。因此，本申请方案提出对经过转义得到
的机器语言进行优化调整，从而实现工作人员通过自然语言与数据库的无障
35 碍交互的方案。转义得到的机器语言准确率更高，更容易被成功执行。能

够有效降低对工作人员技术能力要求，降低人工成本，提升工作效率。

在本申请技术方案中，具体工作过程，将在下述实施例中说明。

如图 1 为本申请实施例提供的语言转义方法的流程示意图。该语言转义方法的执行主体可以是本地计算机、云服务设备等。如图 1 所示，该语言转义方法包括如下步骤：

步骤 101：对获取到的目标语句进行拆分，得到至少一个目标子语句。

步骤 102：确定所述目标子语句中包含的关键词类型和关键词子句；

步骤 103：根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中的关键词子句进行修正。

步骤 104：对修正后的目标子语句进行组合，生成可被执行的机器语言。

这里所说的目标语句可以理解为利用编码器和解码器进行编码以及解码处理后所得到的机器语言（比如，SQL）。一般来说，在得到的目标语句当中可能会包含有多个具有嵌套关系的目标子语句。

在对目标语句进行修正的时候，需要对目标语句进行拆分，使得拆分后所得到的目标子语句均不具有嵌套关系。换言之，若经过一次拆分后得到两个目标子语句，其中一个目标子语句中还包含有嵌套关系，则需要针对该目标子语句进行进一步拆分，直到不再包含有任何嵌套关系为止。具体拆分过程和拆分方案，将在下述实施例中进行举例说明。

这里所说的关键词类型包括：SELECT 类型、WHERE 类型、FROM 类型、HAVING 类型、ORDER BY 类型、GROUP BY 类型、LIMIT 类型等。这里所说的关键词子句，可以理解为关键词及其上下文关联得到的句子。在一个目标子语句中，可能同时包含有多种的关键词类型和对应的多个关键词子句。关键词类型可以根据用户需要进行定义、调整的，比如，可以新增某个关键词类型或减少某个关键词类型。

需要说明的是，不同关键词类型对应不同的语义约束规则。这里所说的语义约束规则可以理解为不同计算机语言中所采用的语法所具有的语义约束规则。比如，某类值能够支持哪些类型操作符，哪些词或句所支持的嵌套关系，某个操作符在语句中的位置等等。在实际应用中，为了获得更加准确的修正效果，以关键词子句作为最小修正单元根据语义约束规则进行修正。在修正的时候，可以按照目标子语句中的各个关键词的逻辑关系进行修正，也可以按照先后顺序进行修正。

在对各个目标子语句中的关键词子句进行修正后，得到修正后的目标子语句，进而，对多个修正后的目标子语句进行组合。在组合的时候，按拆分是的嵌套关系进行组合，得到机器语言（比如，SQL 等）。

在本申请的一个或者多个实施例中，所述对获取到的目标语句进行拆分，得到至少一个目标子语句，包括：利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句；根据嵌套关系对所述目标语句进行拆分，生成无嵌套关系的所述至少一个目标子语句。

如图 2 为本申请实施例提供的语言转义系统的结构示意图。下面将结合附图 2 对本申请技术方案的整体架构进行说明。在本申请技术方案中，包括两个模型编码器和解码器。其中，编码器模型通常由长短期记忆网络（Long Short-Term Memory, LSTM）、Transformer 等网络结构构成，将输入的自然语言转化为一系列的抽象的语义特征向量。解码器通常使用另一个独立的 LSTM 或 Tree-LSTM 构成。需要说明的是，这里所说的编码器和解码器都是预先训练好的各种神经网络模型。

当用户或工作人员通过口语方式或者文字输入方式提供自然语言（这里所说的自然语言通常是指一种自然地随文化演化的语言。例如，汉语、英语、日语等等），进一步将自然语言以文字形式输入到编码器当中，输出语义特征向量。

在编码器输出语义特征向量之后，将该语义特征向量输入到解码器当中，进一步由解码器对该语义特征向量进行解码处理，得到目标语句。在目标语句中，包含有至少一个关键词，以及操作符（比如，=、!=、==等）和预测值（利用机器学习模型预测得到的值）。

如前文所述可知，在得到目标语句之后，将进一步对目标语句进行拆分，拆分的最终目的是使得每个目标子语句都不再有嵌套关系。当然，若目标语句不包含有任何嵌套关系，则可以不对其进行拆分，将该不具有任何嵌套关系的语句作为子语句进行处理，根据其中包含的关键字类型对应的语义约束规则，对关键字子句进行修正。

例如，用户输入的自然语言问题为：“告诉我官方语言不为英语的国家的居民平均寿命分别是多少？”。直接使用解码器（比如，神经网络模型）生成的目标语句 SQL 为：SELECT avg(life_expectancy) FROM country WHERE country_name != (SELECT isofficial FROM country_language WHERE language != “英语”)。在该目标语句中，可以看到包括关键词类型有：SELECT 类型、FROM 类型、WHERE 类型。其中，外层目标子语句中得到的关键词子句有：SELECT avg(life_expectancy)、FROM country、WHERE country_name !=。在 WHERE 子句中操作符有!=，预测值 country_name。

对上述目标语句进行拆分后，得到不具有嵌套关系的两个目标子语句

分别为：①SELECT avg (life_expectancy) FROM country WHERE country_name !=和②SELECT isofficial FROM country_language WHERE language != “英语”。

在本申请的一个或者多个实施例中，如图3为本申请实施例提供的对目标语句进行拆分方法的流程视图。从图3中可以看到，所述根据嵌套关系对所述目标语句进行拆分，生成无嵌套关系的所述至少一个目标子语句，包括如下步骤：

步骤301：根据所述目标语句中包含的嵌套标识，确定存在嵌套关系的第一目标子语句和第二目标子语句；其中，所述第二目标子语句嵌套在所述第一目标子语句中。

步骤302：在所述第一目标子语句中所述嵌套标识所在的嵌套位置添加位置标识。

步骤303：生成由所述第一目标子语句和所述第二目标子语句按照嵌套关系中的层级顺序组成的序列。

在实际应用中，对目标语句进行拆分的时候，基于嵌套标识进行拆分。这里所说的嵌套标识，可以是特定符号，比如括号（）、{}等。还可以是集合操作符，比如，INTERSECT、UNION、EXCEPT等。

由于第二目标子语句嵌套在第一目标子语句中，若通过对第二目标子语句进行检查，发现其中还包含有其他子语句，则还需要对第二目标子语句进行进一步拆分。反之，若第二目标子语句中不再包含有任何嵌套内容，则不需要对第二目标子语句进行进一步拆分。

继续假设在第二目标子语句中不再包含任何嵌套关系，则需要第一目标子语句中嵌套标识的位置添加位置标识，以便后续对第一目标子语句和第二目标子语句进行组合。在组合的是，需要对修正后的按照此前嵌套关系进行复原，因此，需要根据位置标识将修改后的第一目标子语句和第二目标子语句进行组合。

经过拆分可以得到第一目标子语句和第二目标子语句，可以进一步基于第一目标子语句和第二目标子语句之间的嵌套关系中的层级顺序进行组合，得到子语句序列。

例如，用户输入的自然语言问题为：“告诉我官方语言不为英语的国家的居民平均寿命分别是多少？”。直接使用解码器（比如，神经网络模型）生成的目标语句SQL为：SELECT avg (life_expectancy) FROM country WHERE country_name != (SELECT isofficial FROM country_language WHERE language != “英语”)。在该目标语句中，可以看到包括关键词类型有：SELECT 类型、FROM 类型、WHERE 类型。

对上述目标语句进行拆分后，得到不具有嵌套关系的两个目标子语句分别为：①SELECT avg (life_expectancy) FROM country WHERE country_name !=和②SELECT isofficial FROM country_language WHERE language != “英语”。由于第二目标子语句嵌入到第一目标子语句当中，
5 因此，可以得到序列为[SELECT avg(life_expectancy) FROM country WHERE country_name != NESTEDVALUE,SELECT isofficial FROM country_language WHERE language != “英语 “]。其中，NESTEDVALUE可以作为位置标识。

然而，此 SQL 语句是不能执行的，因为存在多个语法、语义错误，正
10 确的 SQL 英为：SELECT avg(life_expectancy) FROM country WHERE country_name not in (SELECT country_name FROM country_language WHERE language != “English” and isofficial = 1)。在下述实施例中，将对修正过程进行具体说明。

在本申请的一个或者多个实施例中，得到至少一个目标子语句之后，
15 还包括：根据所述目标子语句中包含的关键词类型，确定对应的所述语义约束规则。

不同的关键词类型所采用的用于对关键词子句进行修正的语法约束规则不同。举例来说：WHERE 子句中的预测值进行修正，将预测值与数据库中存储的数据表中的值进行相似度计算，从而判断是否需要预测值进行修改。
20 HAVING 子句中，在确定聚合函数与列名匹配的情况下，根据聚合函数不同类型采用不同的修正方式，比如，若聚合函数为 max、min 等保持数值类型不变的函数，处理方式可以是将预测值与数据库中存储的数据表中的值进行相似度计算，从而判断是否需要预测值进行修改；若聚合函数为 count、avg 等聚合结果为计算数值，则将预测值替换为常量，
25 比如，将预测值 single 替换为 1。在无法合理转义的情况下，则可以舍弃该 HAVING 子句。

需要说明的是，在修正的时候，需要借助数据库进行修正。在一可选实施例当中，可以在数据库中建立数据表。如图 4 为本申请实施例举例说明的数据表结构示意图。从图 4 中可以看到，在数据表 1 中存储有列名、
30 值等。其中，列名可以根据需要设定，比如，列名为 country_name 等，对应的值有：China、American、Japan 等。在数据表 2 中存储有列名、值，比如，列名为 city_name, country_name, ...，对应的第一行值：Beijing, China, ...；第二行值：Shanghai, China, ...；第三行值：New York, USA, ...等。

35 在本申请的一个或者多个实施例中，当所述关键词类型为 SELECT 类

型时，所述利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句，包括：获取数据库提供的数据库表；对所述语义特征向量进行解码得到的所述 SELECT 关键词关联的操作内容与所述数据库表中数据列进行匹配度分析；

- 5 根据所述匹配度大小从所述数据库表中确定与所述操作内容匹配的目标数据列；基于所述目标数据列中的数据，生成包含 SELECT 关键词、所述操作内容和所述预测值的目标语句。

10 为了使得解码器在解码过程中所得到更加准确的解码结果，可以将解码得到的聚合函数与数据库表中的列名进行匹配度分析，例如，列名 varchar 与聚合函数 max/min 不匹配，可以将该位置匹配度设置概率值为零；若该聚合函数 max/min 与某个列名匹配，根据匹配度设置对应的概率值；进而，从多个列中选择概率值最大的。并根据所选择的列中所包含的值和/或操作内容生成 SELECT 子句。

15 在本申请的一个或者多个实施例中，当所述关键词类型为 WHERE 类型时。所述利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句，包括：获取数据库提供的数据库表；根据所述语义特征信息进行解码得到的 WHERE 关键词关联的预测值，从所述数据库表中筛选出操作内容；根据所述预测值与所述筛选出操作内容的匹配度大小，确定与所述预测值匹配的
20 所述筛选出操作内容，以得到包含 WHERE 关键词、所述预测值和所述筛选出操作内容的目标语句。

在解码过程中，若解码得到 WHERE 子句，其中包含有预测值。根据 WHERE 关键词关联的预测值，选择对应的数据库表，进而计算预测值与数据库表中的操作符的匹配度，若不匹配，则将对数据库表中数据列位置设置
25 概率值为零。进而，将预测值与其他数据列中的操作符进行匹配度计算，从而选择出概率值高的数据列以便得到所需的 WHERE 子句。在实际应用中，计算预测值与操作符之间匹配度的时候，可以根据预测值的类型对操作符进行进一步限制。例如，得到的预测值为字符串 abc，则限定的操作符范围为 {=, <>, like, not like}，其他类型操作符的位置概率值置为 0。

30 下面将结合具体实施例对各种关键字类型对应的关键字子句修正过程进行说明。

当关键词类型为 WHERE 类型时，若所述预测值为非嵌套值，则计算所述预测值的第一特征向量与所述数据库表中第一数据列中任意一个值的第二特征向量之间的第一相似度值；若所述第一相似度值大于第一阈值，则
35 用所述任意一个值替换所述预测值以得修正后的目标子语句。若所述相似

度值小于第一阈值，则计算所述数据表中除所述第一数据列之外的任意一列中任意一个值的第三特征向量与所述第一特征向量之间的第二相似度值；若所述第二相似度值大于第二阈值，则将所述任意一列中任意一个值替换所述预测值以得到修正后的目标子语句。

- 5 举例来说，当 WHERE 类型子句中的预测值为非嵌套值，比如预测值为字符串，则获取其对应的 word2vec 向量，并与数据表中对应数据列的所有值的 word2vec 向量计算余弦相似度，并找到余弦相似度最大的值。若最大余弦相似度大于第一阈值 M1，则用数据表中的值替换预测值；若数据表对应数据列中所有值与预测值的余弦相似度值均小于第一阈值 M1，则
- 10 对比数据表中其他数据列的值，若存在某一数据列的值与预测值的余弦相似度大于第二阈值 M2，则用该数据列中的值替换该预测值，并将 WHERE 子句中的列名修改为该数据流的列名，若所有值与预测值的余弦相似度均小于 M2，则保持原语句不变。

- 当所述关键词类型为 HAVING 类型时，若执行所述操作内容时的执行
- 15 结果为数值常量，则将 HAVING 子句中的所述预测值替换为常量。例如，当操作内容为聚合函数 count、avg 等，执行该聚合函数后所得到的执行结果是数值形式的常量，则在修正的时候，可以对与该操作内容（聚合函数）具有上下文关系的预测值替换为常量（比如，1）。

- 当执行操作内容不会改变预测值的时候，若所述预测值为非嵌套值，
- 20 则计算所述预测值的第一特征向量与所述数据表中第一数据列中任意一个值的第二特征向量之间的第一相似度值；若所述第一相似度值大于第一阈值，则用所述任意一个值替换所述预测值以得修正后的目标子语句。若所述相似度值小于第一阈值，则计算所述数据表中除所述第一数据列之外的任意一列中任意一个值的第三特征向量与所述第一特征向量之间的第二相似度值；若所述第二相似度值大于第二阈值，则将所述任意一列中任意一个值替换所述预测值以得到修正后的目标子语句。
- 25

- 举例来说，若 HAVING 子句中的聚合函数为 max、min 等保持类型不变的函数的时候，计算预测值（比如，字符串 abc）的 word2vec 向量，并与数据表中对应数据列的所有值的 word2vec 向量计算余弦相似度，找到计
- 30 算后余弦相似度值最大的值。若余弦相似度最大值大于第一阈值 M1，则用数据表中的值替换预测值；若数据表中对应数据列中所有值与预测值的余弦相似度值均小于第一阈值 M1，则对比数据表中其他数据列的值，若存在某一数据列的值与预测值的余弦相似度大于第二阈值 M2，则用该数据列中的值替换该预测值，并将 WHERE 子句中的列名修改为该数据流的
- 35 列名，若所有值与预测值的余弦相似度均小于 M2，则保持原语句不变。

在本申请的一个或者多个实施例中，所述根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中各所述关键词子句进行修正，包括：当所述关键词类型为 ORDER BY 类型时，若所述目标子语句中除 ORDER BY 子句外的其他子句包含有聚合函数，则将所述

5 ORDER BY 子句转换为嵌套查询语句。

ORDER BY 用于对查询结果进行排序。在对 ORDER BY 子句修正时，需要检查当前目标子语句当中的其他关键词子句是否包含有聚合函数、ORDER BY 子句中是否含有聚合函数。若目标子语句中其他关键词子句不具有聚合函数，或者目标子语句中的其他关键词子句与 ORDER BY 子句同时具有聚合函数且有 GROUP BY 语句，则可以不对 ORDER BY 子句做任何修正处理。

若目标子语句中除了 ORDER BY 子句外的其他关键词子句都含有聚合函数，则需要修正处理。举例来说，若其他子句中包含聚合函数，则将其转化为嵌套查询，转化的规则为：在 WHERE 子句的部分中增加一条子句，形式为 WHERE c = (SELECT c ORDER BY c)，其中 c 为原 ORDER BY 语句中的列名。若 ORDER BY 中包含聚合函数且没有 GROUP BY 语句，则删除 ORDER BY 语句。

在本申请的一个或者多个实施例中，所述根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中各所述关键词子句进行修正，包括：当所述关键词类型为 GROUP BY 类型时，若 SELECT 子句不包含聚合函数和 GROUP BY 子句对应的数据列的列名，则将所述 GROUP BY 的列名替换为主键；若未查找到所述主键，则将所述 SELECT 子句对应的数据列复制到所述 GROUP BY 对应的数据列。

如前文所述，为了提升对自然语言的转义速度和效率，会预先建立用于数据表，在数据表中存储有数据列，包含列名和多个值。这些值有的可以作为外键，同时作为另一个数据表的主键，从而方便两个数据表之间加强链接，方便数据的查找、管理。如图 4 所示，在数据表 1 中，country_name 为数据表 1 的主键，同时 country_name 为数据表 2 的外键，主键和外键在这里（关系型数据库）本质上是同一列。从而使得数据表 1 与数据表 2 建立关联关系。

举例来说，若 SELECT 子句中存在不包含聚合函数且不出现在 GROUP BY 中的列名，则将 GROUP BY 的列名替换为 Unique Key（如主键等）。若没有 Unique Key，则将 SELECT 子句中的列复制到 GROUP BY 子句中。

在本申请的一个或者多个实施例中，当所述关键词类型为 FROM 类型时，基于所述修正后的目标子语句对应的数据表和其中的数据列生成有向

图；基于所述有向图中的所述数据表和所述数据列，以所述数据列或外键为根节点构造最小生成树。

例如，如图 5 为本申请实施例举例说明的有向图的示意图。从图 5 中可以看到，首先构造关系模式图（Schema Graph），关系模式图中的每个节点为一个表或一列，标准的模式图有四种形式：

- a. 起始节点为一列，终止节点为列所属的表
- b. 起始节点为外键，终止节点为对应的主键
- c. 与 a 反向的边
- d. 与 b 反向的边

将解码器预测的 FROM 子句中的表名和所有其他关键字子句涉及的表作为节点构造最小生成树，为了避免歧义，树中使用的边只有 a 和 b 这两种形式。在得到树之后，需要进一步对其进行剪枝处理，不断删除生成树中所有出度为 0 的表直至树中所有表的出度均大于 0。

在本申请的一个或者多个实施例中，所述对修正后的目标子语句进行组合，包括：若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句具有相同的列名或主外键关系，则对修改后的目标子语句进行组合；或者，若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句不具有相同的列名或主外键关系，则将 WHERE 子句对应的列名复制到所述 SELECT 子句，对修改后的目标子语句进行组合。

例如，在完成对目标语句中各个关键字子句的修正操作后，进一步对这些修正后的关键字语句重新组合得到修正后的目标语句。对于 WHERE 子句中的嵌套查询内层的 SQL 需要保证 SELECT 子句的列名与外层的条件（WHERE 子句）的列名相同或存在主键-外键关系，若不满足上述条件，则把外层的列名复制到内层的 SELECT 子句。特别地，若内层 SELECT 子句中的列为 boolean 属性，则在内层 WHERE 子句中添加条件：WHERE c = 1，其中 c 为该 boolean 属性的列名。在进行组合的时候，可以根据拆分时候标记的位置标识进行组合，能够有效提高组合效率。

为了便于理解，下面将通过具体实施例对本方案进行举例说明。

首先，利用编码器和解码器中预先训练好的神经网络模型直接预测的 SQL：SELECT avg (life_expectancy) FROM country WHERE country_name != (SELECT isofficial FROM country_language WHERE language != “英语”)会被拆分为序列：

[
SELECT avg(life_expectancy) FROM country WHERE country_name !=
NESTEDVALUE,

SELECT isofficial FROM country_language WHERE language != “英语”
“
]

对于序列中第一个目标子语句中的每个关键词子句进行处理，其中
5 WHERE 子句中 country_name != NESTEDVALUE 中的 != 由于操作符与值
类型不匹配，因此 != 会被重新预测为 not in。

对序列中第二个目标子语句中的每个关键词子句进行处理，其中
WHERE 子句中的 language != “英语” 中的值 “英语” 会通过词向量相似
度计算替换为 “English”。

10 拼接两个修正后关键词子句，其中内层 SELECT 子句中的 isofficial
与外层条件中的列名 country_name 不同且不具备主键外键关系，因此将
country_name 复制到内层 SELECT 子句中，并且由于 isofficial 是 boolean
类型，因此在内层 WHERE 子句中加入条件 isofficial=1。

最终得到修正后的目标语句，也就是可被计算机执行的机器语言：

15 SELECT avg(life_expectancy) FROM country WHERE country_name not in
(SELECT country_name FROM country_language WHERE language !=
“English” and isofficial = 1)。通过上述方式，对编码、解码处理得到的
SQL 语句进行拆分，然后，针对拆分得到的关键词子句进行修正，由于每
个关键词都有其特有的执行语义约束规则，这些规则是由机器语言特性所
20 决定的，而且这些规则在自然语言中是不具的，无法通过自然语句转义直
接得到的。此外，由于规则不具有很强的规律性，关键词种类也很多，是
神经网络模型无法学习到的。因此，为了确保 SQL 语句能够被正确执行，
具体针对每个关键词子句中的值进行逐一修正，使得修正效果更准确。

25 基于同样的思路，本申请实施例还提供一种语言转义装置。如图 6 为
本申请实施例提供的一种语言转义装置的结构示意图。该语言转义装置包
括：

拆分模块 61，用于对获取到的目标语句进行拆分，得到至少一个目标
子语句。

30 确定模块 62，用于确定所述目标子语句中包含的关键词类型和关键词
子句。

修正模块 63，用于根据所述目标子语句中包含的关键词类型对应的语
义约束规则，对所述目标子语句中的关键词子句进行修正。

35 组合模块 64，用于对修正后的目标子语句进行组合，生成可被执行的
机器语言。

可选地，拆分模块 61，用于利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句；

5 根据嵌套关系对所述目标语句进行拆分，生成无嵌套关系的所述至少一个目标子语句。

可选地，拆分模块 61，用于根据所述目标语句中包含的嵌套标识，确定存在嵌套关系的第一目标子语句和第二目标子语句；其中，所述第二目标子语句嵌套在所述第一目标子语句中；

10 在所述第一目标子语句中所述嵌套标识所在的嵌套位置添加位置标识；

生成由所述第一目标子语句和所述第二目标子语句按照嵌套关系中的层级顺序组成的序列。

可选地，确定模块 62，还用于根据所述目标子语句中包含的关键词类型，确定对应的所述语义约束规则。

15 可选地，还包括解码模块 65，当所述关键词类型为 SELECT 类型时，解码模块 65 用于获取数据库提供的数据表；

对所述语义特征向量进行解码得到的所述 SELECT 关键词关联的操作内容与所述数据表中数据列进行匹配度分析；

20 根据所述匹配度大小从所述数据表中确定与所述操作内容匹配的目标数据列；

基于所述目标数据列中的数据，生成包含 SELECT 关键词、所述操作内容和所述预测值的目标语句。

可选地，当所述关键词类型为 WHERE 类型时；

解码模块 65 用于获取数据库提供的数据表；

25 根据所述语义特征信息进行解码得到的 WHERE 关键词关联的预测值，从所述数据表中筛选出操作内容；

根据所述预测值与所述筛选出操作内容的匹配度大小，确定与所述预测值匹配的所述筛选出操作内容，以得到包含 WHERE 关键词、所述预测值和所述筛选出操作内容的目标语句。

30 修正模块 63，用于当所述关键词类型为 HAVING 类型时，若执行所述操作内容时的执行结果为数值常量，则将 HAVING 子句中的所述预测值替换为常量。

修正模块 63，用于若所述预测值为非嵌套值，则计算所述预测值的第一特征向量与所述数据表中第一数据列中任意一个值的第二特征向量之间的第一相似度值；

35

若所述第一相似度值大于第一阈值，则用所述任意一个值替换所述预测值以得修正后的目标子语句。

修正模块 63，用于若所述相似度值小于第一阈值，则计算所述数据表中除所述第一数据列之外的任意一列中任意一个值的第三特征向量与所述
5 第一特征向量之间的第二相似度值；

若所述第二相似度值大于第二阈值，则将所述任意一列中任意一个值替换所述预测值以得到修正后的目标子语句。

修正模块 63，用于当所述关键词类型为 ORDER BY 类型时，若所述目标子语句中除 ORDER BY 子句外的其他子句包含有聚合函数，则将所述
10 ORDER BY 子句转换为嵌套查询语句。

修正模块 63，用于当所述关键词类型为 GROUP BY 类型时，若 SELECT 子句不包含聚合函数和 GROUP BY 子句对应的数据列的列名，则将所述 GROUP BY 的列名替换为主键；

若未查找到所述主键，则将所述 SELECT 子句对应的数据列复制到所述
15 述 GROUP BY 对应的数据列。

可选地，还包括生成模块 66，用于当所述关键词类型为 FROM 类型时，基于所述修正后的目标子语句对应的数据表和其中的数据列生成有向图；

基于所述有向图中的所述数据表和所述数据列，以所述数据列或外键
20 为根节点构造最小生成树。

可选地，组合模块 64 用于若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句具有相同的列名或主外键关系，则对修改后的目标子语句进行组合；或者，若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句不具有相同的列名或主外键关系，则将 WHERE 子句对应的
25 列名复制到所述 SELECT 子句，对修改后的目标子语句进行组合。

本申请一个实施例还提供一种电子设备。如图 7 为本申请实施例提供的一种电子设备的结构示意图。该电子设备包括存储器 701、处理器 702 及通信组件 703；其中，

30 所述存储器 701，用于存储程序；

所述处理器 702，与所述存储器耦合，用于执行所述存储器中存储的所述程序，以用于：

对获取到的目标语句进行拆分，得到至少一个目标子语句；

确定所述目标子语句中包含的关键词类型和关键词子句；

35 根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所

述目标子语句中的关键词子句进行修正；

对修正后的目标子语句进行组合，生成可被执行的机器语言。

可选地，处理器 702 用于利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句；

根据嵌套关系对所述目标语句进行拆分，生成无嵌套关系的所述至少一个目标子语句。

可选地，处理器 702 用于根据所述目标语句中包含的嵌套标识，确定存在嵌套关系的第一目标子语句和第二目标子语句；其中，所述第二目标子语句嵌套在所述第一目标子语句中；

在所述第一目标子语句中所述嵌套标识所在的嵌套位置添加位置标识；

生成由所述第一目标子语句和所述第二目标子语句按照嵌套关系中的层级顺序组成的序列。

可选地，处理器 702 用于根据所述目标子语句中包含的关键词类型，确定对应的所述语义约束规则。

可选地，当所述关键词类型为 SELECT 类型时，处理器 702 用于获取数据库提供的数据表；

对所述语义特征向量进行解码得到的所述 SELECT 关键词关联的操作内容与所述数据表中数据列进行匹配度分析；

根据所述匹配度大小从所述数据表中确定与所述操作内容匹配的目标数据列；

基于所述目标数据列中的数据，生成包含 SELECT 关键词、所述操作内容和所述预测值的目标语句。

可选地，当所述关键词类型为 WHERE 类型时；

处理器 702 用于获取数据库提供的数据表；

根据所述语义特征信息进行解码得到的 WHERE 关键词关联的预测值，从所述数据表中筛选出操作内容；

根据所述预测值与所述筛选出操作内容的匹配度大小，确定与所述预测值匹配的所述筛选出操作内容，以得到包含 WHERE 关键词、所述预测值和所述筛选出操作内容的目标语句。

处理器 702 用于当所述关键词类型为 HAVING 类型时，若执行所述操作内容时的执行结果为数值常量，则将 HAVING 子句中的所述预测值替换为常量。

处理器 702 用于若所述预测值为非嵌套值，则计算所述预测值的第一

特征向量与所述数据表中第一数据列中任意一个值的第二特征向量之间的第一相似度值；

若所述第一相似度值大于第一阈值，则用所述任意一个值替换所述预测值以得修正后的目标子语句。

- 5 处理器 702 用于若所述相似度值小于第一阈值，则计算所述数据表中除所述第一数据列之外的任意一列中任意一个值的第三特征向量与所述第一特征向量之间的第二相似度值；

若所述第二相似度值大于第二阈值，则将所述任意一列中任意一个值替换所述预测值以得到修正后的目标子语句。

- 10 处理器 702 用于当所述关键词类型为 ORDER BY 类型时，若所述目标子语句中除 ORDER BY 子句外的其他子句包含有聚合函数，则将所述 ORDER BY 子句转换为嵌套查询语句。

处理器 702 用于当所述关键词类型为 GROUP BY 类型时，若 SELECT 子句不包含聚合函数和 GROUP BY 子句对应的数据列的列名，则将所述

- 15 GROUP BY 的列名替换为主键；

若未查找到所述主键，则将所述 SELECT 子句对应的数据列复制到所述 GROUP BY 对应的数据列。

处理器 702 用于当所述关键词类型为 FROM 类型时，基于所述修正后的目标子语句对应的数据表和其中的数据列生成有向图；

- 20 基于所述有向图中的所述数据表和所述数据列，以所述数据列或外键为根节点构造最小生成树。

处理器 702 用于若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句具有相同的列名或主外键关系，则对修改后的目标子语句进行组合；或者，

- 25 若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句不具有相同的列名或主外键关系，则将 WHERE 子句对应的列名复制到所述 SELECT 子句，对修改后的目标子语句进行组合。

- 30 上述存储器 701 可被配置为存储其它各种数据以支持在电子设备上的操作。这些数据的示例包括用于在电子设备上操作的任何应用程序或方法的指令。存储器可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器（SRAM），电可擦除可编程只读存储器（EEPROM），可擦除可编程只读存储器（EPROM），可编程只读存储器（PROM），只读存储器（ROM），磁存储器，快闪存储器，磁盘或光盘。

- 35 进一步地，本实施例中的所述处理器 702 可以具体是：可编程交换处

理芯片，该可编程交换处理芯片中配置有数据复制引擎，能对接收到的数据进行复制。

上述处理器 702 在执行存储器中的程序时，除了上面的功能之外，还可实现其它功能，具体可参见前面各实施例的描述。进一步，如图 7 所示，
5 电子设备还包括：电源组件 704 等其它组件。

本申请实施例还提供一种非暂时性机器可读存储介质，所述非暂时性机器可读存储介质上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行图 1 对应实施例所述的方法。

本申请实施例还提供一种计算机程序产品，包括计算机程序/指令，当
10 所述计算机程序/指令被处理器执行时，致使所述处理器能够实现图 1 对应实施例所述的方法。

如图 8 为本申请实施例提供的一种数据库系统的示意图。从图 8 中可以看到，所述系统包括：

编码器 81，用于获取到的自然语言进行编码处理得到的语义特征向量；
15 解码器 82，用于所述编码器输出的所述语义特征向量进行解码处理，生成目标语句；

拆分模块 83，用于对获取到的所述目标语句进行拆分，得到至少一个目标子语句；

修正模块 84，用于根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中的关键词子句进行修正；
20

组合模块 85，用于对修正后的目标子语句进行组合，生成可被执行的机器语言；

数据库 86，用于基于所述可被执行的机器语言执行数据查询任务。

在实际应用中，编码器 81 获取基于用户提供的自然语言进行编码处理得到的语义特征向量；利用解码器 82 对编码器 81 提供的语义特征向量进行解码处理，得到不可执行的目标语句。进而，利用拆分模块 83 对目标语句进行拆分处理，得到至少一个目标子语句。将至少一个目标子语句发送给修正模块 84，根据所述目标子语句序列中包含的关键词类型对应的语义约束规则，对所述目标子语句进行修正。由组合模块 85 对修正后的目标子语句进行组合，生成所述自然语言对应的机器语言。数据库 86 在接收到可被执行的机器语言后，执行相应的数据查询任务，并将查询结果反馈给用户。当然，在实际应用中，编码器、解码器、拆分模块、修正模块和组合模块可以集成在一个工作节点，数据库作为另一个单独节点工作；当然也可以将所有模块和数据库集成在同本地服务器或者云端服务器中。
30

35 基于上述实施例，当用户提供查询作用的自然语言后，可以由编码器

对自然语言进行编码处理从而得到该自然语言的语义特征向量。利用解码器对语言特征向量进行进一步处理，得到自然语言对应的可被数据库识别的目标语句。对目标语句进行拆分处理，得到包含至少一个目标子语句的序列。在每个目标子语句中都包含有关键词，在对目标子语句进行修正的时候，基于目标子语句中包含的关键词类型和对应的语义约束规则，对目标子语句进行修正，并对修正后的目标子语句进行组合，得到能够被数据库识别的机器语言。通过上述方式，能够有效提高自然语言转义准确性以及用户使用数据库查询的工作效率，同时降低对使用数据库查询工作人员的技术能力要求。

10 由于每个关键词都有其特有的执行语义约束规则，这些规则是由机器语言特性所决定的，而且这些规则在自然语言中是不具的，无法通过自然语言转义直接得到的。此外，由于规则不具有很强的规律性，关键词种类也很多，是神经网络模型无法学习到的。因此，为了确保 SQL 语句能够被正确执行，具体针对每个关键词子句中的值进行逐一修正，使得修正效果更准确。

15 以上所描述的装置实施例仅仅是示意性的，其中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模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

20 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如 ROM/RAM、磁碟、光盘等，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）执行各个实施例或者实施例的某些部分所述的方法。

25 最后应说明的是：以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权 利 要 求 书

- 1、一种语言转义方法，所述方法包括：
- 对获取到的目标语句进行拆分，得到至少一个目标子语句；
- 确定所述目标子语句中包含的关键词类型和关键词子句；
- 5 根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中的关键词子句进行修正；
- 对修正后的目标子语句进行组合，生成可被执行的机器语言。
- 2、根据权利要求1所述的方法，所述对获取到的目标语句进行拆分，得到至少一个目标子语句，包括：
- 10 利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句；
- 根据嵌套关系对所述目标语句进行拆分，生成无嵌套关系的所述至少一个目标子语句。
- 3、根据权利要求2所述的方法，所述根据嵌套关系对所述目标语句进行拆分，生成无嵌套关系的所述至少一个目标子语句，包括：
- 15 根据所述目标语句中包含的嵌套标识，确定存在嵌套关系的第一目标子语句和第二目标子语句；其中，所述第二目标子语句嵌套在所述第一目标子语句中；
- 在所述第一目标子语句中所述嵌套标识所在的嵌套位置添加位置标识；
- 20 生成由所述第一目标子语句和所述第二目标子语句按照嵌套关系中的层级顺序组成的序列。
- 4、根据权利要求2所述的方法，得到至少一个目标子语句之后，还包括：
- 25 根据所述目标子语句中包含的关键词类型，确定对应的所述语义约束规则。
- 5、根据权利要求2所述的方法，当所述关键词类型为 SELECT 类型时，
- 所述利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句，包括：
- 30 获取数据库提供的数据表；
- 对所述语义特征向量进行解码得到的所述 SELECT 关键词关联的操作内容与所述数据表中数据列进行匹配度分析；
- 35 根据所述匹配度大小从所述数据表中确定与所述操作内容匹配的目

标数据列；

基于所述目标数据列中的数据，生成包含 SELECT 关键词、所述操作内容和所述预测值的目标语句。

5 6、根据权利要求 2 所述的方法，当所述关键词类型为 WHERE 类型时；

所述利用解码器对基于用户提供的自然语言编码得到的语义特征向量进行解码处理，得到包含有关键词、操作内容和预测值的目标语句，包括：

获取数据库提供的数据表；

10 根据所述语义特征信息进行解码得到的 WHERE 关键词关联的预测值，从所述数据表中筛选出操作内容；

根据所述预测值与所述筛选出操作内容的匹配度大小，确定与所述预测值匹配的所述筛选出操作内容，以得到包含 WHERE 关键词、所述预测值和所述筛选出操作内容的目标语句。

15 7、根据权利要求 4 所述的方法，所述根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中的关键词子句进行修正，包括：

当所述关键词类型为 HAVING 类型时，若执行所述操作内容时的执行结果为数值常量，则将 HAVING 子句中的所述预测值替换为常量。

20 8、根据权利要求 1 至 7 中任一项所述的方法，所述根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中各所述关键词子句进行修正，包括：

若所述预测值为非嵌套值，则计算所述预测值的第一特征向量与所述数据表中第一数据列中任意一个值的第二特征向量之间的第一相似度值；

25 若所述第一相似度值大于第一阈值，则用所述任意一个值替换所述预测值以得修正后的目标子语句。

9、根据权利要求 8 所述的方法，所述方法还包括：若所述相似度值小于第一阈值，则计算所述数据表中除所述第一数据列之外的任意一列中任意一个值的第三特征向量与所述第一特征向量之间的第二相似度值；

30 若所述第二相似度值大于第二阈值，则将所述任意一列中任意一个值替换所述预测值以得到修正后的目标子语句。

10、根据权利要求 5 或 6 所述的方法，所述根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中各所述关键词子句进行修正，包括：

35 当所述关键词类型为 ORDER BY 类型时，若所述目标子语句中除

ORDER BY 子句外的其他子句包含有聚合函数，则将所述 ORDER BY 子句转换为嵌套查询语句。

- 11、根据权利要求 5 或 6 所述的方法，所述根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中各所述关键词子句进行修正，包括：

当所述关键词类型为 GROUP BY 类型时，若 SELECT 子句不包含聚合函数和 GROUP BY 子句对应的数据列的列名，则将所述 GROUP BY 的列名替换为主键；

- 10 若未查找到所述主键，则将所述 SELECT 子句对应的数据列复制到所述 GROUP BY 对应的数据列。

- 12、根据权利要求 5 或 6 所述的方法，还包括：

当所述关键词类型为 FROM 类型时，基于所述修正后的目标子语句对应的数据表和其中的数据列生成有向图；

- 15 基于所述有向图中的所述数据表和所述数据列，以所述数据列或外键为根节点构造最小生成树。

- 13、根据权利要求 5 或 6 所述的方法，所述对修正后的目标子语句进行组合，包括：

若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句具有相同的列名或主外键关系，则对修改后的目标子语句进行组合；或者，

- 20 若修改后的所述目标子语句中的 SELECT 子句与 WHERE 子句不具有相同的列名或主外键关系，则将 WHERE 子句对应的列名复制到所述 SELECT 子句，对修改后的目标子语句进行组合。

- 25 14、一种电子设备，包括存储器及处理器；其中，所述存储器，用于存储程序；

所述处理器，与所述存储器耦合，用于执行所述存储器中存储的所述程序，以用于实现上述权利要求 1 至 13 中任一项所述的方法。

- 30 15、一种非暂时性机器可读存储介质，所述非暂时性机器可读存储介质上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行如权利要求 1 至 13 中任一项所述的方法。

- 16、一种数据库系统，所述系统包括：

编码器，用于获取到的自然语言进行编码处理得到的语义特征向量；

解码器，用于所述编码器输出的所述语义特征向量进行解码处理，生成目标语句；

- 35 拆分模块，用于对获取到的所述目标语句进行拆分，得到至少一个目

标子语句；

修正模块，用于根据所述目标子语句中包含的关键词类型对应的语义约束规则，对所述目标子语句中的关键词子句进行修正；

组合模块，用于对修正后的目标子语句进行组合，生成可被执行的机

5 器语言；

数据库，用于基于所述可被执行的机器语言执行数据查询任务。