(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105095193 A
(43)申请公布日 2015.11.25

(21)申请号 201410192917.6

(22)申请日 2014.05.08

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 马林 腾志扬 熊皓

(74)专利代理机构 深圳中一专利商标事务所 44237
代理人 张全文

(51)Int.Cl.
G06F 17/28(2006.01)

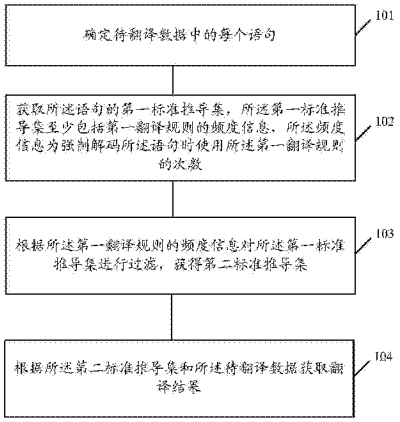
权利要求书3页 说明书12页 附图6页

(54)发明名称

一种机器翻译的方法及其设备

(57)摘要

本发明公开了一种机器翻译的方法及设备，通过确定待翻译数据中的每个语句；获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集，所述第二标准推导集至少包括所述第二翻译规则的频度信息；根据所述第二标准推导集和所述待翻译数据获取翻译结果，从而仅占用少量的计算资源实现语言翻译，替用户节约获取翻译的成本，提高移动终端的用户体验。



CN 105095193 A

1. 一种机器翻译的方法，其特征在于，所述方法包括：

确定待翻译数据中的每个语句；

获取所述语句的第一标准推导集，所述第一标准推导集至少包括第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；

根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；

根据所述第二标准推导集和所述待翻译数据获取翻译结果。

2. 根据权利要求 1 所述的方法，其特征在于，所述第二标准推导集包括第二翻译规则，所述获得第二标准推导集之后，还包括：

确定所述第二翻译规则对应的概率数值；

根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则；

所述根据所述第二标准推导集和所述待翻译数据获取翻译结果包括：

根据所述第三翻译规则和所述语句获得翻译结果。

3. 根据权利要求 2 所述的方法，其特征在于，所述第二翻译规则对应的概率数值包括正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率；

所述根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则包括：

将所述第二翻译规则对应的正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率分别和所述正向短语翻译概率对应的分数、反向短语翻译概率对应的分数、正向词汇翻译概率对应的分数乘和反向词汇翻译概率对应的分数相乘，将相乘获得的积累加为第一概率值；

根据所述第一概率值和预设的聚类方法获取第一码表，所述第一码表包括但不限于 65536 码表；

所述根据所述第三翻译规则和所述语句获得翻译结果，包括：

根据所述第一码表和所述语句获得翻译结果。

4. 根据权利要求 1 至 3 任一所述的方法，其特征在于，所述获取所述语句的第一标准推导集，包括：

根据统计的机器翻译的方法和所述语句获取规则表 Ta，所述规则表 Ta 包括各个翻译规则和所述翻译规则的频度信息；

根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码；

将强制解码所述语句时使用到的翻译规则组合为第一标准推导集。

5. 根据权利要求 4 所述的方法，其特征在于，所述根据翻译规则和所述翻译规则的频度信息对所述语句进行强制解码之后，还包括：

获取强制解码的结果；

当强制解码成功时，根据所述强制解码的结果获取 K 个推导树对应的规则信息，K 为正整数；

所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

将所述 K 个推导树对应的规则信息设置为第一标准推导集。

6. 根据权利要求 5 所述的方法，其特征在于，所述获取强制解码的结果之后，还包括：

当强制解码失败时，则获取强制解码失败时生成的推导树中的根节点到子节点的规则信息；

所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

将所述根节点到子节点的规则信息设置为第一标准推导集。

7. 一种机器翻译的装置，其特征在于，所述装置包括：

第一确定单元，用于确定待翻译数据中的每个语句；

第一获取单元，用于获取所述语句的第一标准推导集，所述第一标准推导集至少包括第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；

过滤单元，用于根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；

第二获取单元，用于根据所述第二标准推导集和所述待翻译数据获取翻译结果。

8. 根据权利要求 7 所述的装置，其特征在于，所述第二标准推导集包括第二翻译规则，所述装置还包括：

第二确定单元，用于确定所述第二翻译规则对应的概率数值；

第三获取单元，用于根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则；

所述第二获取单元，具体用于：

根据所述第三翻译规则和所述语句获得翻译结果。

9. 根据权利要求 8 所述的装置，其特征在于，所述第二翻译规则对应的概率数值包括正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率；

所述第三获取单元具体用于：

将所述第二翻译规则对应的正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率分别和所述正向短语翻译概率对应的分数、反向短语翻译概率对应的分数、正向词汇翻译概率对应的分数乘和反向词汇翻译概率对应的分数相乘，将相乘获得的积累加为第一概率值；

根据所述第一概率值和预设的聚类方法获取第一码表，所述第一码表包括但不限于 65536 码表；

所述第二获取单元具体用于：

根据所述第一码表和所述语句获得翻译结果。

10. 根据权利要求 7 至 9 任一所述的装置，其特征在于，所述第一获取单元，具体用于：

根据统计的机器翻译的方法和所述语句获取规则表 Ta，所述规则表 Ta 包括各个翻译规则和所述翻译规则的频度信息；

根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码；

将强制解码所述语句时使用到的翻译规则组合为第一标准推导集。

11. 根据权利要求 10 所述的装置，其特征在于，所述装置还包括第四获取单元，

所述第四获取单元，用于获取强制解码的结果，当强制解码成功时，根据所述强制解码的结果获取 K 个推导树对应的规则信息，K 为正整数；

所述第一获取单元还用于将所述 K 个推导树对应的规则信息设置为第一标准推导集。

12. 根据权利要求 11 所述的装置，其特征在于，所述第四获取单元还用于：

当强制解码失败时，则获取强制解码失败时生成的推导树中的根节点到子节点的规则信息；

所述第一获取单元还用于将所述根节点到子节点的规则信息设置为第一标准推导集。

一种机器翻译的方法及其设备

技术领域

[0001] 本发明涉及机器翻译领域，尤其涉及到一种机器翻译的方法及设备。

背景技术

[0002] 随着经济全球化的发展，不同国家之间的信息交流和交换变得越来越频繁。同时，蓬勃发展的互联网为获取诸如英语、汉语、法语、德语、日语等各种语言形式的信息提供了极大的便利。公众对于不同语言之间的翻译需求也变得日益强烈。人工翻译耗时较长，成本较高，已经满足不了人们对多语言信息日益增长的需求。机器翻译能够将一种自然语言自动地翻译为另一种自然语言。利用机器翻译快速获取多语言的信息和资源已成为必然趋势。这使得能提供多语言、高质量、易获取的翻译服务的机器翻译系统和设备也变得越来越重要。近年来在一些国际组织机构（如欧洲联盟）、新闻媒体、全球性的网络平台、跨国贸易与旅游等政治、网络、文化、教育以及商务环境中，机器翻译已逐渐成为了一种获取信息和传播信息的重要基础手段。

[0003] 统计机器翻译是目前主流的机器翻译技术。它能够根据数学模型和算法自动地从平行语料库中学习到翻译知识。统计机器翻译并不需要相关的语言学家参与，并且与具体的语言相对独立。另外，统计机器翻译系统开发部署周期较短，翻译速度较快，翻译质量较为鲁棒。

[0004] 统计机器翻译模型学习到的翻译知识通常用规则表表示。规则表的质量对翻译质量起到了关键性的作用。根据规则的不同，统计机器翻译模型通常可以分为短语模型、句法模型、语义模型等。在实际应用中，短语模型和层次短语模型较为成熟，被广泛使用。一个可实用的统计机器翻译模型通常需要在千万级别的双语句对语料上训练。这使得经由自动学习算法得到的规则表十分庞大。一方面，庞大的规则表会占用较多的资源，如存储时需要较大硬盘空间，加载时需要较多的内存；另一方面过大的规则表也会增加翻译解码过程搜索空间，降低翻译速度。

[0005] 现有技术采用基于强制解码的方法对规则表进行过滤。强制解码技术指的是将训练预料中一个句对的源语言端 f 作为输入提供给解码器，用对应的目标语言端 e 硬性约束翻译解码过程的推导 d，其中要求 d(f) = e。d(f) 表示 f 的一个翻译推导 d 对应的目标语言输出，满足这样条件的推导 d，我们称为标准推导。强制解码采用期望最大化算法在压缩标准推导森林(standard derivative forest compression)中估计每个规则的概率，根据设置的阈值过滤掉概率较低的规则，从而减少搜索空间。

[0006] 从上可知，由于规则表中规则数量庞大，强制解码采用期望最大化算法在压缩标准推导森林中估计每个规则的概率会使得计算代价较大，并且降低用户体验。

发明内容

[0007] 本发明实施例提供了一种机器翻译的方法及设备，旨在解决如何过滤规则使得使用过滤后的规则可以提高机器翻译的质量。

[0008] 第一方面，一种机器翻译的方法，所述方法包括：

[0009] 确定待翻译数据中的每个语句；

[0010] 获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；

[0011] 根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；

[0012] 根据所述第二标准推导集和所述待翻译数据获取翻译结果。

[0013] 结合第一方面，在第一方面的第一种可能的实现方式中，所述第二标准推导集包括第二翻译规则，所述获得第二标准推导集之后，还包括：

[0014] 确定所述第二翻译规则对应的概率数值；

[0015] 根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则；

[0016] 所述根据所述第二标准推导集和所述待翻译数据获取翻译结果包括：

[0017] 根据所述第三翻译规则和所述语句获得翻译结果。

[0018] 结合第一方面的第一种可能的实现方式，在第一方面的第二种可能的实现方式中，所述第二翻译规则对应的概率数值包括正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率；

[0019] 所述根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则包括：

[0020] 将所述第二翻译规则对应的正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率分别和所述正向短语翻译概率对应的分数、反向短语翻译概率对应的分数、正向词汇翻译概率对应的分数乘和反向词汇翻译概率对应的分数相乘，将相乘获得的积累加为第一概率值；

[0021] 根据所述第一概率值和预设的聚类方法获取第一码表，所述第一码表包括但不限于 65536 码表；所述方法还包括：

[0022] 所述根据所述第三翻译规则和所述语句获得翻译结果，包括：

[0023] 根据所述第一码表和所述语句获得翻译结果。

[0024] 结合第一方面或者第一方面的第一种可能的实现方式或第一方面的第二种可能的实现方式，在第一方面的第三种可能的实现方式中，所述获取所述语句的第一标准推导集，包括：

[0025] 根据统计的机器翻译的方法和所述语句获取规则表 Ta，所述规则表 Ta 包括各个翻译规则和所述翻译规则的频度信息；

[0026] 根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码；

[0027] 将强制解码所述语句时使用到的翻译规则组合为第一标准推导集。

[0028] 结合第一方面的第三种可能的实现方式，在第一方面的第四种可能的实现方式中，所述根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码之后，还包括：

[0029] 获取强制解码的结果；

[0030] 当强制解码成功时，根据所述强制解码的结果获取 K 个推导树对应的规则信息，K

为正整数；

[0031] 所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

[0032] 将所述 K 个推导树对应的规则信息设置为第一标准推导集。

[0033] 结合第一方面的第四可能的实现方式，在第一方面的第五种可能的实现方式中，所述获取强制解码的结果后，还包括：

[0034] 当强制解码失败时，则获取强制解码失败时生成的推导树中的根节点到子节点的规则信息；

[0035] 所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

[0036] 将所述根节点到子节点的规则信息设置为第一标准推导集。

[0037] 第二方面，一种机器翻译的装置，所述装置包括：

[0038] 第一确定单元，用于确定待翻译数据中的每个语句；

[0039] 第一获取单元，用于获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；

[0040] 过滤单元，用于根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；

[0041] 第二获取单元，用于根据所述第二标准推导集和所述待翻译数据获取翻译结果。

[0042] 结合第二方面，在第二方面的第一种可能的实现方式中，所述第二标准推导集包括第二翻译规则，所述装置还包括：

[0043] 第二确定单元，用于确定所述第二翻译规则对应的概率数值；

[0044] 第三获取单元，用于根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则；

[0045] 所述第二获取单元，具体用于：

[0046] 根据所述第三翻译规则和所述语句获得翻译结果。

[0047] 结合第二方面的第一种可能的实现方式，在第二方面的第二种可能的实现方式中，所述第二翻译规则对应的概率数值包括正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率；

[0048] 所述第三获取单元具体用于：

[0049] 将所述第二翻译规则对应的正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率分别和所述正向短语翻译概率对应的分数、反向短语翻译概率对应的分数、正向词汇翻译概率对应的分数乘和反向词汇翻译概率对应的分数相乘，将相乘获得的积累加为第一概率值；

[0050] 根据所述第一概率值和预设的聚类方法获取第一码表，所述第一码表包括但不限于 65536 码表；

[0051] 所述第二获取单元具体用于：

[0052] 根据所述第一码表和所述语句获得翻译结果。

[0053] 结合第二方面或者第二方面的第一种可能的实现方式或第二方面的第二种可能的实现方式，在第二方面的第三种可能的实现方式中，所述第一获取单元，具体用于：

[0054] 根据统计的机器翻译的方法和所述语句获取规则表 Ta，所述规则表 Ta 包括各个

翻译规则和所述翻译规则的频度信息；

[0055] 根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码；

[0056] 将强制解码所述语句时使用到的翻译规则组合为第一标准推导集。

[0057] 结合第二方面的第三种可能的实现方式，在第二方面的第四种可能的实现方式中，所述装置还包括第四获取单元，

[0058] 所述第四获取单元，用于获取强制解码的结果；当强制解码成功时，根据所述强制解码的结果获取 K 个推导树对应的规则信息，K 为正整数；

[0059] 所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

[0060] 将所述 K 个推导树对应的规则信息设置为第一标准推导集。

[0061] 结合第二方面的第四种可能的实现方式，在第二方面的第五种可能的实现方式中，所述第四获取单元还用于：

[0062] 当强制解码失败时，则获取强制解码失败时生成的推导树中的根节点到子节点的规则信息；

[0063] 所述第一获取单元还用于将所述根节点到子节点的规则信息设置为第一标准推导集。

[0064] 本发明通过确定待翻译数据中的每个语句；获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；根据所述第二标准推导集和所述待翻译数据获取翻译结果，从而仅占用少量的计算资源实现语言翻译，替用户节约获取翻译的成本，提高移动终端的用户体验。

附图说明

[0065] 为了更清楚地说明本发明实施例的技术方案，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，下面描述中的附图仅仅是本发明的一些实施例，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。

[0066] 图 1 是本发明实施例提供的一种机器翻译的方法流程图；

[0067] 图 2 至图 4 是本发明实施例提供的一种推导树的结构图；

[0068] 图 5 是本发明实施例提供的一种基于强制解码的低频规则过滤的方法示意图；

[0069] 图 6 是本发明实施例提供的一种规则压缩的方法示意图；

[0070] 图 7 是本发明实施例提供的一种机器翻译的装置结构图；

[0071] 图 8 是本发明实施例提供的一种机器翻译的装置结构图；

[0072] 图 9 是本发明实施例提供的一种机器翻译的装置结构图；

[0073] 图 10 是本发明实施例提供的一种机器翻译的装置结构图。

具体实施方式

[0074] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于

本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0075] 参考图 1，图 1 是本发明实施例提供的一种机器翻译的方法流程图。如图 1 所示，所述方法包括以下步骤：

[0076] 步骤 101，确定待翻译数据中的每个语句；

[0077] 步骤 102，获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；

[0078] 可选地，所述获取所述语句的第一标准推导集，包括：

[0079] 根据统计的机器翻译的方法和所述语句获取规则表 Ta，所述规则表 Ta 包括各个翻译规则和所述翻译规则的频度信息；

[0080] 根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码；

[0081] 将强制解码所述语句时使用到的翻译规则组合为第一标准推导集。

[0082] 所述根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码之后，还包括：

[0083] 获取强制解码的结果；

[0084] 当强制解码成功时，根据所述强制解码的结果获取 K 个推导树对应的规则信息，K 为正整数；

[0085] 所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

[0086] 将所述 K 个推导树对应的规则信息设置为第一标准推导集。

[0087] 所述获取强制解码的结果之后，还包括：

[0088] 当强制解码失败时，则获取强制解码失败时生成的推导树中的根节点到子节点的规则信息；

[0089] 将所述根节点到子节点的规则信息设置为第一标准推导集。

[0090] 其中，所述统计的机器翻译的方法 (statistical machine translation) 是现有技术中常用的一种翻译方法。

[0091] 具体的，假定第一标准推导集中的规则如表 1 所示，并且给定训练预料中的一个对齐的双语句对的源语言 f 为“电脑和手机是上个世纪的发明”，目标语言 e 为“Computers and cell phones are the invention of the last century”。所述规则的源语言和目标语言如下表 1 所示：

[0092]

规则序号	规则的源语言和目标语言
R1	<手机, cell phones>
R2	<电脑 和 X_1, computers and X_1>
R3	<电脑 和 手机, computers and cell phones >
R4	<X_1 是 X_2, X_1 is X_2>

[0093]

R5	<上个 世纪, the last century>
R6	<X_1 的 发明, inventions of X_1>
R7	<手机, mobile phones>
R8	<上个世纪, last century>
R9	<电, electricity>
R10	<脑, brain>

[0094] 表 1

[0095] 对于一条规则“<X1 是 X2, X1is X2>”,其中“X1 是 X2”称为规则的源语言,“X1is X2”称为规则的目标语言,X 表示泛化的变量,下标表示变量替换时的对应关系。

[0096] 参考图 2- 图 4 的推导树,结点 S 表示一颗推导树的起点,X 表示推导所用的规则,为了统计方便,所有用到的规则,均以用规则的编号作好了标注。各个规则在前 3 个标准推导中的频度统计信息如下表 2 所示,根据表 2 可以从 Ta 生成 Tb。

[0097]

规则	在 3-best 标准推导集上规则的频度
R1	2
R2	2
R3	1
R4	3
R5	3

R6	3
R7	0
R8	0
R9	0
R10	0

[0098] 表 2

[0099] 根据规则表 Ta 对所述训练语料的每个句对进行强制解码生成 K-best 标准推导集，其中，K 最小为 1，最大可以为无穷大。

[0100] 具体的，对于每一个句子，我们可以参考第一标准推导集，得到相应的翻译结果。在产生翻译结果的过程中，有很多不同的规则组合，可以产生相同的翻译结果。根据概率的不同，可以将这些规则组合排序。所谓的 k-best 推导集就是选取前 k 个最好的翻译规则组合。

[0101] 可选地，在上述实施案例的基础上，通过融入标准推导集上的规则概率特征，进行重新训练。以源语言短语 f 到目标语言 e 为例，翻译概率 P(e|f) 的最大似然估计为：

[0102] $$P(e|f)=\frac{count(e,f)}{\sum_{ef} count(e',f)}$$

[0103] 这种评估方式考虑了语料库上的 e 和 f 的所有互译次数，并且考虑语料库中所有跟 f 互译的短语 e'。本实例中在原始的计算上加入两个新的概率特征，这些概率特征只在强制解码所得到的标准推导集中统计，标准推导集中的翻译概率 $P_{gd}(e|f)$ 的计算公式为：

[0104] $$P_{gd}(e|f)=\frac{count_{gd}(e,f)}{\sum_{ef} count_{gd}(e',f)}$$

[0105] 从目标语言 e 到 f 的翻译概率 Pgd(f|e) 的计算方式与之类似。

[0106] 步骤 103，根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；

[0107] 具体的，参考表 2，例如将过滤的频度的阈值设置为 1（包括频度 1），剩下的规则为 R1、R2、R4、R5、R6，过滤掉的规则为 R3、R7、R8、R9、R10。R8 是一条错误的规则，没有被使用上，所以被过滤。R9 和 R10 虽然是正确的规则，但是不符合本句话翻译的语义，也被过滤掉了。R7 虽然是正确且语义相符合的规则，但是由于不符合强制解码的条件，所以被过滤掉了。R3 虽然正确、符合语义也满足强制解码的条件，但是由于出现的频度较低，所以也被过滤了。R3 和 R7 被过滤，并不会影响翻译的质量，源语言 f 仍然能够正确地被翻译到 e，翻译的推导仍然能够正常进行。

[0108] 具体的，参考图 5，图 5 是本发明实施例提供的一种基于强制解码的低频规则过滤的方法示意图。对于给定的初始规则表 Ta、翻译模型参数和训练语料，首先判断一个规则是不是能出现在强制推导的标准集中，如果不满足这个条件，那么该规则对应的频度就设为 0，如表 2 的例子所示，有些频度为 0 的规则就是不满足条件的规则。对于满足条件的所有规

则，通过翻译模型的解码算法，不考虑语言模型的得分，生成前 k 个较好的标准推导集。如果由于剪枝等原因，造成了强制解码失败，那么此时须采用回退策略，回退倒能够覆盖源语言词数最多的部分标准推导。强制解码完后，对规则的频度进行统计，依次判断每一条规则是否在标准推导集中常见。对于不常见的规则，那么就过滤掉。而剩下的规则则被保留下来。

[0109] 可选地，所述第二标准推导集包括第二翻译规则，所述获得第二标准推导集之后，还包括：

[0110] 确定所述第二翻译规则对应的概率数值；

[0111] 根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则；

[0112] 所述根据所述第二标准推导集和所述待翻译数据获取翻译结果包括：

[0113] 根据所述第三翻译规则和所述语句获得翻译结果。

[0114] 可选地，所述根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，包括：

[0115] 所述第二翻译规则对应的概率数值包括正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率；

[0116] 将所述第二翻译规则对应的正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率分别和所述正向短语翻译概率对应的分数、反向短语翻译概率对应的分数、正向词汇翻译概率对应的分数乘和反向词汇翻译概率对应的分数相乘，将相乘获得的积累加为第一概率值；

[0117] 根据所述第一概率值和预设的聚类方法获取第一码表，所述第一码表包括但不限于 65536 码表；

[0118] 所述根据所述第三翻译规则和所述语句获得翻译结果，包括：

[0119] 根据所述第一码表和所述语句获得翻译结果。

[0120] 其中，所述压缩是指对词语进行数字化的表示，使得占用的空间更小。

[0121] 例如，假设对于规则 R1<手机，cell phones>其翻译模型的分值分别为 0、-10.824、-10.2205、-0.847298，翻译模型的权重为 0.186212、0.0568202、0.144704、0.0193515，那么预合并的结果为 score(R1) = 0*0.186212-10.824*0.0568202-10.2205*0.144704-0.847298*0.193515 = -2.58。得到预合并的分值之后，原始的 4 个翻译模型的分值就不用再存储了，在评估规则的翻译模型分数时，仅使用预合并后的结果即可。

[0122] 首先将所有的分数都放在一起，做一次聚类。聚类的数目为 2 的 16 次方 (2 个字节能放得下)，也就是聚成 65536 个类，每个类有一个中心值，将这 65536 个值做成一个码表。每个概率存储的时候只存储这个码表的索引，然后从码表中取值。

[0123] 例如，假设有四条规则 R1、R2、R3 和 R4，其对应的得分为 0.1、0.2、0.7 和 0.8，聚类的数目为 2。那么很容易可以得到两个聚类的中心点 C1 = 0.15，C2 = 0.75。因此 R1 和 R2 对应的类别索引为 1，分值为 0.15，R3 和 R4 对应的类别索引为 2，分值为 0.75。

[0124] 预设的聚类方法可采用 k-means。由于规则表中的分数值数目量十分庞大，直接的 k-means 方法在数亿级别的数据中聚类较慢。由于数据是一维的，因此在 k-means 之前对数据先做了一遍排序。经过预排序后的数据，k-means 聚类速度较快。

[0125] 具体的，参考图 6，图 6 是本发明实施例提供的一种规则压缩的方法示意图。如图 6 所示，首先利用源语言和目标语言的压缩方法对初始的规则表进行词级的压缩。然后对于规则表中剩余的分数项，首先根据解码过程中的需要判断该分数项是否可以提前跟其他的分数合并，如果可以合并，则将这些分数合并为一个分数。如果不可以合并，则先将该数值项对应的所有分数排序，并进行 k-means 聚类，聚成 65535 个类别。根据聚类的信息，将每个数值项对应的类别索引和中心点的分数值记录成码表。生成完码表后，即可利用该码表将对应的数值项转换成对应类别的索引。到此结束压缩的过程。

[0126] 在解码阶段，对于压缩后的规则表，解码器首先获得的是规则的分数项所在的类别的索引，需要根据该索引获取对应类别的中心点的值，这一点与普通的解码器有所不同。

[0127] 步骤 104，根据所述第二标准推导集和所述待翻译数据获取翻译结果。

[0128] 本发明结合规则过滤中强制解码方法和频度过滤方法的优点，具体说来指的是根据强制推导所生成的标准推导集上的规则频度进行过滤。其基本原理是假设规则的频度分布在整个训练集的标准推导集上符合长尾定律。少量的规则在标准推导中被反复使用，而且大量的规则在标准推导中被少量使用。标准推导代表了最准确的翻译过程，在标准推导中很难用到的规则，在翻译搜索的空间中也应该很难搜索到。因此将这类翻译模型很难搜索到的规则去掉，既能减少规则表的大小，又对翻译模型的质量影响不大。

[0129] 同时针对强制解码失败的句对，本发明并不会直接忽略，而是提出一种回退策略进行处理。当强制解码失败的时候，我们将保留其中成功强制解码的最大跨度片段所对应的部分标准推导 (Partial Gold Derivation)。

[0130] 本发明通过确定待翻译数据中的每个语句；获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集，所述第二标准推导集至少包括所述第二翻译规则的频度信息；根据所述第二标准推导集和所述待翻译数据获取翻译结果，从而仅占用少量的计算资源实现语言翻译，节约计算成本；能轻便地与电话系统、会议系统、手机操作系统、嵌入式操作系统等系统深度集成，为各个系统组件提供智能翻译服务，实现系统级别的“所见即所译”。

[0131] 参考图 7，图 7 是本发明实施例提供的一种机器翻译的装置结构图。如图 7 所示，所述装置包括：

[0132] 第一确定单元 701，用于确定待翻译数据中的每个语句；

[0133] 第一获取单元 702，用于获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；

[0134] 可选地，所述第一获取单元 702，具体用于：

[0135] 根据统计的机器翻译的方法和所述语句获取规则表 Ta，所述规则表 Ta 包括各个翻译规则和所述翻译规则的频度信息；

[0136] 根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码；

[0137] 将强制解码所述语句时使用到的翻译规则组合为第一标准推导集。

[0138] 可选地，所述装置还包括：

[0139] 第四获取单元 801，用于获取强制解码的结果；当强制解码成功时，根据所述强制解码的结果获取 K 个推导树对应的规则信息，K 为正整数；

[0140] 所述第一获取单元 702 还用于将所述 K 个推导树对应的规则信息设置为第一标准推导集。

[0141] 可选地，第四获取单元 801 还用于：

[0142] 当强制解码失败时，则获取强制解码失败时生成的推导树中的根节点到子节点的规则信息；

[0143] 所述第一获取单元 702 还用于：

[0144] 将所述根节点到子节点的规则信息设置为第一标准推导集。

[0145] 具体的，参考表 1 和表 2、图 2- 图 4 的描述，在此不再赘述。

[0146] 过滤单元 703，用于根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集，所述第二标准推导集至少包括所述第二翻译规则的频度信息；

[0147] 具体的，参考图 5 的描述，在此不再赘述。

[0148] 可选地，所述装置还包括：

[0149] 第二确定单元 901，用于确定所述第二翻译规则对应的概率数值；

[0150] 第三获取单元 902，用于根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则；

[0151] 所述第二获取单元 704，用于：

[0152] 根据所述第三翻译规则和所述语句获得翻译结果。

[0153] 可选地，所述第二翻译规则对应的概率数值包括正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率；

[0154] 所述第三获取单元 902 具体用于：

[0155] 将所述第二翻译规则对应的正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率分别和所述正向短语翻译概率对应的分数、反向短语翻译概率对应的分数、正向词汇翻译概率对应的分数乘和反向词汇翻译概率对应的分数相乘，将相乘获得的积累加为第一概率值；

[0156] 根据所述第一概率值和聚类的方法获取第一码表，所述第一码表包括但不限于 65536 码表；

[0157] 所述第二获取单元 704 具体用于：

[0158] 根据所述第一码表和所述语句获得翻译结果。

[0159] 第二获取单元 704，用于根据所述第二标准推导集和所述待翻译数据获取翻译结果。

[0160] 具体的，参考图 6 的描述，在此不再赘述。

[0161] 本发明通过确定待翻译数据中的每个语句；获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集，所述第二标准推导集至少包括所述第二翻译规则的频度信息；根据所述第二标准推导集和所述待翻译数据获取翻译结果，从而仅占用少量

的计算资源实现语言翻译，节约计算成本；能轻便地与电话系统、会议系统、手机操作系统、嵌入式操作系统等系统深度集成，为各个系统组件提供智能翻译服务，实现系统级别的“所见即所译”。

[0162] 图 10 是本发明实施例提供的一种机器翻译的装置结构图。参考图 10，图 10 是本发明实施例提供的一种机器翻译的装置 1000，本发明具体实施例并不对所述机器翻译的装置的具体实现做限定。所述机器翻译的装置 1000 包括：

[0163] 处理器 (processor)1001，通信接口 (Communications Interface)1002，存储器 (memory)1003，总线 1004。

[0164] 处理器 1001，通信接口 1002，存储器 1003 通过总线 1004 完成相互间的通信。

[0165] 通信接口 1002，用于与其他设备进行通信；

[0166] 处理器 1001，用于执行程序。

[0167] 具体地，程序可以包括程序代码，所述程序代码包括计算机操作指令。

[0168] 处理器 1001 可能是一个中央处理器 (central processing unit，CPU)，或者是特定集成电路 ASIC(Application Specific Integrated Circuit)，或者是被配置成实施本发明实施例的一个或多个集成电路。

[0169] 存储器 1003，用于存储程序。存储器 1003 可以是易失性存储器 (volatile memory)，例如随机存取存储器 (random-access memory，RAM)，或者非易失性存储器 (non-volatile memory)，例如只读存储器 (read-only memory，ROM)，快闪存储器 (flash memory)，硬盘 (hard disk drive，HDD) 或固态硬盘 (solid-state drive，SSD)。处理器 1001 根据存储器 1003 存储的程序指令，执行以下方法：

[0170] 确定待翻译数据中的每个语句；

[0171] 获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；

[0172] 根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；

[0173] 根据所述第二标准推导集和所述待翻译数据获取翻译结果。

[0174] 所述获得第二标准推导集之后，还包括：

[0175] 确定所述第二翻译规则对应的概率数值；

[0176] 根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，获取第三翻译规则；

[0177] 所述根据所述第二标准推导集和所述待翻译数据获取翻译结果包括：

[0178] 根据所述第三翻译规则和所述语句获得翻译结果。

[0179] 所述获取所述语句的第一标准推导集，包括：

[0180] 根据统计的机器翻译的方法和所述语句获取规则表 Ta，所述规则表 Ta 包括各个翻译规则和所述翻译规则的频度信息；

[0181] 根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码；

[0182] 将强制解码所述语句时使用到的翻译规则组合为第一标准推导集。

[0183] 所述根据所述翻译规则和所述翻译规则的频度信息对所述语句进行强制解码之后，包括：

[0184] 获取强制解码的结果；

[0185] 当强制解码成功时，根据所述强制解码的结果获取 K 个推导树对应的规则信息，K 为正整数；

[0186] 所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

[0187] 将所述 K 个推导树对应的规则信息设置为第一标准推导集。

[0188] 所述获取强制解码的结果之后，还包括：

[0189] 当强制解码失败时，则获取强制解码失败时生成的推导树中的根节点到子节点的规则信息；

[0190] 所述将强制解码所述语句时使用到的翻译规则组合为第一标准推导集包括：

[0191] 将所述根节点到子节点的规则信息设置为第一标准推导集。

[0192] 所述第二翻译规则对应的概率数值包括正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率；

[0193] 所述根据所述第二翻译规则对应的概率数值对所述第二翻译规则进行压缩，包括：

[0194] 将所述第二翻译规则对应的正向短语翻译概率、反向短语翻译概率、正向词汇翻译概率和反向词汇翻译概率分别和所述正向短语翻译概率对应的分数、反向短语翻译概率对应的分数、正向词汇翻译概率对应的分数乘和反向词汇翻译概率对应的分数相乘，将相乘获得的积累加为第一概率值；

[0195] 根据所述第一概率值和预设的聚类方法获取第一码表，所述第一码表包括但不限于 65536 码表；

[0196] 所述根据所述第三翻译规则和所述语句获得翻译结果，包括：

[0197] 根据所述第一码表和所述语句获得翻译结果。

[0198] 本发明通过确定待翻译数据中的每个语句；获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数；根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集；根据所述第二标准推导集和所述待翻译数据获取翻译结果，从而仅占用少量的计算资源实现语言翻译，替用户节约获取翻译的成本，提高移动终端的用户体验。

[0199] 以上所述，仅为本发明较佳的具体实施方式，但本发明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发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求的保护范围为准。

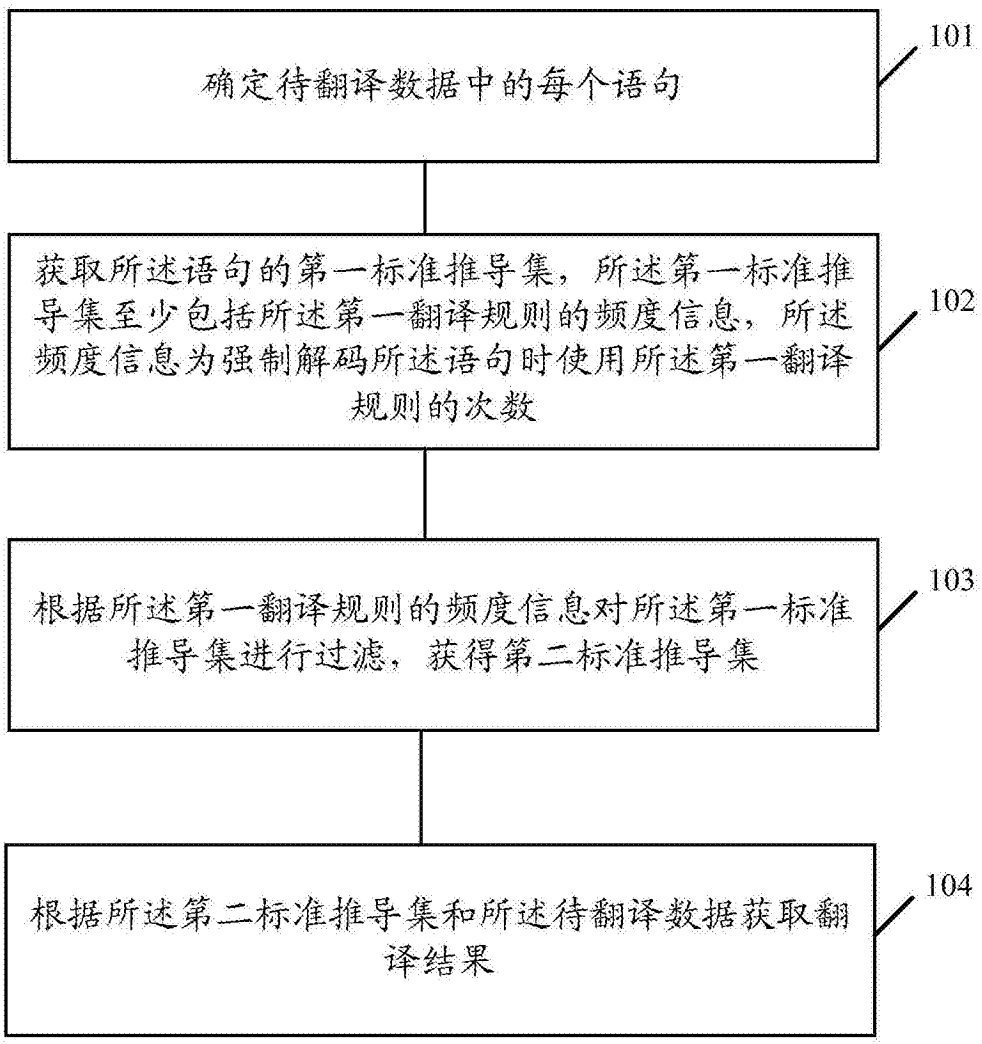

确定待翻译数据中的每个语句
101
获取所述语句的第一标准推导集，所述第一标准推导集至少包括所述第一翻译规则的频度信息，所述频度信息为强制解码所述语句时使用所述第一翻译规则的次数
102
根据所述第一翻译规则的频度信息对所述第一标准推导集进行过滤，获得第二标准推导集
103
根据所述第二标准推导集和所述待翻译数据获取翻译结果
104

图 1

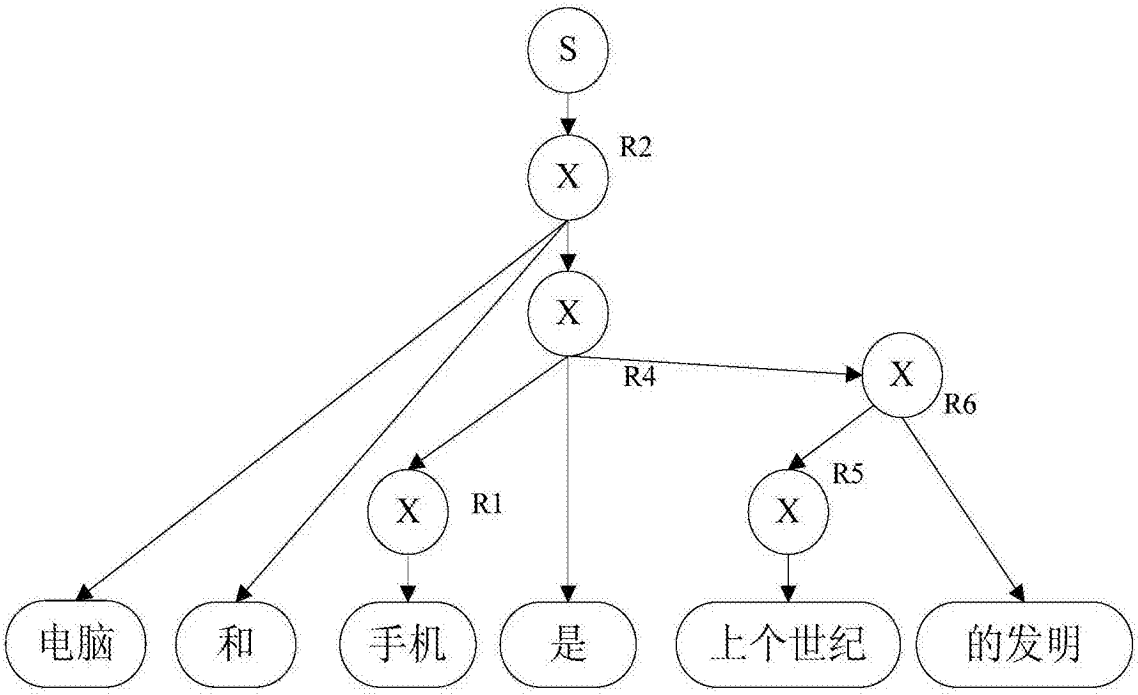

S
X
R2
X
X
R4
X
R6
X
R1
X
R5
电脑
和
手机
是
上个世纪
的发明

图 2

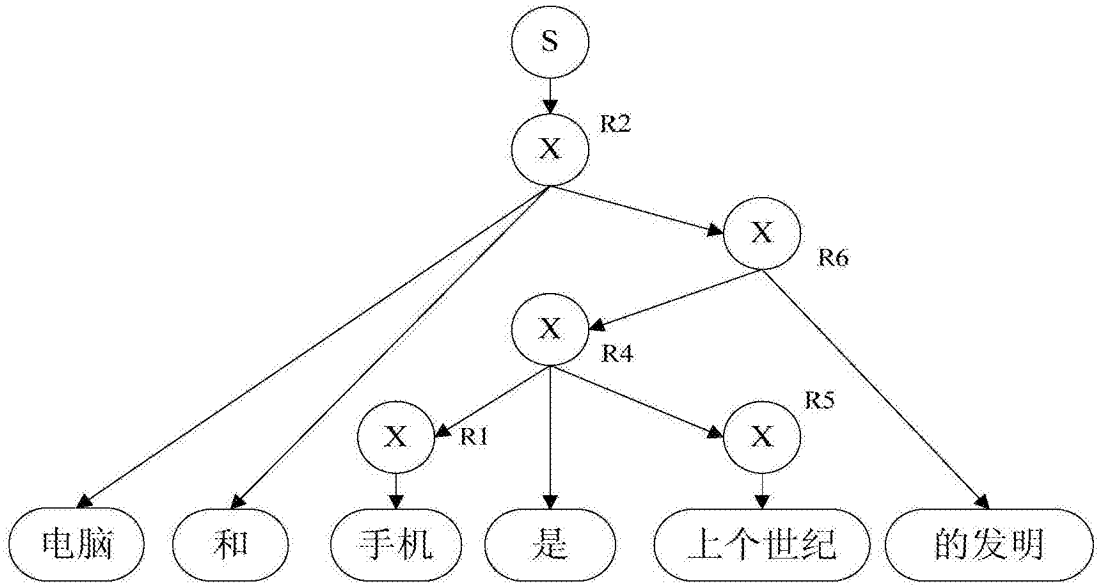

S
X
R2
X
R6
X
R4
X
R1
X
R5
电脑
和
手机
是
上个世纪
的发明

图 3

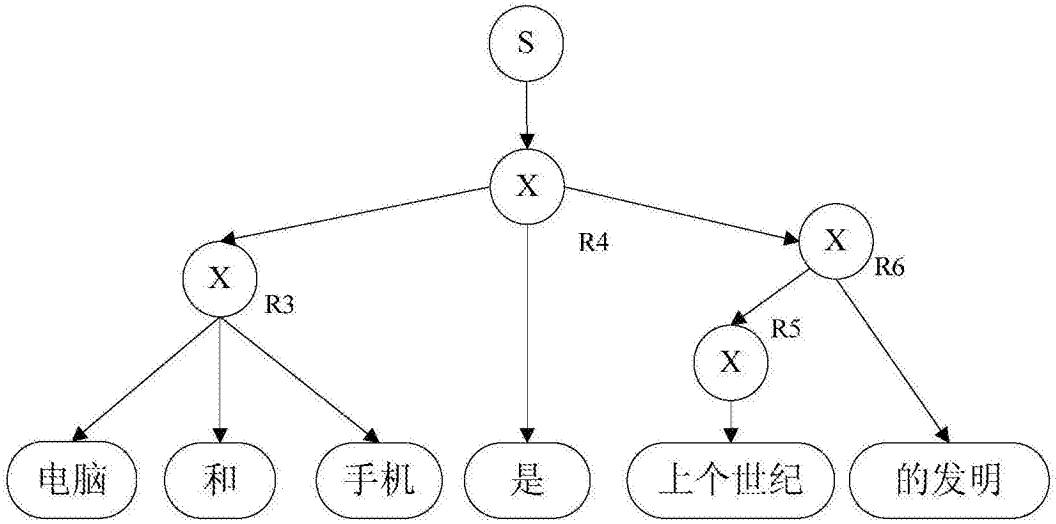


图 4

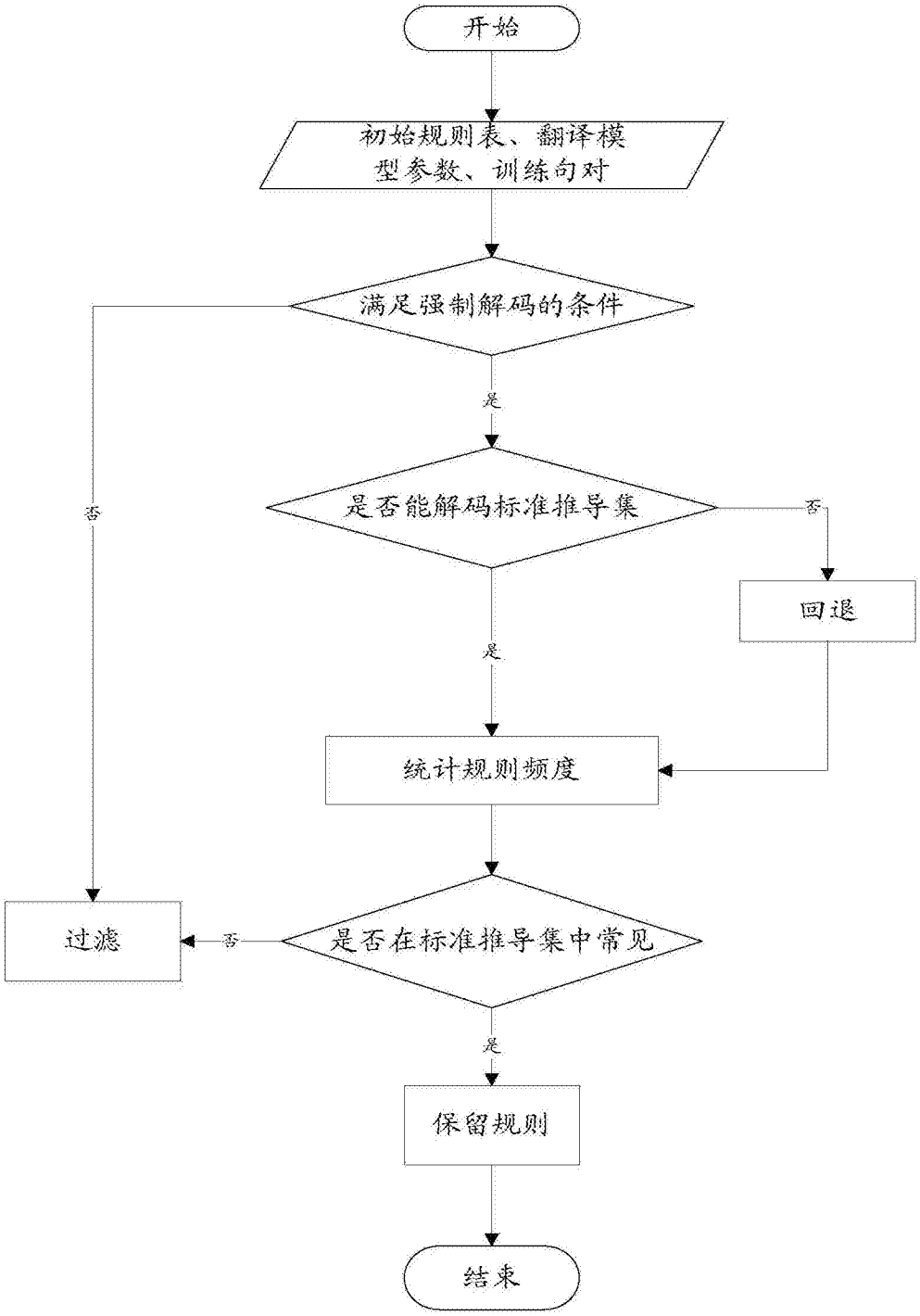

开始
初始规则表、翻译模型参数、训练句对
满足强制解码的条件
是
否
是否能解码标准推导集
否
回退
是
统计规则频度
过滤
否
是否在标准推导集中常见
是
保留规则
结束

图 5

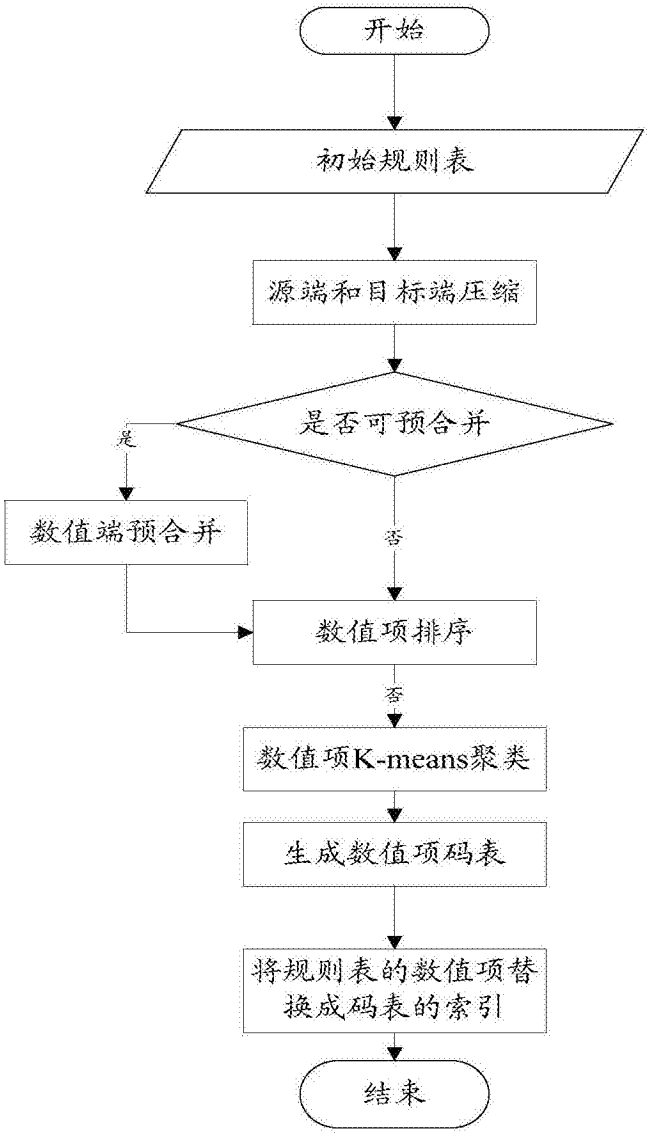

开始
初始规则表
源端和目标端压缩
是否可预合并
是
否
数值端预合并
数值项排序
否
数值项K-means聚类
生成数值项码表
将规则表的数值项替换成码表的索引
结束

图 6

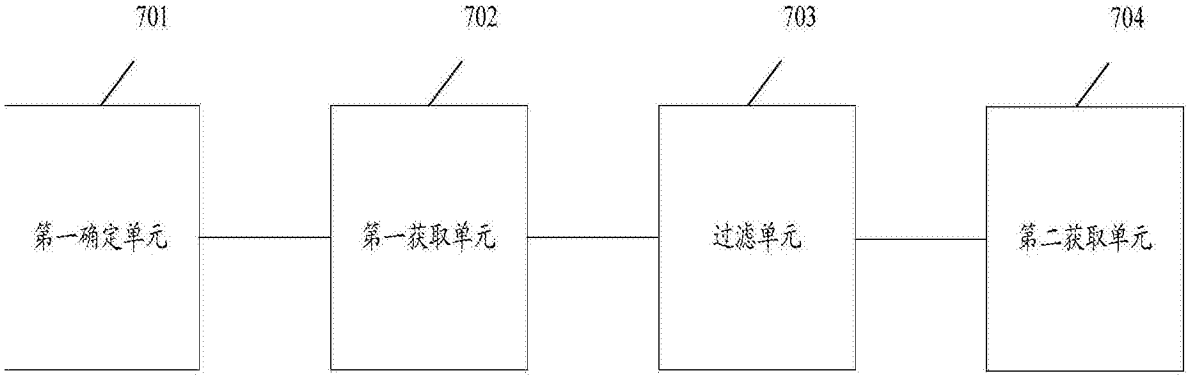

701
702
703
704
第一确定单元
第一获取单元
过滤单元
第二获取单元

图 7

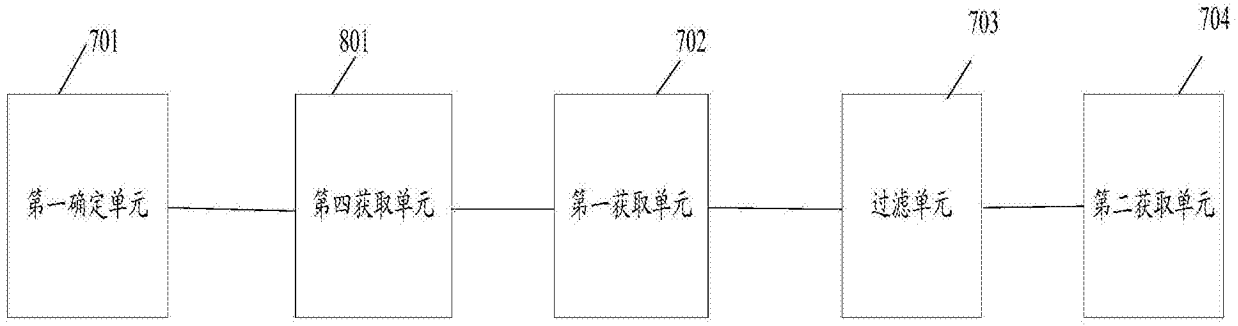

701
801
702
703
704
第一确定单元
第四获取单元
第一获取单元
过滤单元
第二获取单元

图 8

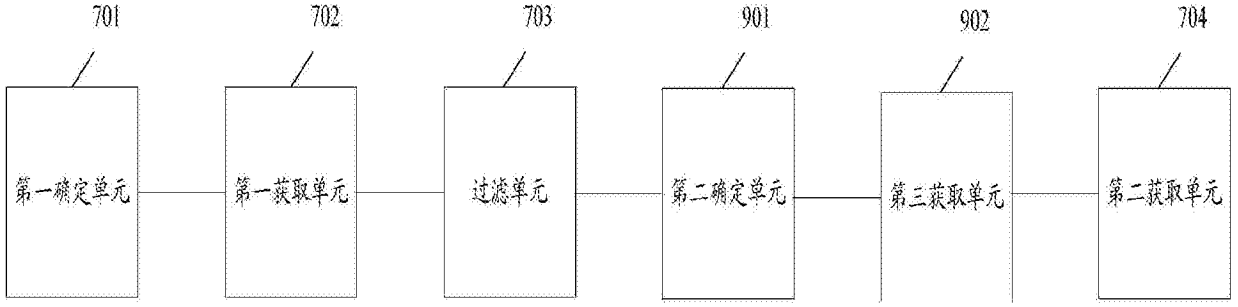

701
702
703
901
902
704
第一确定单元
第一获取单元
过滤单元
第二确定单元
第三获取单元
第二获取单元

图 9

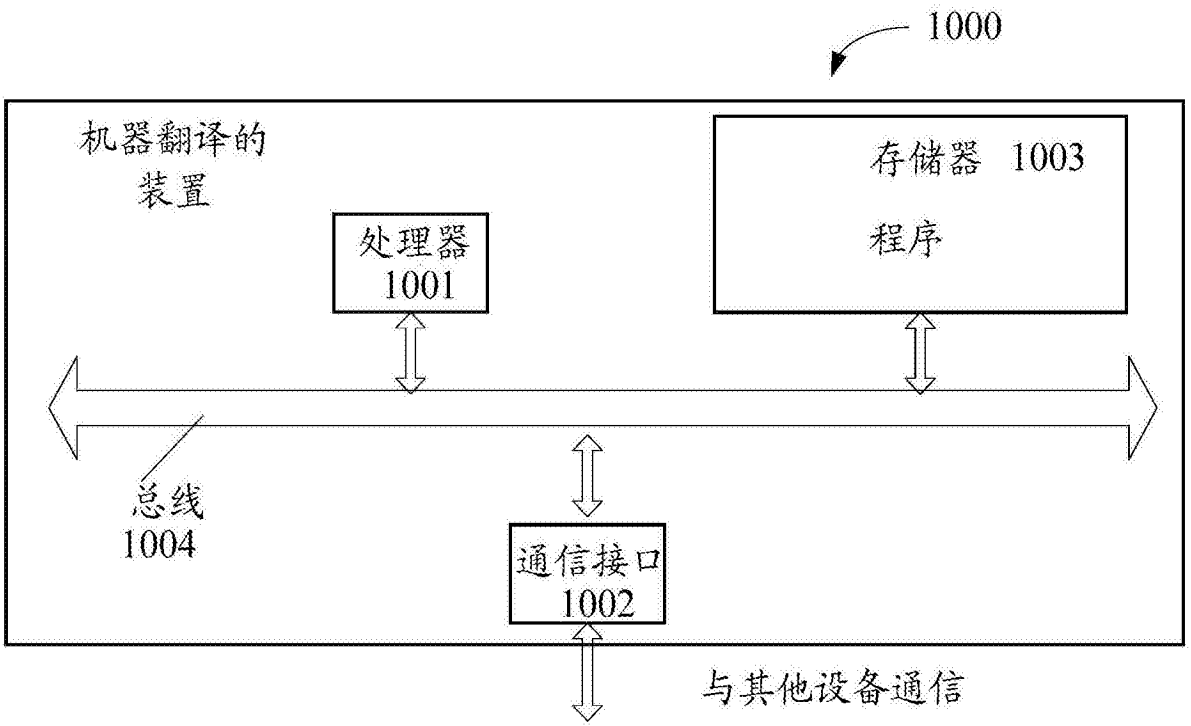

1000
机器翻译的
装置
处理器
1001
存储器 1003
程序
总线
1004
通信接口
1002
与其他设备通信

图 10