



(12) 发明专利申请

(10) 申请公布号 CN 102369524 A

(43) 申请公布日 2012.03.07

(21) 申请号 201080014350.3

(74) 专利代理机构 中科专利商标代理有限责任

(22) 申请日 2010.03.23

公司 11021

(30) 优先权数据

2009-081431 2009.03.30 JP

代理人 王波波

(85) PCT申请进入国家阶段日

(51) Int. Cl.

G06F 17/27(2006.01)

2011.09.29

(86) PCT申请的申请数据

PCT/JP2010/054920 2010.03.23

(87) PCT申请的公布数据

W02010/113691 JA 2010.10.07

(71) 申请人 日本电气株式会社

地址 日本东京都

(72) 发明人 安藤真一 定政邦彦

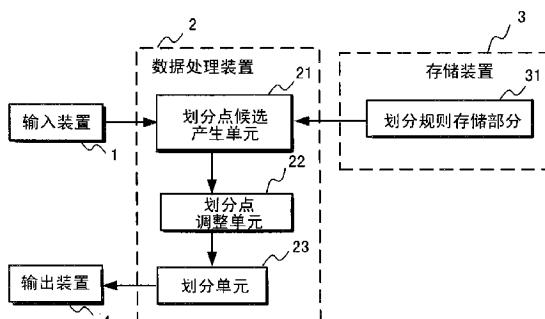
权利要求书 2 页 说明书 9 页 附图 5 页

(54) 发明名称

语言分析装置、语言分析方法和语言分析程序

(57) 摘要

本发明的语言分析装置包括：划分规则，根据在应用时引起分析准确度问题的风险程度，每种划分规则被分类至一种等级；划分点候选产生单元 21，当输入了长度大于预定的最大输入长度的字符串时，通过按照引起问题的风险等级递增的顺序逐一依次应用所述划分规则，来产生针对输入字符串的划分点候选；划分点调整单元 22，当划分点候选产生单元 21 所产生的所述划分点候选所获得的划分单元候选的长度小于所述最大输入长度时，从通过应用相同等级的划分规则同时确保每个划分单元的长度不大于所述最大输入长度而获得的划分点候选中，选择划分点的组合；以及划分单元 23，在所述划分点调整单元所确定的划分点处，将输入字符串划分。



1. 一种语言分析装置,包括:

划分规则,根据在应用时引起分析准确度问题的风险程度,每种划分规则被分类至一种等级;

划分点候选产生单元,当输入了长度大于预定的最大输入长度的字符串时,通过按照引起问题的风险等级递增的顺序逐一依次应用所述划分规则,来产生针对输入字符串的划分点候选;

划分点调整单元,当通过在划分点候选产生单元所产生的所述划分点候选处将所述字符串划分而获得的划分单元候选的长度小于所述最大输入长度时,从通过应用相同时级的划分规则同时确保每个划分单元的长度不大于所述最大输入长度而获得的划分点候选中,选择划分点的组合;以及

划分单元,在所述划分点调整单元所确定的划分点处,将输入字符串划分。

2. 根据权利要求 1 所述的语言分析装置,其中,当所述划分点调整单元确定所述划分单元候选的长度大于所述最大输入长度时,所述划分点候选产生单元通过应用低于先前划分规则等级的划分规则,来产生针对所述划分单元候选的新的划分点候选。

3. 根据权利要求 1 或 2 所述的语言分析装置,其中,从位于前端的划分单元候选开始,所述划分点调整单元计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整单元不将位于相邻划分点候选之间的划分点候选选为划分点。

4. 根据权利要求 1 或 2 所述的语言分析装置,其中,从位于末尾的划分单元候选开始,所述划分点调整单元计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整单元不将位于相邻划分点候选之间的划分点候选选为划分点。

5. 根据权利要求 1 或 2 所述的语言分析装置,其中,从具有最小长度的划分单元候选开始,所述划分点调整单元计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整单元不将位于相邻划分点候选之间的划分点候选选为划分点。

6. 一种语言分析方法,包括:

划分点候选产生步骤,当输入了长度大于预定的最大输入长度的字符串时,通过按照引起问题的风险等级递增的顺序逐一依次应用划分规则,来产生针对输入字符串的划分点候选,根据在应用时引起分析准确度问题的风险程度,每种划分规则已被分类至一种等级;

划分点调整步骤,当通过在划分点候选产生步骤中所产生的所述划分点候选处将所述字符串划分而获得的划分单元候选的长度小于所述最大输入长度时,从通过应用相同时级的划分规则同时确保每个划分单元的长度不大于所述最大输入长度而获得的划分点候选中,选择划分点的组合;以及

划分步骤,在所述划分点调整步骤所确定的划分点处,将输入字符串划分。

7. 根据权利要求 6 所述的语言分析方法,其中,当所述划分点调整步骤确定所述划分单元候选的长度大于所述最大输入长度时,所述划分点候选产生步骤通过应用低于先前划分规则等级的划分规则,来产生针对所述划分单元候选的新的划分点候选。

8. 根据权利要求 6 或 7 所述的语言分析方法,其中,从位于前端的划分单元候选开始,所述划分点调整步骤计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整步骤不将位于相邻划分点候选之间的划分点候选选为划分点。

9. 根据权利要求 6 或 7 所述的语言分析方法,其中,从位于末尾的划分单元候选开始,所述划分点调整步骤计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整步骤不将位于相邻划分点候选之间的划分点候选选为划分点。

10. 根据权利要求 6 或 7 所述的语言分析方法,其中,从具有最小长度的划分单元候选开始,所述划分点调整步骤计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整步骤不将位于相邻划分点候选之间的划分点候选选为划分点。

11. 一种语言分析程序,通过在计算机上运行来执行语言分析过程,所述语言分析程序使计算机执行:

划分点候选产生过程,当输入了长度大于预定的最大输入长度的字符串时,通过按照引起问题的风险等级递增的顺序逐一依次应用划分规则,来产生针对输入字符串的划分点候选,根据在应用时引起分析准确度问题的风险程度,每种划分规则已被分类至一种等级;

划分点调整过程,当通过在划分点候选产生过程中所产生的所述划分点候选处将所述字符串划分而获得的划分单元候选的长度小于所述最大输入长度时,从通过应用相同等级的划分规则同时确保每个划分单元的长度不大于所述最大输入长度而获得的划分点候选中,选择划分点的组合;以及

划分过程,在所述划分点调整过程所确定的划分点处,将输入字符串划分。

12. 根据权利要求 11 所述的语言分析程序,其中,当所述划分点调整过程确定所述划分单元候选的长度大于所述最大输入长度时,所述划分点候选产生过程通过应用低于先前划分规则等级的划分规则,来产生针对所述划分单元候选的新的划分点候选。

13. 根据权利要求 11 或 12 所述的语言分析程序,其中,从位于前端的划分单元候选开始,所述划分点调整过程计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整过程不将位于相邻划分点候选之间的划分点候选选为划分点。

14. 根据权利要求 11 或 12 所述的语言分析程序,其中,从位于末尾的划分单元候选开始,所述划分点调整过程计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整过程不将位于相邻划分点候选之间的划分点候选选为划分点。

15. 根据权利要求 11 或 12 所述的语言分析程序,其中,从具有最小长度的划分单元候选开始,所述划分点调整过程计算两个相邻划分单元候选的长度之和,并且当所述长度之和不大于所述最大输入长度时,所述划分点调整过程不将位于相邻划分点候选之间的划分点候选选为划分点。

语言分析装置、语言分析方法和语言分析程序

技术领域

[0001] 本发明涉及通过对自然语言进行语法分析来执行语言分析的方法。更具体地，本发明涉及语言分析装置、语言分析方法和语言分析程序，在将长句子划分为较短的句子时高效地执行语言分析过程。

背景技术

[0002] 典型地，通过首先将输入字符串划分为句子，然后对这些句子中的每个句子执行分析过程，来执行基于语法分析的语言分析。然而，当分析极长的句子（如经常在专利申请的说明书中看到的句子）时，基于逐句的简单分析过程可能会遇到某些问题。

[0003] 典型地，语言分析装置（如，用于语法分析的语言分析装置）通过将输入字符串划分为句子，然后研究每个句子中包含的每个单词对之间的关系，来执行分析过程。这意味着，所要考虑的单词对的数目随输入句子长度的增加成指数增长。

[0004] 如果要分析极长的句子，必须计算巨量的单词对。这将导致各种问题，包括：分析所需的较长的分析时间和大量的存储容量。

[0005] 此外，可能的解释方式的数目随所要考虑的单词对的数目的增加而增加。这进而提高了分析差错的可能。为避免如此，已经提出了各种方法：在执行分析过程之前，如果输入的句子过长，对输入的句子进行划分。

[0006] 例如，在专利文献 1 中，公开了一种方法，其中，如果机器翻译过程所花的时间大于预定时间，应用之前给定的划分规则将输入的句子划分为较小的单元，并对每个单元执行机器翻译过程。

[0007] 专利文献 2 中提出的方法与适应性单词计数相关联地存储划分规则，并按照适应性单词计数递减的顺序依次应用划分规则，使得输入的句子能够被划分为更合适的单元。

[0008] 专利文献 1：日本专利待审公开 No. 61-255468

[0009] 专利文献 2：专利号 003173514

[0010] 以下，将描述对输入的句子进行划分以执行基于语法分析的语言分析的上述方法存在的问题。

[0011] 第一个问题，是当给定了分析过程中可接受的最大输入长度（以下称“最大输入长度”）时，无法根据这样的最大输入长度将长句子划分为适当长度的处理单元。

[0012] 划分规则大致分为两类。一类划分规则关注提供相对宽松的中断的语言表述，另一类关注提供相对细致的中断的语言表述。一般而言，前一类划分规则允许分析得到正确地执行，即使不予改变地（即，在通过应用前一类划分规则获得的划分点处将句子划分后不作任何调整）对每个划分单元执行分析过程。然而，该规则关注于相对稀有的特定语言表述。由于可能未必从所有输入的句子中获得划分点，这可能是存在问题的，并且当实际获得划分点时，每个得到的划分单元可能不够短。

[0013] 另一方面，后一类划分规则通过关注于相对常用的语言表述来获得划分点。因此，该类划分规则允许从相对大量的句子获得划分点。此外，得到的划分单元可能足够短。然

而,由于各个划分单元可能变得过短以至于无法对每个划分单元执行正确的分析,这将引起分析准确度常常降低的问题。

[0014] 专利文献 2 中公开的划分方法试图通过与适应性单词计数相关联地存储划分规则,并按照适应性单词计数递减的顺序依次应用划分规则,来解决上述问题。然而,该方法也存在分析准确度降低的问题。一个原因在于,难以针对划分规则设置合适的适应性单词计数。另一个原因在于,当达到需要应用具有较小适应性单词计数的划分规则的阶段时,得到的划分单元变得过短以至于无法确保正确分析。

[0015] (本发明的目的)

[0016] 本发明的目的在于,提供语言分析装置和语言分析方法,根据分析过程中可接受的最大输入长度将长句子划分为合适长度的处理单元。

发明内容

[0017] 根据本发明的第一示例方面,一种语言分析装置包括:

[0018] 划分规则,根据在应用时引起分析准确度问题的风险程度,每种划分规则被分类至一种等级;

[0019] 划分点候选产生单元,当输入了长度大于预定的最大输入长度的字符串时,通过按照引起问题的风险等级递增的顺序逐一依次应用所述划分规则,来产生针对输入字符串的划分点候选;

[0020] 划分点调整单元,当通过在划分点候选产生单元所产生的所述划分点候选处将所述字符串划分而获得的划分单元候选的长度小于所述最大输入长度时,从通过应用相同等级的划分规则同时确保每个划分单元的长度不大于所述最大输入长度而获得的划分点候选中,选择划分点的组合;以及

[0021] 划分单元,在所述划分点调整单元所确定的划分点处,将输入字符串划分。

[0022] 根据本发明的第二示例方面,一种语言分析方法包括:

[0023] 划分点候选产生步骤,当输入了长度大于预定的最大输入长度的字符串时,通过按照引起问题的风险等级递增的顺序逐一依次应用划分规则,来产生针对输入字符串的划分点候选,根据在应用时引起分析准确度问题的风险程度,每种划分规则已被分类至一种等级;

[0024] 划分点调整步骤,当通过在划分点候选产生步骤中所产生的所述划分点候选处将所述字符串划分而获得的划分单元候选的长度小于所述最大输入长度时,从通过应用相同等级的划分规则同时确保每个划分单元的长度不大于所述最大输入长度而获得的划分点候选中,选择划分点的组合;以及

[0025] 划分步骤,在所述划分点调整步骤所确定的划分点处,将输入字符串划分。

[0026] 根据本发明的第三示例方法,一种语言分析程序通过在计算机上运行来执行语言分析过程,所述语言分析程序使计算机执行:

[0027] 划分点候选产生过程,当输入了长度大于预定的最大输入长度的字符串时,通过按照引起问题的风险等级递增的顺序逐一依次应用划分规则,来产生针对输入字符串的划分点候选,根据在应用时引起分析准确度问题的风险程度,每种划分规则已被分类至一种等级;

[0028] 划分点调整过程,当通过在划分点候选产生过程中所产生的所述划分点候选处将所述字符串划分而获得的划分单元候选的长度小于所述最大输入长度时,从通过应用相同等级的划分规则同时确保每个划分单元的长度不大于所述最大输入长度而获得的划分点候选中,选择划分点的组合;以及

[0029] 划分过程,在所述划分点调整过程所确定的划分点处,将输入字符串划分。

[0030] 根据本发明,如果设置了分析过程中可接受的最大输入长度,可以根据这样的最大输入长度将长句子划分为适当长度的处理单元。

[0031] 这是由于:依次应用划分规则,根据在应用时引起分析准确度问题的风险程度,每种划分规则已被分类至一种等级;以及将每个得到的划分点调整为使得每个划分单元将具有最大可能长度但不会超过最大输入长度。

附图说明

[0032] 图 1 是示出了根据本发明的第一示例实施例的语言分析装置的结构的框图;

[0033] 图 2 是示出了根据本发明的第一示例实施例的语言分析装置的操作的流程图;

[0034] 图 3 是示出了根据本发明的第一示例实施例的语言分析装置的操作的流程图;

[0035] 图 4 是示出了根据本发明的第二示例实施例的语言分析装置的结构的框图;

[0036] 图 5 是示出了根据本发明的第一示例实施例的划分规则存储部分中的示例数据结构的图;

[0037] 图 6 是示出了根据与根据本发明的第一示例实施例相对应的示例 1 的示例操作的图;以及

[0038] 图 7 是示出了根据本发明的第一实施例的语言分析装置的数据处理设备的示例硬件结构的框图。

具体实施方式

[0039] (第一示例实施例)

[0040] 下面将参考附图描述本发明的第一示例实施例。

[0041] 参照图 1 可见,根据本发明的第一示例实施例的语言分析装置包括:输入装置 1,如键盘或鼠标;数据处理装置 2,在程序指令的控制下操作;存储装置 3,存储信息;以及输出装置 4,如显示装置和打印装置。

[0042] 向存储装置 3 提供划分规则存储部分 31。划分规则存储部分 31 存储划分规则,划分规则将被应用于输入字符串以识别输入字符串中可能充当划分点的点。

[0043] 在根据在应用时引起分析准确度问题的风险程度将划分规则分组为至少两个等级后,划分规则存储部分 31 存储划分规则。

[0044] 例如,按以下方式执行将划分规则分组为某一等级。首先,将通过应用该划分规则获得的对每个划分单元执行语言分析的结果与在不划分输入字符串的情况下执行语言分析的结果进行比较。如果比较表明:划分将不会实质上引起与每个划分单元相对应的分析结果的改变,划分规则被认为是“无风险的”并被分类为“高级”。如果比较表明:划分以较低的概率引起该分析结果的改变,划分规则被认为是“低风险的”并被分类为“中级”。如果比较表明:划分以较高的概率引起该分析结果的改变,划分规则被认为是“高风险的”并被

分类为“低级”。

[0045] 数据处理装置 2 包括：划分点候选产生单元 21、划分点调整单元 22 和划分单元 23。

[0046] 如果作为处理目标输入的字符串的长度大于预定的最大输入长度，划分点候选产生单元 21 按照风险度递增的顺序，从存储在划分规则存储部分 31 中的划分规则中读取划分规则，并通过应用所读取的划分规则产生输入字符串内的划分点候选。

[0047] 此处，术语“划分点候选”指：在输入字符串中识别的可能充当划分点的点。术语“最大输入长度”指：输入字符串的可接受的最大长度。最大输入长度是根据语言分析的所需处理时间和可接受的存储开销而确定的值。例如，可以基于输入字符串中包含的字符或语素的数目来定义该值。

[0048] 划分点调整单元 22 接收划分点候选产生单元 21 所产生的划分点候选。划分点调整单元 22 依次逐一检查通过在每个划分点候选处将输入字符串划分可得到的划分单元候选，并确定其是否将划分点候选用作划分点。

[0049] 划分点调整单元 22 通过必要时选择划分点候选，以确保最终获得的全部独立划分单元的总数等于或小于最大输入长度，来决定划分点。

[0050] 划分单元 23 接收划分点调整单元 22 所确定的划分点，并通过在该划分点处将输入字符串划分来产生划分单元。

[0051] (示例实施例的操作)

[0052] 下面，将参照图 1 和 2 的流程图来详细描述本示例实施例的操作。

[0053] 当从输入装置 1 接收到输入字符串时，数据处理装置 2 的划分点候选产生单元 21 首先计算输入字符串的长度。接着，划分点候选产生单元 21 将所计算的长度与先前给定的最大输入长度进行比较，以确定输入字符串的长度是否大于最大输入长度（步骤 A1）。应当注意的是，虽然来自输入装置 1 的输入可以是简单的字符串，更优选地，输入包含构成输入字符串的语素以及这些语素的属性信息，如词根形式和词性。

[0054] 如果在步骤 A1 中输入字符串的长度等于或小于最大输入长度，则无需划分字符串，因此划分点候选产生单元 21 终止整个过程。

[0055] 如果输入字符串的长度大于最大输入长度，划分点候选产生单元 21 将该输入字符串设置为划分目标（步骤 A2）。

[0056] 接着，划分点候选产生单元 21 初始化划分规则等级，并将其设置为风险最低的等级，即“高级”（步骤 A3）。

[0057] 在步骤 A4 中，划分点候选产生单元 21 使用在以上步骤 A3 中设置的等级的划分规则，对在以上步骤 A2 中被设置为划分目标的字符串启动划分点产生过程。下面，将描述划分点产生过程。

[0058] 划分单元 23 基于通过划分点候选产生单元 21 执行的划分点产生过程获得的划分点，对作为划分目标的字符串进行划分（步骤 A5）。

[0059] 下面，将参照图 3 的流程图，详细描述划分点候选产生单元 21 和划分点调整单元 22 执行的划分点产生过程。

[0060] 划分点候选产生单元 21 从划分规则存储部分 31 获得所设置等级的划分规则，并通过对被设置为划分目标的字符串应用所获得的划分规则来产生划分点候选（步骤 B1）。

[0061] 如果通过应用划分规则未从划分目标中获得划分点候选,划分点候选产生单元 21 将划分规则等级降低一级(步骤 B2 和步骤 B3),并通过应用新等级的划分规则再次尝试产生划分点候选。更具体地,划分点候选产生单元 21 通过使用风险等级小一级的划分规则(即,风险度大于初始划分规则等级的划分规则),来产生划分点候选。

[0062] 如果不能再降低划分规则等级(图 3 中未示出),将终止划分点产生过程。

[0063] 如果已在图 3 的步骤 B1 中设置了划分点候选,划分点候选产生单元 21 将所设置的划分点候选传递至划分点调整单元 22。

[0064] 当接收到划分点候选产生单元 21 设置的划分点候选时,划分点调整单元 22 在划分点候选处将输入字符串划分为划分单元候选(步骤 B4)。

[0065] 接着,划分点调整单元 22 从所获得的划分单元候选中选择尚待检查的一个划分单元候选(步骤 B5)。

[0066] 选择划分单元候选的方法的一个示例可以是:首先选择在已从当前划分目标获得的并仍待检查的所有划分单元候选中、距字符串前端最近的划分单元候选,并依次移动。相反,可以首先选择仍待检查的所有划分单元候选中、距字符串末尾最近的划分单元候选。另一种可选方法是:以长度递增的顺序从仍待检查的所有划分单元候选中选择划分单元候选。

[0067] 接着,在步骤 B6 中,划分点调整单元 22 验证在步骤 B5 中是否成功选择了仍待检查的划分单元候选。

[0068] 如果未能在 B5 中选择仍待检查的划分单元候选,这意味着对所有划分单元候选完成了检查过程。在该情况下,划分点调整单元 22 采用剩下的未移除的划分点候选作为划分点,输出所获得的划分点,并终止过程(步骤 B7)。

[0069] 如果在步骤 B6 中验证能够选择仍待检查的划分单元候选,划分点调整单元 22 计算所选择的划分单元候选的长度,并将所计算的长度与预定的最大输入长度进行比较,以确定划分单元候选的长度是否大于最大输入长度(步骤 B8)。

[0070] 如果步骤 B8 中的比较指示所选择的划分单元候选的长度大于最大输入长度,划分点调整单元 22 将该划分单元候选设置为新的划分目标(步骤 B9),并将划分规则等级降低一级(步骤 B10)。

[0071] 接着,划分点调整单元 22 将过程交付划分点候选产生单元 21,划分点候选产生单元 21 进而通过使用新等级的划分规则为划分单元候选产生划分点候选(步骤 B11)。

[0072] 当该过程从划分点候选产生单元 21 返回时,划分点调整单元 22 回到步骤 B5 并继续该过程。

[0073] 如果在步骤 B8 中,所选择的划分单元候选的长度已被确定为小于最大输入长度,划分点调整单元 22 从当前划分目标中获得与所选择的划分单元候选相邻的新的划分单元候选(步骤 B12)。

[0074] 接着,划分点调整单元 22 验证是否能够成功获得相邻的划分单元候选(步骤 B13),如果不能则返回步骤 B5 并继续该过程。

[0075] 如果能够获得相邻划分单元候选,划分点调整单元 22 计算所选择的划分单元候选的长度和所获得的相邻划分单元候选的长度之和。接着,划分点调整单元 22 将得到的长度与预定的最大输入长度进行比较,以确定长度之和是否大于最大输入长度(步骤 B14)。

[0076] 如果在步骤 B14 中所计算的长度之和被确定为大于最大输入长度,划分点调整单元 22 将当前选择的划分单元候选确立为“被检查的”划分单元候选,并返回步骤 B2 以继续过程。

[0077] 如果在步骤 B14 中所计算的长度之和被确定为小于最大输入长度,划分点调整单元 22 移除位于所选择的划分单元候选和所获得的相邻划分单元候选之间的划分点候选(步骤 B15)。接着,划分点调整单元 22 将通过连接两个划分单元候选(即,所选择的划分单元候选和所获得的相邻划分单元候选)获得的划分单元设置为新的处理目标(步骤 B16),并返回步骤 B12 以继续过程。

[0078] (第一示例实施例的效果)

[0079] 下面将描述该示例实施例的效果。

[0080] 第一示例实施例被配置为:依次应用划分规则,根据在应用时引起分析准确度问题的风险程度,每种划分规则已被分类至多种等级中的一种等级,并且调整每个划分点,使得每个得到的划分单元的长度不是太短但不超过最大输入长度。因此,在设置了分析过程中可接受的最大输入长度时,可以根据这样的最大输入长度将长句子划分为适当长度的处理单元。

[0081] 此外,该示例实施例被配置为:其确立宽松的分类规则,其中,每一个划分规则与根据引起分析准确度问题的风险度定义的等级相关联;以及从而其调整每个划分点,使得每个得到的划分单元的长度不是太短但不大于最大输入长度。因此,在该示例实施例中,可以相对容易地创建划分规则,这是由于不必向其分类规则添加任何严格的优先级信息,例如,要对其应用划分规则的单词数目的下限。

[0082] (第二示例实施例)

[0083] 下面,将参照附图详细描述本发明的第二示例实施例。

[0084] 参照图 4,与本发明的第一示例实施例类似,本发明的第二示例实施例包括:输入装置 1、数据处理装置 6、存储装置 3 和输出装置 4。

[0085] 在被读入数据处理装置 6 之后,语言分析程序 5 控制数据处理装置 6 的操作,并在存储装置 3 中产生划分规则存储部分 32。

[0086] 在语言分析程序 5 的控制下,数据处理装置 6 也执行与根据第一示例实施例的数据处理装置 2 所执行过程相同的过程。

[0087] 数据处理装置 6 具有图 7 所示的硬件结构。

[0088] 如图 7 所示,可以采用与通用计算机装置类似的硬件结构来实现数据处理装置 6,并且数据处理装置 6 包括:CPU(中央处理单元)61;主存储部分 62,为主存储器(如, RAM(随机存取存储器))并被用作工作区和/或临时存储区;输入/输出接口部分 63,与输入装置 1、输出装置 4 和存储装置 3 相连,以发送和接收数据;系统总线 64,连接在上述组件之间。例如,存储装置 3 通过使用硬盘装置来实现,所述硬盘装置包括非易失性存储器,如 ROM(只读存储器)、磁盘和半导体存储器。

[0089] 不用说,可以通过实现电路组件,将根据该示例实施例的数据处理装置 6 的操作作为硬件实现,所述电路组件是硬件组件(如,LSI(大规模集成电路))并且并入了上述语言分析程序 5。此外,还可以通过在存储装置 3 中存储语言分析程序 5,将该程序加载至主存储部分 62 中,并在 CPU61 上执行该程序,将其作为软件实现。

[0090] (第一示例)

[0091] 下面将参照附图描述本发明的示例 1。该示例对应于本发明的第一示例实施例。

[0092] 示例 1 包括：作为输入装置 1 的键盘；作为数据处理装置 2 的个人计算机；作为存储装置 3 的磁盘记录设备；以及作为输出装置 4 的显示器。

[0093] 个人计算机具有 CPU，该 CPU 执行划分点候选产生单元 21、划分点调整单元 22 和划分单元 23 的功能。在磁盘记录设备中，用作划分规则存储部分 31 的存储区是受保护的。

[0094] 以下描述假定图 5 所示的划分规则存储在划分规则存储部分 31 中。图 5 以示意的表格形式示出了存储在划分规则存储部分 31 中的划分规则。每行包含划分规则。

[0095] 在图 5 中，第一列“等级”存储等级信息，指示应用对应的划分规则将引起分析精度问题的风险程度。下一列“划分点标识模式”存储模式信息，基于模式信息可识别划分点。最后一列“划分点”存储将要被识别为划分点的位置指示为划分点识别模式的相对位置的信息。

[0096] 例如，图 5 的第一行包含“等级 1”的划分规则，其应用将涉及引起分析准确度问题的最低风险。对于“等级 1”的划分规则，指示：如果输入字符串包含作为划分点识别模式的句号“。”，可以将紧接着句号（“模式”）之后的位置识别为划分点。

[0097] 图 5 的第二行包括“等级 2”的划分规则，其应用将涉及引起分析准确度问题的相对较低的风险。对于“等级 2”的划分规则，指示：如果输入字符串包含以连词“が”和逗号“、”的顺序出现的划分点识别模式，可以将紧接着该模式之后的位置识别为划分点。类似地，在图 5 的表中的第三和第四行，描述了等级 3 和等级 4 的划分规则。在划分规则的等级中，等级 1 是最高的，其次是等级 2、等级 3 和等级 4。

[0098] 下面描述假定最大输入长度已被设置为“100”。此外，将描述以下示例操作，其中，输入长度为 300 的字符串（日文句子）“～～～する。～～～するので、～～し、～～し、さらに～～する。”，并且使用图 5 所示的划分规则。

[0099] 当经由键盘等（即，输入装置 1）向作为数据处理装置 2 的个人计算机输入图 6 所示的输入句子时，划分点候选产生单元 21 接受输入句子，并计算其长度。基于该结果，划分点候选产生单元 21 检测到输入句子的长度是“300”，即，大于最大输入长度“100”，并将输入句子设置为划分目标。划分点候选产生单元 21 还初始化划分规则等级，并将其设置为最高的“等级 1”，并且执行划分点产生过程。

[0100] 在划分点产生过程期间，划分点候选产生单元 21 首先从划分规则存储部分 31 获得“等级 1”划分规则，并将其应用于划分目标，即图 6a 所示的整个输入句子。接着，划分点候选产生单元 21 检测到划分目标内的句号“。”，因此将后续部分设置为划分点候选。

[0101] 接着，划分点调整单元 22 接收划分点候选产生单元 21 所设置的划分点候选，并将划分目标输入句子划分为划分单元候选。该结果如图 6b 所示。

[0102] 接着，划分点调整单元 22 从划分目标中选择尚待处理的一个划分单元候选。作为此处所使用的选择尚待检测的划分单元候选的方法，此处将采用以下方法：首先选择所有划分单元候选中距划分单元候选前端最近的划分单元候选，并依次移动。

[0103] 首先，选择图 6 中所示的长度为“60”的首个划分单元候选，即“～～～する。”。接着，划分点调整单元 22 计算划分单元候选的长度，并确定其长度“60”小于最大输入长度“100”。

[0104] 接着,划分点调整单元 22 从划分目标中获得与当前选择的划分单元候选“～～～する。”相邻的划分单元候选“～～～するので,～～し、～～し、さらに～～する。”。

[0105] 由于这两个长度之和为“300”,划分点调整单元 22 将当前选择的划分单元候选“～～～する。”确立为“被检查的”划分单元候选,并选择下一个尚待处理的划分单元候选“～～～するので、～～し、～～し、さらに～～する。”作为新的处理目标。

[0106] 该划分单元候选的长度为“240”,大于最大输入长度“100”。因此,划分点调整单元 22 将该划分单元候选设置为划分目标,将划分规则等级降低一级至“等级 2”,并递归调用划分点产生过程。

[0107] 与以上类似,划分点候选产生单元 21 从划分规则存储部分 31 获得图 5 所示的“等级 2”的划分规则,并将其应用于划分目标,即,图 6b 中所示的第二划分单元候选。

[0108] 划分点候选产生单元 21 检测到以连词“ので”和逗号“、”的顺序连续出现在划分目标中的模式,因此将后续部分设置为划分点候选。该结果如图 6c 所示。

[0109] 采用以上相同的方式,划分点调整单元 22 将长度小于最大输入长度的划分单元候选“～～～するので、”(长度为“80”)确立为“被处理的”划分单元候选。接着,划分点调整单元 22 将长度大于最大输入长度的划分单元候选“～～し、～～し、さらに～～する。”(长度为“160”)设置为划分目标,将划分规则等级降低一级至“等级 3”,并递归调用划分点产生过程。

[0110] 与以上类似,划分点候选产生单元 21 从划分规则存储部分 31 获得“等级 3”的划分规则,并将其应用于划分目标,即,图 6c 中所示的第三划分单元候选。划分点候选产生单元 21 在两个位置检测到以接续形式的动词“レ”和逗号“、”的顺序连续出现在划分目标中的模式,因此将其相应的后续部分设置为划分点候选。该结果如图 6d 所示。

[0111] 划分点调整单元 22 将首个划分单元候选“～～し、”(长度为“50”)设置为尚待处理的划分单元候选,并确定其长度“50”小于最大输入长度“100”。

[0112] 接着,划分点调整单元 22 从划分目标中获得与首个划分单元候选“～～し、”相邻的第二划分单元候选“～～し、”。

[0113] 划分点调整单元 22 还计算这两个划分单元候选的长度之和,并确定该长度之和“80”小于最大输入长度“100”。

[0114] 划分点调整单元 22 移除位于这两个划分单元候选之间的划分点候选,并将已通过联合上述两个划分单元候选所获得的划分单元候选“～～し、～～し、”(长度为“80”)设置为新的处理目标。该结果如图 6e 所示。

[0115] 划分点调整单元 22 选择下一个相邻的划分单元候选“さらに～～する。”(长度为“80”)。此次,所要处理的划分单元候选的和是“160”,超过了最大输入长度“100”。相应地,划分点调整单元 22 将划分单元候选“～～し、～～し、”(长度为“80”)确立为“被处理的”划分单元候选,并选择下一个尚待处理的划分单元候选“さらに～～する。”(长度为“80”)作为新的处理目标。

[0116] 由于该划分单元候选的长度小于最大输入长度“100”,并且不存在尚待处理的相邻划分单元候选,划分点调整单元 22 将所获得的划分点候选确立为最终选择并产生划分点。

[0117] 划分点调整单元 22 以递归方式返回过程,发现不存在尚待处理的划分单元候选。

因此,其逐一确立所有所获得的划分点候选,并产生对应的划分点。

[0118] 最后,划分单元 23 使用所获得的划分点对输入的句子进行划分,并输出得到的 4 个划分单元:“～～～する。”、“～～～するので、”、“～～し、～～し、”和“さらに～～する。”。

[0119] 虽然以优选的示例实施例和示例为例对本发明进行了描述,应当意识到:本发明不限于这样的示例实施例和示例,相反,可以在不背离其技术原理的前提下以各种修改方案予以实现。

[0120] 相关申请的交叉引用

[0121] 本申请基于并要求于 2009 年 3 月 30 日递交的日本专利申请 No. 2009-081431 的优先权,其公开内容全部被并入于此作为参考。

[0122] 工业实用性

[0123] 根据本发明的语言分析装置能够适用于各种应用,包括语法分析装置和机器翻译装置,所述语法分析装置对以第一语言描述的文档进行语法分析以输出每个句子的句法,所述机器翻译装置将以第一语言(某一语言)描述的文档翻译成第二语言(另一语言)。

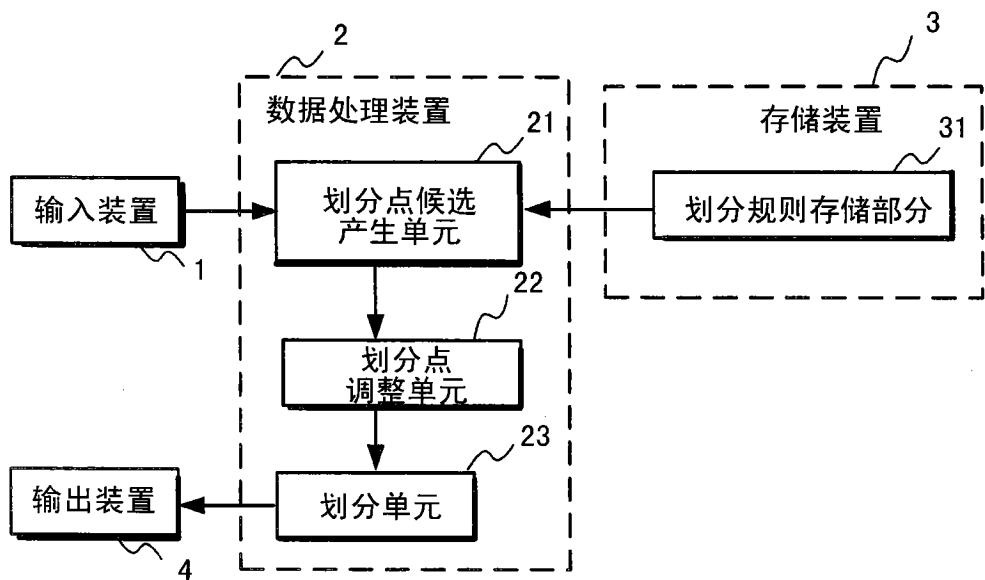


图 1

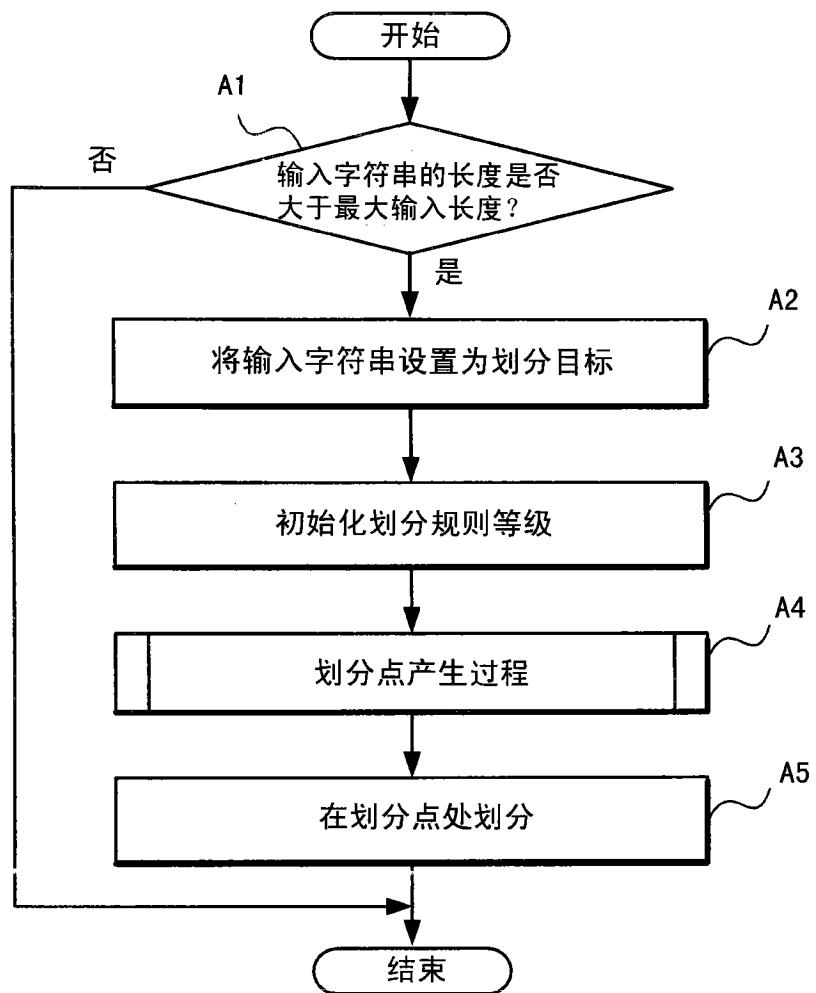


图 2

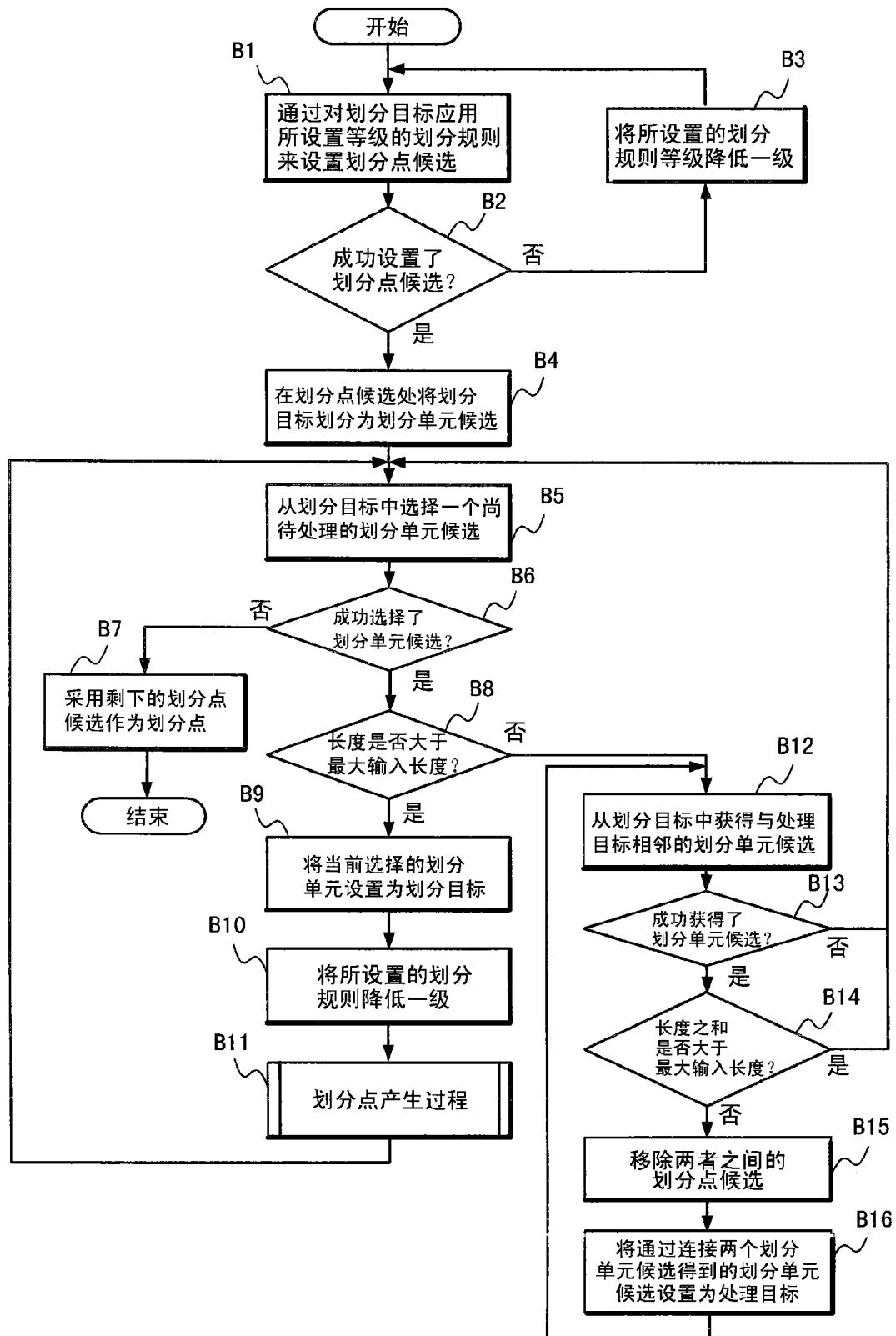


图 3

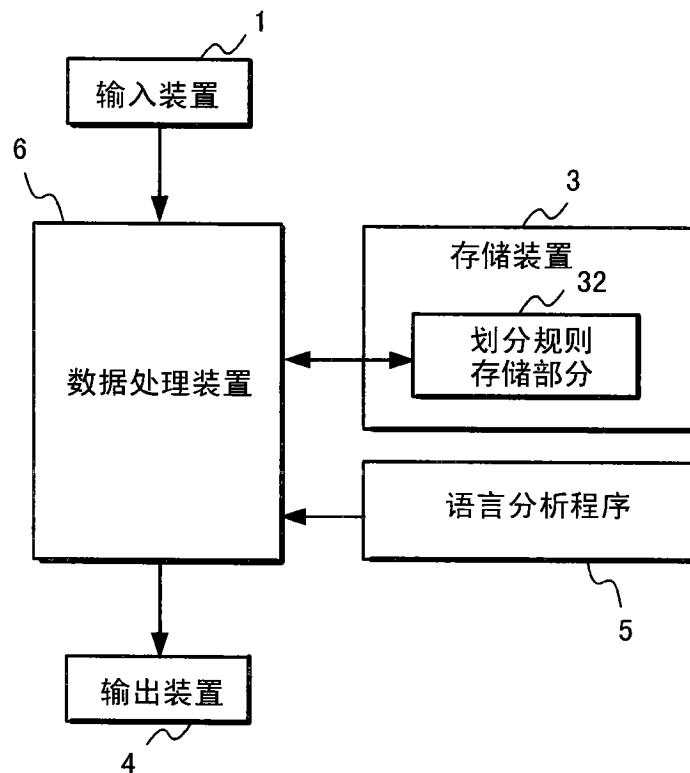


图 4

等级	划分点识别模式	划分点
1	句号“。”	紧接着模式之后
2	连词“が” + 逗号“、”	紧接着模式之后
	连词“ので” + 逗号“、”	紧接着模式之后
3	连词“し” + 逗号“、”	紧接着模式之后
	接续形式的动词 + 逗号“、”	紧接着模式之后
4	从句	紧接着模式之后

图 5

最大输入长度是 100

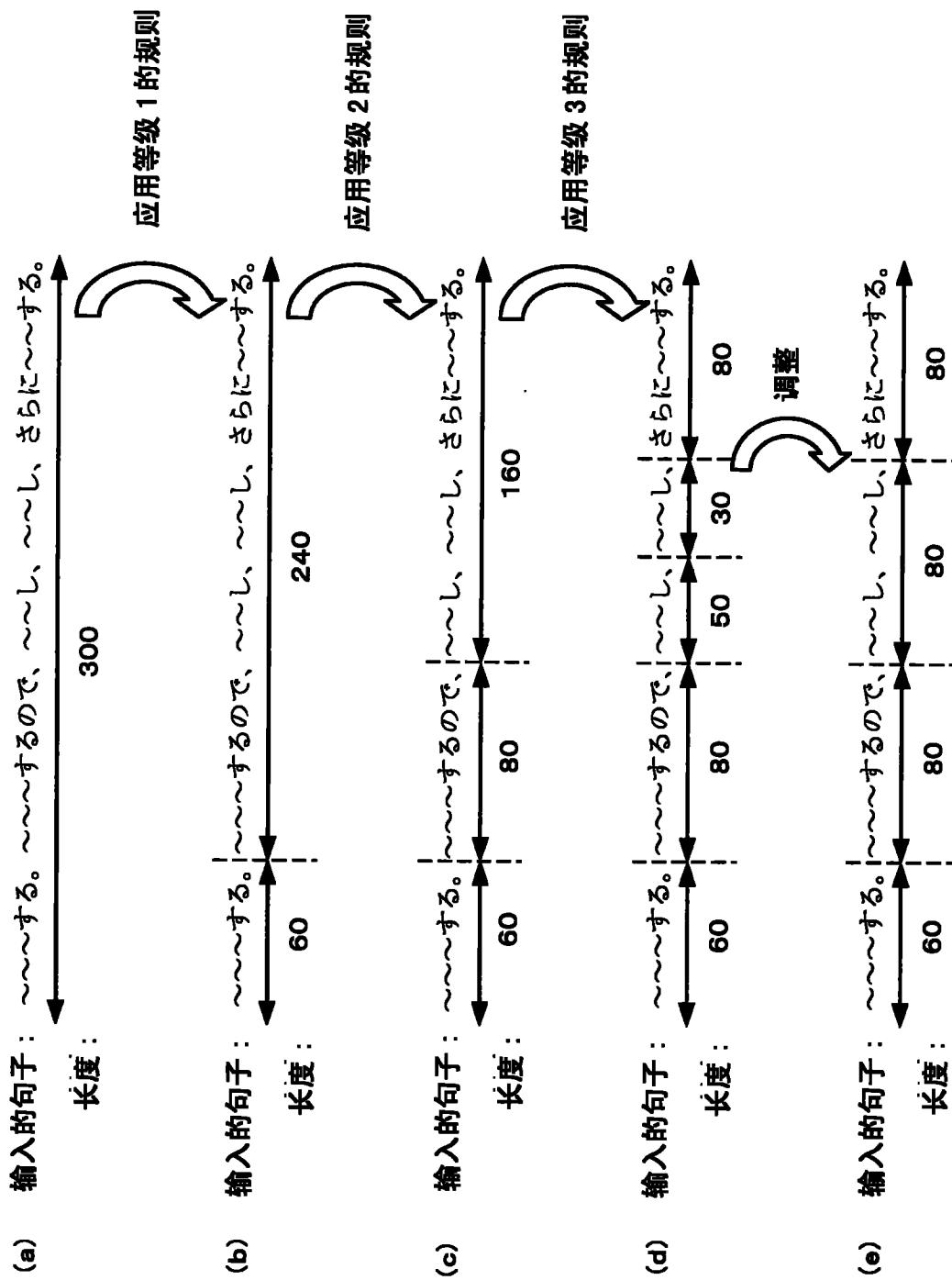


图 6

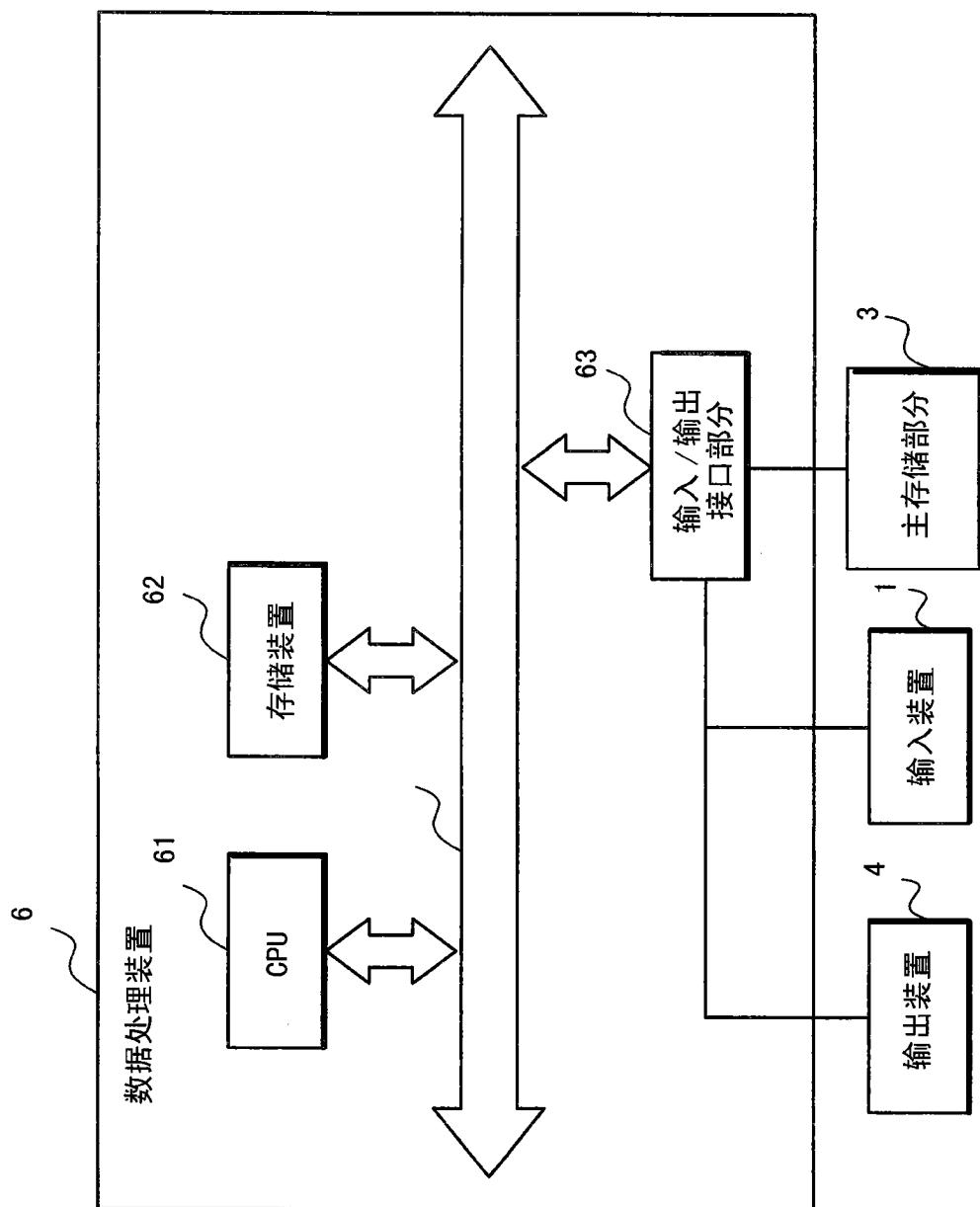


图 7