[19]中华人民共和国专利局

[11] 公开号 CN 1110882A

[12] 发明专利申请公开说明书

[21]申请号 94190391.5

[51]Int.Cl[6]

[43]公开日 1995 年 10 月 25 日

G06F 15/38

[22]申请日 94.6.17

[30]优先权

[32]93.6.18 [33]GB[31]9312598.7

[86]国际申请 PCT/GB94/01321 94.6.17

[87]国际公布 WO95/00912 英 95.1.5

[85]进入国家阶段日期 95.2.17

[71]申请人 欧洲佳能研究中心有限公司

地址 英国吉尔福德郡

共同申请人 欧洲佳能有限公司

[72]发明人 T·F·奥唐诺休

[74]专利代理机构 中国专利代理(香港)有限公司

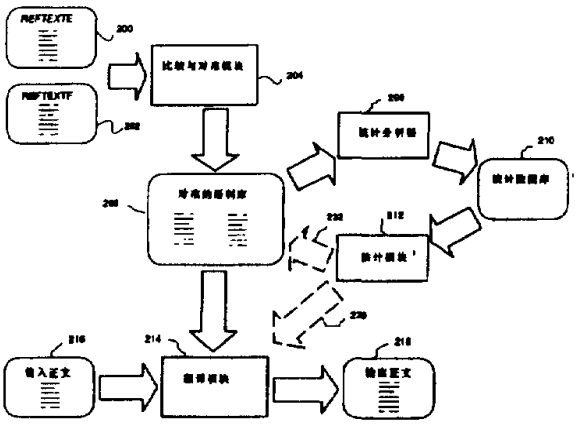
代理人 董江雄 萧掬昌

说明书页数: 附图页数:

[54]发明名称 处理两种文字对照的数据库的方法与装置

[57]摘要

生成或从外部源接收对准的语料库（206，CORPE，CORPF）。各语料库包括与另一语料库的相应部分对准的部分，它们是两种自然语言的互为译文。编制一个统计数据库（210），估计模块（212）为从各语料库中选出一个字的字对计算相互关联评分。给定一个正文部分对（每一种语言中一个部分），估计模块（212）组合字对相互关联评分以得出该正文部分的对准评分。这些对准评分可用于检验一个译文（230）与/或修正对准的语料库（206）以消除不可信的对准。



(BJ)第 1456 号

权利要求书　CPEL 955046P

1．一种操作用于处理一个两种文字对照的数据库的处理装置的方法，包括下述步骤：

在该装置中存储一个包括第一与第二对准的语料库（正文信息本体）的数据库，各语料库是分成多个部分的，使得第一语料库的部分的至少一个子集是用与第二语料库的对应部分的一种对应关系标识的，各部分是进一步分成一个或多个基元部分（“字”）的；以及

分析对准的语料库以得出一个统计数据库；

其特征在于：

使用该统计数据库，为选自对准的语料库的对准部分估计一个相互关联测定值；

从对准的部分的相互关联测定值中，标识未必可信的对准的事例；以及

通过修正对准的语料库来至少消除最不可信的对准，而纠正所标识的不正确对准。

2．权利要求1中提出的一种方法，还包括按照对准的语料库的修正，更新统计数据库。

3．一种处理一个两种文字对照的数据库的方法，包括：

用根据权利要求2的一种方法修正对准的语料库，以及

重复该方法至少一次，用更新后的统计数据库再一次修正经过修正的语料库。

4．权利要求1、2或3中提出的一种方法，其中的修正步骤包括与一位翻译人员的交互对话，来检验标识的未必可信的对准是不正

确的。

5．前面任何一项权利要求中所提出的一种方法，还包括接收正文的更多的对准的部分对，并扩大对准的语料库与统计数据库，将接收的部分包括进来。

6．权利要求5中提出的一种方法，其中为各更多的部分对估计所述相互关联的测定值，并且取决于所述估计的结果，有条件地扩大对准的语料库。

7．前面任何一项权利要求中所提出的一种方法，其中所述估计步骤包括：

利用统计数据库，为从每一个语料库中选取一个字的一个字对推算出观测到的相互关联的一个测定值；以及

通过组合包含在该部分对中的字对的推算出的相互关联测定值，估计这两个部分的相互关联的所述测定值。

8．权利要求7中提出的一种方法，其中所述字对相互关联测定值是在不推算该对中的任一个字是另一个的真实译文的概率的情况下推算的。

9．权利要求7或8中提出的一种方法，其中字对的相互关联测定值的组合是不考虑一对中的字在它们各自的部分中的位置而执行的。

10．权利要求7或8中提出的一种方法，其中字对的相互关联测定值的组合是限制在出现在它们各自的正文部分中的大致上对应的位置上的字对的。

11．权利要求7、8、9或10中提出的一种方法，其中为了字对相互关联测定值的组合，标识及略去了某些常用字。

12．权利要求7至11中任何一项所提出的一种方法，其中推

算字对相互关联测定值的步骤包括：

为一对中的各字推算在其各自的语料库中找到该字的观测概率的测定值；

为各选择的字对推算在语料库的对准部分中找到该字对的观测概率的测定值；以及

用单个字的概率组合对的概率，以推算对中的字之间的相互关联的所述测定值。

13．前面的任何权利要求中所提出的一种方法，其中该统计数据库包括：

用于各语料库的一张字频表；

用于作为一个整体的对准的语料库的一张字对频率表，计数一个给定的字对（各语料库一个字）出现在语料库的对准部分中的次数。

14．权利要求13中提出的一种方法，其中所述字对频率是不考虑各部分中字的位置，而对各对正文部分的所有字对计数的。

15．权利要求13中提出的一种方法，其中的字对的计数是限制在出现在语料库的它们各自的对准部分中的大致上对应的位置上的那些字对的。

16．权利要求13、14或15中提出的一种方法，其中为了字对的计数，而标识与略去某些常用字。

17．一种操作一个处理装置将一篇源正文从一种源语言自动翻译成一种目标语言的方法，包括：

在该装置中存储一个包括源与目标语言的对准的语料库的两种文字对照的数据库；

用根据权利要求1－16中任何一项的的一种方法修正对准的语

料库；

将源正文划分成部分；

在源语言的对准的语料库中搜索与源正文部分匹配的部分；以及

（ｉ）对于找到匹配的源正文部分，输出目标语言语料库中的对应部分，作为供包含进一个输出正文中的译文；以及（ｉｉ）对于找不到匹配的部分，输出一则查询报文，表示未进行翻译。

１８．一种操作用于处理一个两种文字对照的数据库的处理装置的方法，包括下述步骤：

在该装置中存储一个包括第一与第二对准的语料库（正文信息本体）的数据库，各语料库是分成部分的，使得第一语料库的部分的至少一个子集被与第二语料库的对应部分的一种对应关系所标识，各部分是进一步分成一个或多个基元部分（“字”）的；

分析对准的语料库以得到一个统计数据库；以及

利用该统计数据库，为从各语料库中选出一个字的一个字对，推算观测的相互关联的一个测定值，

其特征在于该方法还包括通过组合包含在各部分对中的字对的推算出的相互关系测定值，为一个给定的正文部分对（两种自然语言中每种一个部分）估计部分的相互关联的一个测定值。

１９．权利要求１８中提出的一种方法，其中所述字对相互关联测定值是在不推算一对中的任一字真实地是另一个字的译文的概率的情况中推算的。

２０．权利要求１８或１９中提出的一种方法，其中的字对相互关联测定值的组合是在不考虑一对中的字在它们各自的部分中的位置的情况中执行的。

21．权利要求18或19中提出的一种方法，其中的字对相互关联测定值的组合是限制在出现在它们各自的正文部分中大致上对应的位置上的那些字对的。

22．权利要求18至21中任何一项中提出的一种方法，其中为了组合字对相互关联的测定值，标识与略去了某些常用字。

23．权利要求18至22中任何一项中提出的一种方法，其中的推算字对相互关联测定值的步骤包括：

为一对中的各字推算在其各自语料库中找到该字的观测到的概率的一个测定值；

为各选择的字对，推算在语料库的对准部分中找到该字对的观测到的概率的一个测定值；以及

用单个字的概率组合对的概率，以推算对中的字之间的相互关联的所述测定值。

24．权利要求18至23中任何一项中提出的一种方法，其中的统计数据库包括：

各语料库的一张字频表；

作为一个整体的对准语料库的一张字对频率表，计数一个给定的字对（每一个语料库一个字）出现在语料库的对准部分中的次数。

25．权利要求24中提出的一种方法，其中所述字对频率是在不考虑各部分内的字的位置的情况中，为各正文部分对的所有字对计数的。

26．权利要求24中提出的一种方法，其中的字对计数是限制在出现在它们各自的语料库的对准部分中大致上对应的位置上的字对的。

27．权利要求24、25或26中提出的一种方法，其中为了字对的计数，标识与略去了某些常用字。

28．一种操作一个处理装置将一个源正文从一种源语言自动翻译成一种目标语言的方法，包括：

在该装置中存储包含源与目标语言的对准的语料库的两种文字对照的数据库；

将源正文划分成部分；

得出一个候选目标语言部分形式的各源正文部分的候选译文；

通过用按照权利要求18至27中任何一项的方法，为源语言部分与候选目标语言部分估计一个相互关联测定值，而检验该译文；

输出经过检验的候选部分，供包含进一个目标正文中。

29．权利要求28中提出的一种方法，其中的检验步骤包含将一个阈值作用在一个候选部分的相互关联测定值上。

30．权利要求28或29中提出的一种方法，还包括输出一则表示并未为源语言部分作出经过检验的译文的查询报文。

31．权利要求17、28、29或30中提出的一种方法，其中翻译部分与查询报文（如果有的话）是作为一个单一的输出正文的一部分连接与一起输出的，以供其它装置定稿。

32．一种处理装置，包括用于实现按照前面的任何一项权利要求的一种方法的步骤的装置。

33．一种存储设备，其中存储了经过按照权利要求1至16中任何一项所提出的操作的一种处理装置修正过的一个两种文字对照的数据库。

34．一种自动化翻译系统，包括权利要求32中所提出的一种

装置。

35．一种自动化翻译系统，其中存储了经过按照权利要求1至16中任何一项操作的一个装置修正过的数据库。

处理两种文字对照的数据库的方法与装置

技术领域

本发明涉及处理包括对准的语料库的两种或多种文字对照的数据库的方法与装置，用这种数据库自动翻译的方法与装置。

背景技术

对准的语料库为分成对准的部分的两种（或以上）正文体，使第一种语言语料库中的各部分映射到第二种语言语料库的对应部分上。各部分通常可包括一个单句或短语，但也可包括一个字或者甚至整个段。对准的语料库可用作自动翻译系统中的一个数据库，其中给定了第一种语言中的一个字、短语或句子时，如果它与已经存在在数据库中的一个部分匹配或以某种方式相似时，便能自动地得出第二种语言中的对应译文。这一原理可扩展到使两种以上的语料库对准，以便翻译成多种语言。

在五六十年代，普遍认为在不久的将来研制通用翻译系统是可能的。但由于需要大量背景信息与“智能”，所以稍后便意识到这种系统是甚为遥远的甚至有可能是根本不能实现的。但也意识到对准的语料库可用于小型专业化领域内的自动翻译。这是因为具有许多不同意义的“问题字”在一个专业化的活动领域的范围内，会趋于具有十分有限的意义范围。

然而，在建立这种专业化翻译系统中，尤其是假定为一种活动领域生成的数据库是理想地依据大量以前翻译的文件而可能不适用于另一领域的应用时，生成高质量的对准的语料库的问题仍然是第一位的。首先，要求工作在各领域中的用户生成他们自己的数据库，而这倾向于否定这种自动化系统的使用，从而仍然依赖于人类翻译人员。例如，美国专利5,140,522 描述了一种机器翻译系统，其中在使用过程中建立起一个以前翻译过的句子的数据库，但并未公开不用人类翻译人员的初始努力而得到这一数据库的任何方法。

为了论述上述问题，现在作为GB－A－2272091公布的共同未决英国专利申请描述了生成对准的语料库的一种自动化系统。在这里引入GB－A－2272091的内容。该自动化系统响应字处理装置在大多数文件中插入的格式码，诸如指明新的一章的标题或表中的新项。对于包含诸如电子装置的使用说明书等在内的各种文本，这些格式码之间的正文部分小到足以用作对准语料库中的对准部分。从而，在上述申请中所描述的系统比较简单，这在于无需判断字的意义，也无需将正文剖析成句子或更小的单元。另一方面，由于种种原因，得到的对准是不完美的，使得数据库中包含不正确的对准的形式的“噪声”。

在计算语言学会第２９次年会（Berkeley，Caliqornia）会报中，诸如W A Gale与K W Church在“两种文字对照的语料库中对准句子的程序”，以及P．F．Brown 等人在“在并行语料库中对准句子”中已经描述了自动化生成对准的语料库的替代方法。Brown 等人提出的系统在欧洲专利申请EP A－0525470 中有更全面的描述。在这些系统中，所使用的部分对应于句子，并且对准是通过比较句子的长度，或者以字数（Brown 等人）或者以字符数（Gale与Chureh）来进行的。

当然，用这些方法得出的对准语料库也会包含错误的对准，根据文献，至少达到百分之几的水平。

Brown 等人的文献描述了在自动生成的对准语料库的一个小的抽样（一百万对句子中的一千对）上人工进行的一次随机检验。这一工作揭示存在着一定观察到的概率的错误，但是在假定人工检验整个数据库是不现实的情况下，并未提出检测与纠正任何明显的错误部分的任何可实行的方法。再者，由于“锚定点”的一次比较揭示了节之间的不匹配，已经丢弃了大量的节（大约语料库的１０％）。由于Gale与Church提出的自动对准方法是基于句子长度的或然相关性的，所以这两位作者建议只需简单地略去最小可能性的对准便能消除许多错误的对准。这种取舍可能是有价值的，但数据库的质量仍然受到句子长度的相互关联是对准的唯一关键这一假设的限制。

Brown 等人的翻译系统（EP－A－0525470）利用在其间进行翻译的源与目标语言的相对地复杂的统计模型，使得数据库中低程度的“噪声”是能够容忍的。然而，对于在US 5,140,522及GB-A-2272091中所描述的那种较简单的基于存储器的系统，对于一个给定的句子中的各种不正确对准可以导致输出完全不正确的译文。

发明的公开

本发明采用统计技术来检测对准正文部分中可能存在的错误。可在使用前用于消除数据库中的错误对准，与　或应用一个“有噪声的”数据库或某种其它方法来排除已得出的错误译文。在特定的实施例中，本发明能够推导出一个评分来测定两种文字对照的字对的相互关联。然后可将字对的评分综合以推导出对准部分的任何提出的对的评分。这些部分可以是从外部接收的，或者是来自数据库本身的对准部分。

与作为一个整体的数据库的统计数字相比，可以从数据库中消除呈现为错误的对准。

因此，本发明能以最少的人工干预及处理要求改进包含对准的语料库的数据库。特别是，即使所实现的处理是基于统计的，并且处理器对语料库的语义与语法保持不注意的情况，在实践中也发现能用相对廉价的处理设备快捷地进行高质量的对准语料库的生成。

因为不论用什么技术从前面翻译的文件生成对准语料库时都能独立地实现本技术，它便能用于改进现存的数据库，及检测出执行原始对准的装置不能提示的错误。

注意到EP－A－0499366（英国与外国圣经协会）描述了一种检验由翻译生成的两种文字对照的语料库的过程。这一过程计算字对的评分，并通过一个重复的过程，建立一个翻译"字典"。然后用它来突出某些字的翻译中的可能不符合性。

本发明还提供翻译方法与装置、经过处理的数据库之类，如从属权利要求中所提出的。

下面参照附图用实例的方法描述本发明的实施例，附图中：

图 1 示出体现本发明的一种翻译系统的硬件；

图 2 示出图 1 的系统的操作结构；

图 3 示出包括一对对准的语料库的数据库结构；

图 4 示出用于图 3 的数据库的字频表；

图 5 示出用于图 3 的数据库的配对频率表；

图 6 为该系统的统计分析程序的操作的示意性流程图；

图 7 为该系统中的一个估计模块的部分操作的流程图；

图 8 为该估计模块的另一部分操作的流程图；

图 9 为使用该估计模块的数据库的改型的流程图；

图 1 0 为使用该改型后的数据库翻译一个正文的流程图；以及

图 1 1 为一个示例数据库中的对准评分的直方图。

具有部分 A 至 G 的附录提出英文与荷兰文的对准语料库的一个实例，以及分析程序与估计模块对该实例的操作结果。

在图 1 的系统中，用户的人机对话是用键盘 1 0 与显示屏 1 2 进行的，处理器单元 1 4 包括全都是传统构造的一个中央处理器（CPU）、半导体存储器（RAM与ROM）与接口电路。一个磁与/或光盘存储器 1 6 为存储多种文字对照的数据库、要翻译的正文及用于控制整个系统的操作的程序的大容量存储器。设置了一个可装卸的盘存储器 1 8 用于与系统进行新的数据与程序的通信。

本实施例的一个优点是该系统的上述硬件是可以从个人计算机或工作站型计算机购买的。图 2 示出图 1 的系统的操作结构。该系统存储有包含已经存在的两种或两种以上语言的一篇或多篇参照正文的源数据。例如，图 2 中200 所示的一篇正文REFTEXT1E 为存储在202 的一个法文文件REFTEXT1F 的一篇英文译文。提供了一个对准模块204，它能读取这种正文对并生成对应的对准的语料库对，如图 2 中206 所示。对准的语料库206 构成一个供翻译新文件用的两种文字对照数据库的主要部分。

提供了一个为对准语料库206 生成一个统计数据库210 的分析程序模块208 。提供了一个用统计数据库210 中的信息来测定对准语料库或其它正文中的对准质量的估计模块212。提供了一个读取一个输入正文（例如通过磁盘驱动器 1 8 ）来生成一个输出正文文件218 的翻译模块214 。对于任何模块，与一位操作人员的人机对话是可能的，

例如，使翻译模块214 能够咨询一位熟练的人类翻译家。

应当理解，在实践中各种结构都是可能的，例如，翻译模块及分析与估计模块204、208与212 可设置在分立的装置中，而数据库信息206与 ′或210则生成在一个装置中并与翻译用的一个第二装置进行通信。这在实践中的用处在于可在中心生成与估计及维护对准的语料库，然后将要翻译的文件分配给想要用个人计算机之类在他们家中或小型办公室中用诸如翻译模块214 远程地工作的熟练的翻译家与编辑。

如上所述，对准语料库发生器204 可具有已知的设计，诸如前面的申请GB-A-2272091或者在Gale与Church或Brown 等人的论文中所描述的。再者，生成原始对准语料库的方法与对本发明的理解无关，因此不再详述。图 3 示意性地示出在包含一对对准语料库的情况下的两种文字对照的数据库206 的结构。图 3 的例子是小的并且只是示意性地提出的，但是一个小而全的实例将参照附录在以下描述。

在图 3 中，一个英文语料库CORPE 包括正文的多个部分，它们是可用数字寻址的并称作“块”CORPE[I]，其中 I = 1， 2， 3 等。取决于生成对准语料库的方法，各块可大致上对应于一个原始源文件的一个句子，或正文的一个或长或短部分。两种情况中，各块CORPE[I]包括用CORPE[I][J]参照的可变数量的较小基元。在本例中，这些较小的基元为英文正文中的单字。从而，诸如块CORPE[1]包括两个字：CORPE[1][1]为“Good”而CORPE[1][2]为“day”。CORPE[3][1]为字“No”，而CORPE[6][1]为字“Yes”。字CORPE[5][2]与CORPE[4][3]在图中加上了标号供进一步示例。

在图 3 的另一侧为一个第二（法语）语料库CORPF，它包含相等数目的块CORPF[I]，每一个对应于英文语料库中的相同编号的块

CORPE[I]。特别是，一个关系REL至少在某种名义的意义上，确定各英文块CORPE[I]为对应的法文块CORPF[I]的一种译文。虽然各块是与对方语料库中正好一块对准的，但对准的块内的字数则不必要相等。例如，法文语料库中第一块包括单字“Bonjour ”但英文语料库中第一块则包括两个字“Good”与“Day ”，如所示。对比文件中还描述了包括一个句子的一个部分与包含两个句子的一个部分对准的实例。

应能理解，在对准的语料库CORPE与CORPF中的块的对准仅此而已。这些对准尚未经过一位熟练的翻译人员逐条核对，并且只是对比文件中所描述的那种自动化比较的“推测”。很清楚，这种对准并不表明两个块互为译文，甚至并不开始表示对准的块中的单个字之间的任何特定关系。因此，在数据库中并不意味字“Yes”与“Oui”互为译文，只是它们两者碰巧作为第一个字出现在对准的语料库中的对应块中。

图4与5示出统计分析程序208 的输出，在本实施例中它们生成下述频率表。表FREQE（图4）为英文语料库CORPE的字频表。表FREQE中一项的索引为来自英文语料库中的一个单字。而在这一字下存储的项则是该字在语料库中出现的次数。已知有若干种传统的程序设计语言提供这种所谓的“关联寻址”。这些语言中包括Lisp、POP-11、PERL、AWK。 当然，在不提供关联寻址的环境中，可以由系统设计员明确地实现。

第二字频表FREQF中包含法文语料库CORPF的字频。这些表在本实施例中并不是大小写敏感的，因此“Yes”与“yes”作为同一个字对待。这里用“大括号”即波形括号{}表示关联寻址。

第三表PAIRFREQ（图5）存储对准语料库的字对频率。这是一张概念上的二维表，其各项可用一个字对关联寻址：一个字来自英文语

料库CORPE 而另一个字来自法文语料库CORPF 。对于一个给定的字对，诸如“good”与“bonjour”，表项PAIRFREQ{good，boujour}存储这两个字出现在对准的语料库的对应块中的次数。

图 5 中加影线的框表示对应于图 3 的示例语料库中所示的少数字的项。由于这两个字出现在对准的块CORPE[1]与CORPF[1]的第一对中，因此项PAIRFREQ{good,bonjour}中包含一个至少为 1 的值。类似地，由于这一字对也出现在第一对对准的块中，因此PAIRFREQ{day，bonjour}中包括至少为 1 的一个值。

对于各语料库，存储了字的总数，它等于本例中表FREQE或FREQF的所有项的和。类似地，记录了所有字对的总数，它自然是二维对频率表PAIRFREQ中的所有项之和。

熟悉本技术的人员会理解，为了将表PAIRFREQ真正地实现为一个二维表会得出一个非常稀疏的数组。可采用更有效的实现方法，其中该表为相似于数组FREQE与FREQF的一个一维关联数组。这可以通过将一对中的字连接成一个单一的串来检索该表而容易地做到。因而，例如不将“good”与“bonjour”作为一个二维地址{good，bonjonr}的独立成分，而将整个串“good－bonjour”作为用于关联寻址表PAIRFREQ的对应项的一个单一的一维项对待。

图 6 示意性地示出在统计，分析程序208 从对准语料库CORPE与CORPF中生成统计数据库210时的操作的流程图。在一个初始化步骤600中，为对频率表PAIRFREQ保留空间，并归零其所有项。类似地为字频表FREQE与FREQF保留空间，并且也将它们的项归零。建立一个对计数变量PAIRTOTAL 并将其设置为 0，同样建立和归零字计数变量ETOTAL与FTOTAL。

进程的其余部分包括一系列嵌套的循环。为每一对对准的块CORPE[1]与CORPF[1]执行一次主循环602，其中I每通过该循环一次便从1向上增加1，直到考虑了每一对对准的块为止。在主循环 602内一个深层循环604为英文语料库的当前块内的各字CORPE[I][J]执行一次。在循环604内，一个更深层的循环606 为法文语料库CORPF中的对应块的各字CORPF[I][K]执行一次。在这一内循环606中，将对应于对频率表PAIRFREQ中的当前字对的一个项增加1。如上面所指出的，数组PAIRFREQ是可以通过参照一个英文－法文字对关联地寻址的。还在循环606 内，将计数器变量PAIRTOTAL增加1。

在循环604与606之外，但仍在主循环602 之内，为英文语料库的当前块中每一个字CORPE[I][J]执行一次另一个循环608。将字频表FREQE中的一个项增加1，并同时将英文语料库的总的字计数ETOTAL增加1，完成循环608之后，为法文语料库内的每一个字CORPF[I][K]执行一次主循环602内的另一个循环610。在这一循环中，将字频表FREQF 中的项增加1来记录字 CORPF[I][K]的出现，并同时将法文语料库的总的字计数FTOTAL增加1

因此，一旦为对准的语料库CORPE与CORPF中每一对对准的块执行过主循环602 之后，表PAIRFREQ中包含对准的块中每一个唯一的字对的出现次数的记录，表FREQE记录了英文语料库中各唯一的字的出现次数，而表FREQF则记录了法文语料库中各唯一的字的出现次数。同时，字对的总数则记录在变量PAIRTOTAL中，英文语料库中的字的总数记录在变量 ETOTAL 中而法文语料库 CORPF 中的字的总数则记录在变量FTOTAL中。

图7与8示出图1中所示的估计模块212 的操作，具体地，图7

示出了各字对的相互关联测定值或“评分”的计算，而图 8 则示出使用包含在块中的字对的对评分，计算对准的各块对的评分。

在图 7 中，操作从步骤700 开始，在其中接收到一对字WORDE与WORDF。在步骤702中，使用统计数据库210的表计算三个概率值。通过将记录在对频表(图 5)中的一个字对频率PAIRFREQ{WORDE,WORDF}除以该表中所记录的字对的总数PAIRTOTAL为该字对计算一个对概率值PAIRPROB。从而，PAIRPROB测定所接收的字对出现在数据库的对准的语料库中任何两个对准的块中的观测到的概率。

通过将英文语料库中单独的英文字WORDE的出现频率除以英文语料库中的总字数而计算一个值EPROB。这便是，将表项FREQE{WORDE}除以值ETOTAL。从而，值EPROB测定接收的对的英文字单独出现在英文语料库中的概率。类似地通过将表项FREQF{WORDF}除以法文语料库中的字的总数FTOTAL，而为所接收的字对的法文字WORDF计算概率值FPROB。值FPROB根据法文语料库 CORPF的内容测定字 WORDF的出现概率。

最终，在步骤704中，通过将对概率值PAIRPROB除以各个字概率值 EPROB与 FPROB而为所接收字对 WORDE，WORDF 计算一个估计值 PAIRSCORE。等于 1 的值PAIRSCORE表示这一字对出现在对准的块中的频率不大于根据单个的字出现在它们各自的语料库中观测概率的随机概率所期望的频率。反之，大于 1 的值PAIRSCORE表示这对字的出现频率大于从单个字频所期望的。从而，对评分是对中的两个字之间的相互关系的一种度量。

参见图 8，统计数据库与图 7 的对评分方法可用于测定各包括对应的语言（诸如英文与法文）中的一个或多个字的两块正文的对准质

量，图 8 的操作从步骤800开始，其中为模块212接收两块正文CHUNKE与CHUNKF以估计它们的对准评分。在初始化步骤802 中，将评分变量S设置成 1，并将计数变量N设置成 0。

一旦为S与N建立了初始值，这些变量便由一对嵌套的循环804与806修改。外循环804为在步骤800 中所接收的正文的英语块中的每一个字CHUNKE[J]执行一次。内循环806为在步骤800 中所接收的法语块中的第一个字CHUNKF[K]执行一次。因此，在嵌套的循环806与804 内，将英文块中的第一个字与法文块中的每一个字一起考虑。将值S乘以各个考虑的字对的对评分PAIRSCORE（CHUNKE[J]，CHUNKF[K]）。这一对评分是用图 7 的步骤计算的。此时，将计数变量N增加 1，以维护组合在变量S中的评分的数目的计数。

考虑了两个接收块中的各字对并将其对评分组合进乘积S之后，在步骤808中通过求乘积S的N次根计算对准评分ALSCORE。用数学语言讲，对准评分ALSCORE为所接收的块中的全体字对的对评分PAIRSCORE的“几何平均”值。从而，一对正文块的对准评分ALSCORE是组合这两块中的所有可能的字对的对评分的一个“似然”测定。由于在步骤808中求出了几何平均值，便用类似于对评分的方法正规化值ALSCORE，使得为 1 的值ALSCORE，根据记录在统计数据库210中的字频与字对频率，表示在800中接收的两块只是从单个字概率所期望的那样可能对应。反之，大于 1 的对准评分提示在两块的字之间存在着平均上大于随机概率与观测到的单独字频所提示的相互关联程度。

乘积S可能达到很大的值，并且在一个自动处理装置中计算许多乘法与除法通常是烦琐的。在实践中，对于步骤808中的几何平均值的计算，用对评分的对数的算术平均值来计算可能是有利的。可用对

数的加减来实现乘除。可将S的对数除以N来计算S的N次根的对数。

应当注意，虽然统计数据库210的频率表与计数值是从源语料库200与202以及表示它们之间的对准中推导出的，但在步骤700中接收的字对及在步骤800 中接收的块对可以从它们本身的对准语料库，或者从正在估计其对准的任何正文对中推算出。从而，所接收的块CHUNKE可以是诸如一位翻译所接收的CHUNKF的翻译人员的成果，并用对准评分ALSCORE将其与现存的对准语料库CORPE与CORPF的统计数据库相比较。大于1的一个值表示翻译人员广泛地与现存的对准语料库相符，而一个比1小得多的评分则表示不符合，例如，由于语料库的对准中的错误，翻译人员的错误、或者只是在现存的数据库与翻译人员头脑中所考虑的主题的两种领域中的差别。存在着估计模块212 的许多应用，图6、7与8的技术将作为下述实例示出。

图9示出利用估计模块212 来改进或“过滤”现存的数据库，即对准的语料库206 与统计数据库210的一种方法。这一进程表示一种反锁方式，并用图2中的虚线箭头230与232指示。在步骤900中，为了与对准评分值进行比较而设置一个阈值，有许多种选择该阈值的方法，如下所述。对于现在的描述，简单地将阈值设定为1便足够了，但是一般地说，最佳的阈值是取决于实际数据的，进程接着执行一个循环902，对于在法文语料库中存在着一个对准的块CORPF[I]的英文语料库中的每一个块CORPE[I]执行一次。在循环902中，步骤904读取对准的块CORPF[I]，并在步骤904中利用图8的过程为当前的块对估计对准评分ALSCORE(CORPE[I],CORPF[I]。

在步骤908中，将这一对准评分每步骤900中设定的阈值进行比较。如果这一对准评分超过阈值，控制便进行到912，在其中为下一个I

值执行循环902，即对准的语料库206中的下一对对准的块。如果在步骤908中，对准评分低于阈值，控制进行到步骤910，并从对准的语料库中删除对准的块的当前对，或者至少作上可疑的标记供以后删除。后一种选择在一种给定的实现中可能是方便的，并且在采取最后的决定之前，可允许与一位翻译人员对话。

完成步骤910之后，控制再一次进入点912，在其中将 I 增加 1 并为下一对块执行循环902。当在循环902中考虑过对准的语料库206中所有的对准块之后，控制进入步骤914，在其中计入步骤910中执行的删除（如果有的话）来更新统计数据库210。

注意，图 9 的方法可以在数据库上重复任意次数，逐步过滤掉存在在某些快上的不精确对准中的“噪声”。噪声源是多种多样的，但通常是实现在模块204中用来生成对准的语料库206的自动进程对它正在处理的语言缺少知识，以及在选择正文中哪些块该对准时不注意语法与语义这些事实的后果。并且，即使对于正确地对准的块，原始译文也不总是严恪的译文，并且毫无疑问，即使将一个短的句子翻译成一种给定的语言，也有若干种译法。

在从诸如照相复印机与传真机等电子设备的操作手册中导出的语料库的情况中，通常存在着完全不对应的部分，这是因为不同国家中的法律要求提供不同的安全信息。另一种常见的噪声源在于各 语料库的一部分是按字母表次序排列的项目的一张表时。在两种不同的语言中，这些项目的次序将不相同，即使项目的数目及其总的外观对于对准模块204 可能是难区分的。

然而，假定这些问题局限于源文件的相对地小的部分，则已经发现统计数据库是仍然有用的，并且由估计模块212 生成的对准评分将

会成功地标识出有问题的区。

除了过滤对准的语料库之外，在利用对准的语料库翻译一个新的正文时，也能使用一对块的对准评分，如图 1 0 中所示。

在图10中步骤1000上接收到要由翻译模块214 从英文译成法文的一个新的正文ETEXT （图 2 中216）。步骤1002中标识英文正文的第一块ECHUNK，并在步骤1004中搜索现存的英文语料库CORPE是否出现这一块。如果发现对于某一值 I，英文对准的语料库中的块CORPE[I]等于接收的块ECHUNK，控制进入在步骤1006，读取法文语料库中的对应块CORPF[I]。在步骤1010中，保存该块作为所需要的法文译文（输出正文218）的一个对应的块FCHUNK。

如果在步骤1004中未找到当前块ECHUNK的等价物，则控制进入一个用户对话步骤1008。在此要求一位翻译人员提供英文块ECHUNK的译文，并在步骤1010中作为译文FCHUNK保存之。在步骤1012中，在接收的正文ETEXT中标识下一个英语块，并将控制返回到搜索步骤1004。当翻译完整个输入正文ETEXT时，便在步骤1014中将在步骤1010中保存的所有块FCHUNK连接在一起作为经过翻译的法语正文FTEXT输出。

注意，用户在步骤1008中提供的译文也可用来扩大现存的数据库，这是通过将不熟悉的英语块CHUNKE与用户提供的法文译文加入对准的语料库206 中而实现的。这种性能在诸如US 5,140,522中有所描述，在这一阶段还可以更新统计数据库210，同时注意，作为在执行图10的方法中的“现场”用户对话的一种替代，也可简单地在输出文件 FTEXT 中加入某些问题，供翻译人员以后去考虑。此外，在保存翻译后的块FCHUNK的步骤1010中可包含诸如为块ECHUNK与FCHUNK估计一个对准评分的一个步骤，以便确认这的确是一个可能的译文。如果对准

评分下降到一个预定的阈值以下，可以“现场”进入用户对话或者在输出文件FTEXT 中加入适当的问题。这可起到校正未被图 9 中的过滤进程消除的错误对准的作用。

熟悉本技术的读者能够在上述实施例上发现许多变型，以及本公开中所建议的分析与估计装置的许多其它应用。

适用于大型数据库（尤其是块中包括相对大数量的字时）的一种变型通过考虑比所有可能的字对少的字对而限制处理工作量。这可用若干方法做到，但一种简单的步骤便是限制图 6 的流程图中最内层循环606 的范围K，例如对于某一整数d，使K从J－d变化到J＋d。

然后，不是计数每一个与当前英文字CORPE[I][J]成对的法文块的字，而是只考虑与计数字CORPF[I][K]的一个有限的“窗口”。当当前的英文字下标J 随每次外循环604 的重复而前进时，法文字的“窗口”也随之前进。当然这一实现最适用于典型句子中的字的次序服从类似规则的一对语言。对于对准部分中的字数明显地不同的语言，可以用适当速率来安排窗口（K值的范围）的前进，使其相对于K的最大值的位置粗略地与相对于J的最大值的J值匹配。

另一种减少所考虑的对的数目的技术为省略诸如“the”、“and”等极为常用的字。低频度的字假定为携带较大量的信息。作为一个例子，英文句子：“The man killed a big dog”可缩减为“man killed big dog”而损失很少的含义。

为了在图 6 的流程图中实现这一变化，在字对频率表（PAIRFREQ，步骤602、604、606）之前生成两个语料库的字频表（FREQE、FREQF步骤602、608、610）是较为方便的。然后，便能用这两个字频表来标识要在生成对频率表中省略的最常用的字。作为替代，可以利用对

相同的语料库或对整个相关语言的事先存在的字频表。

如果愿意，可以将这两种（及其它）变化组合在一起。对应的技术可实现在图 8 的内循环806 中，以减少组合字对评分去得到一对正文块的对准评分的工作量。

下面用对附录的讨论来结束本说明，在部分 A 至 G 中提出了两个相对地小的对准的语料库的一个实例，并且其估计是由上述系统执行的。附录中的语料库包括一种传真装置的操作手册的内容清单，第一是英语的，而第二则是荷兰语的。这两个语料库分别提出在附录的 A 与 B 中，行号 1 至 3 0 表示两个语料库中对准的块对 1 至 3 0 。在行 3 0 中出现一个不正确的对准，其中“sending documents”不是荷兰短语“problemen oplossen”的英文译文。通常是这样的，对准语料库中别处包含字“problemen”与“oplossen”的全部都是正确对准的，即“troubleshooting”（块对 2 3 与 2 7 ）。

在附录的 C 中，提供了英文语料库的字频。可以看出，例如字“sending”出现 7 次而字“confidential”只出现一次。英文语料库中的总字数为118。因此，英文语料库中字“sending”的出现概率为 7 除以118或0.059322。

在附录的 D 中，提供了荷兰文语料库的字频表。在荷兰文语料库中总共有106 个字。从而，诸如字“problemen”，具有106个字中 4 个的频率，即观测概率0.037736. 注意，在示例系统中，统计数据库不是对大小写敏感的：这便是，在语料库的块23中的“problemen”与块 6 中的“problemen”之间没有差别。

在附录中 E － 1 至 E － 4 处，提供了对准的语料库的字对频率表，其中总共有427个唯一的字对。所有字对频率的总数为510。注意，在

正常情况中采用较大词汇表的语料库中，出现的字对的数目会惊人地上升。

附录D的字对频率表 示出的例子如有4对块，其中英文块中包含字“part”而对应的荷兰文块中则包含字“en”。对两个语料库（附录中的A与B）的一次快速检视发现这一对字出现在对准的块2、3、5与6中。但是注意，数据库并不表示字“part”与“en”互为译文。这两个字只是偶然同时出现在它们各自的语料库中，因此，存在着在任何一对块中的纯粹随机出现两个字的合理的概率。

在附录中F－1至F－4处，为该示例性语料库中的427个不同的字对计算与示出了测定两个字之间的相互关联的字对评分。而在附录的部分D中，字对是按对频率的次序排列的，在部分E中。它们是按对评分的降序排列的。与频率表进行比较，便可注意到对于实际上互为译文的字有十分大的趋势得到高的评分。评分从24.525490下降到0.383211。然而，并无利用来自附录的部分A的单个字对的评分来检验任何逐字翻译的精确性的可能性。

反之，在附录中的G－1至G－2处，独立地提供了对准的块以及各块的对准评分。这些块对准评分是以图8的方法通过累计各块对中的所有字对的对评分而得到的。块1具有10.071629 的评分，表示与整个数据库的统计相比，看起来这两个对准的块作为互为译文是真正有用的。能够看出，所有的对准的块对，除了最后一个以外，都具有显著地超过1的评分。反之，读者已知其为错误的块对号30只有0.819339的评分。从而，即使在包含至少一个不正确对准的块对的这一非常小的数据库中，这里所提出的估计技术与装置也已提供了对不正确译文的清楚的突出指示。

图 1 1 为用图形示出该示例性语料库的 3 0 个块的对准评分的分布的直方图。垂直轴标出频率，而水平轴则为了方便而标出对准评分的对数（以 2 为底）。例如，对准评分的对数中的 3 至 4 的范围（水平轴）对应于对准评分本身 8 至 1 6 的范围。垂直虚线1100表示以对准评分的对数表示的阈值 0，它对应于上面提到的对准评分本身的阈值 1（log 1＝0）。在阈值线1100的右方，频率分布的主体清楚地与阈值左方的一个较小的峰值1102分开，这一峰值标示块对号 3 0 的低对准评分。

熟悉本技术的人会理解，取决于统计数据库的内容，其它的阈值可能是理想的甚至是必要的。在许多情况中，如在本实例中，有可能区分出对准语料库中的一个对准评分的分布主体及起因于错误的对准的一个次要分布。如果这两种密度明显地分离，如在本例中那样，在两者之间设置阈值便是一件简单的工作。

在其它情况中，可能有必要采用更精细的方法来设定阈值。这种方法之一为设定一个百分比的阈值，例如，通过选择最少可能性的 5 个百分点的对准加以拒绝。然后，相应地设定对准评分阈值，或者可以简单地蕴含在删除最坏的 5 个百分点的对准的操作中。

在某些情况中。甚至希望完全不设置一个硬性的阈值，而采用与一位翻译人员的对话来决定哪些对准是正确的。然后，系统通过向翻译人员提出从具有最底的对准评分的块开始的对准的块对来进行操作。通常，提出的第一对将是容易地作为错误的而加以拒绝的。然后，当提出的块的对准评分达到较高的值时，便开始向翻译人员提出虽然正确但碰巧具有相对地低的对准评分的对（例如附录中 G 处的例子中的对号 2 ）。

在这一过程中继续下去，提出给用户的大多数对将是正确的，而这是系统设计员与/或操作员的一种选择，在哪一点上作出切割，并认为其余的对准是正确的。在任何情况下，利用本发明，实际上由一位翻译人员检验的数据库的比例已经减少到容易处理的一小部分，而检验整个数据库则是不现实地昂贵与费时的。

对于剩下的那些错误，使用修正后的语料库及一个更新后的统计数据库的另一次迭代可能比为少数剩下的错误而强制操作员在搜索中去检验多得多的对来消除它们更有效。此外，在消除了少数错误并更新了统计数据库之后，正确但起先评分低的对的对准评分，可在随后的迭代中改进，因为这时的统计数据库本身是一个较少噪声的数据库的产物。从而，在第二次迭代上，较少可能向翻译人员提出实际上正确地对准的对。

再者，上述实现上的许多变型是在熟悉本技术的人员的能力与想象力范围内的。例如，作为体现在图 6 与 7 中用于获得字对评分的方法的一种替代，可采用诸如EP－A－0499366所用的方法来得到类似的效果。取决于用来得到字对评分的实际方法，还可能需要适用的组合字对评分以得出对准的句子的评分的方法，例如上述涉及对数的方法。类似地，在统计学或词法知识的基础上，标识具有共同“词干”的字的预处理技术也可采用，如各种对比文献中所描述的。上述实例仅供例示之用。

［附录如下］

APPENDIX A THE ENGLISH CORPUS

1 Part 1 Before Starting
2 Part 2 Sending and Receiving Documents
3 Part 3 Using the Telephone and Copying Features
4 Part 4 Using the Memory and Network Features
5 Part 5 Reports and User Switches
6 Part 6 Maintenance and Troubleshooting
7 Installing Your FAX
8 A Look at the FAX-260E
9 Identifying the Documents You Send
10 Before Sending Documents
11 Sending Documents
12 Receiving Documents
13 Different Ways of Dialling
14 Using the Telephone with the FAX-260E
15 Sending at a Preset Time
16 Sending through a Relay Unit
17 Sending Confidential Documents
18 Polling (Requesting documents from other units)
19 Printing Reports and Registration Lists
20 Setting the Operating Guidelines
21 Caring for Your Fax
22 Error Messages and Codes
23 Troubleshooting
24 Specifications
25 Index
26 Error Messages and Codes
27 Troubleshooting
28 Index
29 Setting the Operating Guidelines
30 Sending Documents

APPENDIX B THE DUTCH CORPUS

1 Deel 1 Voordat u begint
2 Deel 2 Verzenden en ontvangen
3 Deel 3 De FAX-260E gebruiken als telefoonkiezer en copier
4 Deel 4 FAX-functies
5 Deel 5 Rapporten en gebruikersschakelaars
6 Deel 6 Onderhoud en problemen oplossen
7 Installatie van uw FAX-260E
8 De onderdelen van uw FAX-260E
9 Identificatie van uw verzonden documenten
10 Originelen
11 Verzenden
12 Ontvangen
13 Snel en eenvoudig kiezen
14 Gebruik van de FAX-260E als telefoonkiezer
15 Verzenden op ingesteld tijdstip
16 Verzenden via transit fax-apparaat
17 Vertrouweijk verzenden
18 Polling (op verzoek documenten van andere fax-apparaten ontvangen)
19 Afdrukken van rapporten en lijsten
20 Instellen van gebruikersschakelaars
21 Onderhoud
22 Foutmeldingen en codes
23 Problemen oplossen
24 Technische gegevens
25 Trefwoordenlijst
26 Foutmeldingen en codes
27 Problemen oplossen
28 Trefwoordenlijst
29 Vastleggen van gebruikersinstellingen
30 Problemen oplossen

APPENDIX C ENGLISH WORD FREQUENCIES

8	and
8	documents (TOTAL=118)
8	the
7	sending
6	part
3	a
3	troubleshooting
3	using
2	at
2	before
2	codes
2	error
2	fax
2	fax-260e
2	features
2	guidelines
2	index
2	messages
2	operating
2	receiving
2	reports
2	setting
2	telephone
2	your
1	1
1	2
1	3
1	4
1	5
1	6
1	caring
1	confidential
1	copying
1	dialling
1	different
1	for
1	from
1	identifying
1	installing
1	lists
1	look
1	maintenance
1	memory
1	network
1	of
1	other
1	polling
1	preset
1	printing
1	registration
1	relay
1	requesting
1	send
1	specifications
1	starting
1	switches
1	through
1	time
1	unit
1	units
1	user
1	ways
1	with
1	you

APPENDIX D DUTCH WORD FREQUENCIES

8	en
8	van
6	deel (TOTAL=106)
5	verzenden
4	fax-260e
4	oplossen
4	problemen
3	de
3	ontvangen
3	uw
2	als
2	codes
2	documenten
2	foutmeldingen
2	gebruikersschakelaars
2	onderhoud
2	op
2	rapporten
2	telefoonkiezer
2	trefwoordenlijst
1	1
1	2
1	3
1	4
1	5
1	6
1	afdrukken
1	andere
1	begint
1	copier
1	eenvoudig
1	fax-apparaat
1	fax-apparaten
1	fax-functies
1	gebruik
1	gebruiken
1	gebruikersinstellingen
1	gegevens
1	identificatie
1	ingesteld
1	installatie
1	instellen
1	kiezen
1	lijsten
1	onderdelen
1	originelen
1	polling
1	snel
1	technische
1	tijdstip
1	transit
1	u
1	vastleggen
1	vertrouweijk
1	verzoek
1	verzonden
1	via
1	voordat

APPENDIX E - WORD PAIR FREQUENCIES

(TOTAL=510)

7	and en
6	part deel
6	the van
5	and deel
5	sending verzenden
4	part en
4	the de
4	the fax-260e
3	documents ontvangen
3	documents verzenden
3	the als
3	the telefoonkiezer
3	troubleshooting oplossen
3	troubleshooting problemen
2	a verzenden
2	and codes
2	and foutmeldingen
2	and rapporten
2	codes codes
2	codes en
2	codes foutmeldingen
2	documents documenten
2	documents van
2	error codes
2	error en
2	error foutmeldingen
2	fax-260e de
2	fax-260e fax-260e
2	fax-260e van
2	features deel
2	guidelines van
2	index trefwoordenlijst
2	messages codes
2	messages en
2	messages foutmeldingen
2	operating van
2	receiving ontvangen
2	reports en
2	reports rapporten
2	setting van
2	telephone als
2	telephone de
2	telephone fax-260e
2	telephone telefoonkiezer
2	the deel
2	the gebruik
2	the uw
2	using als
2	using de
2	using deel
2	using fax-260e
2	using telefoonkiezer
1	1 1
1	1 begint
1	1 deel

1 1 u
1 1 voordat
1 2 2
1 2 deel
1 2 en
1 2 ontvangen
1 2 verzenden
1 3 3
1 3 als
1 3 copier
1 3 de
1 3 deel
1 3 en
1 3 fax-260e
1 3 gebruiken
1 3 telefoonkiezer
1 4 4
1 4 deel
1 4 fax-functies
1 5 5
1 5 deel
1 5 en
1 5 gebruikersschakelaars
1 5 rapporten
1 6 6
1 6 deel
1 6 en
1 6 onderhoud
1 6 oplossen
1 6 problemen
1 a de
1 a fax-260e
1 a fax-apparaat
1 a ingesteld
1 a onderdelen
1 a op
1 a tijdstip
1 a transit
1 a uw
1 a van
1 a via
1 and 2
1 and 3
1 and 4
1 and 5
1 and 6
1 and afdrukken
1 and als
1 and copier
1 and de
1 and fax-260e
1 and fax-functies
1 and gebruiken
1 and gebruikersschakelaars
1 and lijsten

1 and onderhoud
1 and ontvangen
1 and oplossen
1 and problemen
1 and telefoonkiezer
1 and van
1 and verzenden
1 at de
1 at fax-260e
1 at ingesteld
1 at onderdelen
1 at op
1 at tijdstip
1 at uw
1 at van
1 at verzenden
1 before 1
1 before begint
1 before deel
1 before originelen
1 before u
1 before voordat
1 caring onderhoud
1 confidential vertrouweijk
1 confidential verzenden
1 copying 3
1 copying als
1 copying copier
1 copying de
1 copying deel
1 copying en
1 copying fax-260e
1 copying gebruiken
1 copying telefoonkiezer
1 dialling eenvoudig
1 dialling en
1 dialling kiezen
1 dialling snel
1 different eenvoudig
1 different en
1 different kiezen
1 different snel
1 documents 2
1 documents andere
1 documents deel
1 documents en
1 documents fax-apparaten
1 documents identificatie
1 documents op
1 documents oplossen
1 documents originelen
1 documents polling
1 documents problemen
1 documents uw
1 documents vertrouweijk

1 documents verzoek
1 documents verzonden
1 fax fax-260e
1 fax installatie
1 fax onderhoud
1 fax uw
1 fax van
1 fax-260e als
1 fax-260e gebruik
1 fax-260e onderdelen
1 fax-260e telefoonkiezer
1 fax-260e uw
1 features 3
1 features 4
1 features als
1 features copier
1 features de
1 features en
1 features fax-260e
1 features fax-functies
1 features gebruiken
1 features telefoonkiezer
1 for onderhoud
1 from andere
1 from documenten
1 from fax-apparaten
1 from ontvangen
1 from op
1 from polling
1 from van
1 from verzoek
1 guidelines gebruikersinstellingen
1 guidelines gebruikersschakelaars
1 guidelines instellen
1 guidelines vastleggen
1 identifying documenten
1 identifying identificatie
1 identifying uw
1 identifying van
1 identifying verzonden
1 installing fax-260e
1 installing installatie
1 installing uw
1 installing van
1 lists afdrukken
1 lists en
1 lists lijsten
1 lists rapporten
1 lists van
1 look de
1 look fax-260e
1 look onderdelen
1 look uw
1 look van
1 maintenance 6

1 maintenance deel
1 maintenance en
1 maintenance onderhoud
1 maintenance oplossen
1 maintenance problemen
1 memory 4
1 memory deel
1 memory fax-functies
1 network 4
1 network deel
1 network fax-functies
1 of eenvoudig
1 of en
1 of kiezen
1 of snel
1 operating gebruikersinstellingen
1 operating gebruikersschakelaars
1 operating instellen
1 operating vastleggen
1 other andere
1 other documenten
1 other fax-apparaten
1 other ontvangen
1 other op
1 other polling
1 other van
1 other verzoek
1 part 1
1 part 2
1 part 3
1 part 4
1 part 5
1 part 6
1 part als
1 part begint
1 part copier
1 part de
1 part fax-260e
1 part fax-functies
1 part gebruiken
1 part gebruikersschakelaars
1 part onderhoud
1 part ontvangen
1 part oplossen
1 part problemen
1 part rapporten
1 part telefoonkiezer
1 part u
1 part verzenden
1 part voordat
1 polling andere
1 polling documenten
1 polling fax-apparaten
1 polling ontvangen
1 polling op

1 polling polling
1 polling van
1 polling verzoek
1 preset ingesteld
1 preset op
1 preset tijdstip
1 preset verzenden
1 printing afdrukken
1 printing en
1 printing lijsten
1 printing rapporten
1 printing van
1 receiving 2
1 receiving deel
1 receiving en
1 receiving verzenden
1 registration afdrukken
1 registration en
1 registration lijsten
1 registration rapporten
1 registration van
1 relay fax-apparaat
1 relay transit
1 relay verzenden
1 relay via
1 reports 5
1 reports afdrukken
1 reports deel
1 reports gebruikersschakelaars
1 reports lijsten
1 reports van
1 requesting andere
1 requesting documenten
1 requesting fax-apparaten
1 requesting ontvangen
1 requesting op
1 requesting polling
1 requesting van
1 requesting verzoek
1 send documenten
1 send identificatie
1 send uw
1 send van
1 send verzonden
1 sending 2
1 sending deel
1 sending en
1 sending fax-apparaat
1 sending ingesteld
1 sending ontvangen
1 sending op
1 sending oplossen
1 sending originelen
1 sending problemen
1 sending tijdstip

1 sending transit
1 sending vertrouweijk
1 sending via
1 setting gebruikersinstellingen
1 setting gebruikersschakelaars
1 setting instellen
1 setting vastleggen
1 specifications gegevens
1 specifications technische
1 starting 1
1 starting begint
1 starting deel
1 starting u
1 starting voordat
1 switches 5
1 switches deel
1 switches en
1 switches gebruikersschakelaars
1 switches rapporten
1 telephone 3
1 telephone copier
1 telephone deel
1 telephone en
1 telephone gebruik
1 telephone gebruiken
1 telephone van
1 the 3
1 the 4
1 the copier
1 the documenten
1 the en
1 the fax-functies
1 the gebruiken
1 the gebruikersinstellingen
1 the gebruikersschakelaars
1 the identificatie
1 the instellen
1 the onderdelen
1 the vastleggen
1 the verzonden
1 through fax-apparaat
1 through transit
1 through verzenden
1 through via
1 time ingesteld
1 time op
1 time tijdstip
1 time verzenden
1 troubleshooting 6
1 troubleshooting deel
1 troubleshooting en
1 troubleshooting onderhoud
1 unit fax-apparaat
1 unit transit
1 unit verzenden

1	unit via
1	units andere
1	units documenten
1	units fax-apparaten
1	units ontvangen
1	units op
1	units polling
1	units van
1	units verzoek
1	user 5
1	user deel
1	user en
1	user gebruikersschakelaars
1	user rapporten
1	using 3
1	using 4
1	using copier
1	using en
1	using fax-functies
1	using gebruik
1	using gebruiken
1	using van
1	ways eenvoudig
1	ways en
1	ways kiezen
1	ways snel
1	with als
1	with de
1	with fax-260e
1	with gebruik
1	with telefoonkiezer
1	with van
1	you documenten
1	you identificatie
1	you uw
1	you van
1	you verzonden
1	your fax-260e
1	your installatie
1	your onderhoud
1	your uw
1	your van

APPENDIX F - WORD PAIR CORRELATION SCORES

24.525490	1 1
24.525490	1 begint
24.525490	1 u
24.525490	1 voordat
24.525490	2 2
24.525490	3 3
24.525490	3 copier
24.525490	3 gebruiken
24.525490	4 4
24.525490	4 fax-functies
24.525490	5 5
24.525490	6 6
24.525490	confidential vertrouweijk
24.525490	copying 3
24.525490	copying copier
24.525490	copying gebruiken
24.525490	dialling eenvoudig
24.525490	dialling kiezen
24.525490	dialling snel
24.525490	different eenvoudig
24.525490	different kiezen
24.525490	different snel
24.525490	from andere
24.525490	from fax-apparaten
24.525490	from polling
24.525490	from verzoek
24.525490	identifying identificatie
24.525490	identifying verzonden
24.525490	installing installatie
24.525490	lists afdrukken
24.525490	lists lijsten
24.525490	look onderdelen
24.525490	maintenance 6
24.525490	memory 4
24.525490	memory fax-functies
24.525490	network 4
24.525490	network fax-functies
24.525490	of eenvoudig
24.525490	of kiezen
24.525490	of snel
24.525490	other andere
24.525490	other fax-apparaten
24.525490	other polling
24.525490	other verzoek
24.525490	polling andere
24.525490	polling fax-apparaten
24.525490	polling polling
24.525490	polling verzoek
24.525490	preset ingesteld
24.525490	preset tijdstip
24.525490	printing afdrukken
24.525490	printing lijsten
24.525490	registration afdrukken
24.525490	registration lijsten
24.525490	relay fax-apparaat

24.525490 relay transit
24.525490 relay via
24.525490 requesting andere
24.525490 requesting fax-apparaten
24.525490 requesting polling
24.525490 requesting verzoek
24.525490 send identificatie
24.525490 send verzonden
24.525490 specifications gegevens
24.525490 specifications technische
24.525490 starting 1
24.525490 starting begint
24.525490 starting u
24.525490 starting voordat
24.525490 switches 5
24.525490 through fax-apparaat
24.525490 through transit
24.525490 through via
24.525490 time ingesteld
24.525490 time tijdstip
24.525490 unit fax-apparaat
24.525490 unit transit
24.525490 unit via
24.525490 units andere
24.525490 units fax-apparaten
24.525490 units polling
24.525490 units verzoek
24.525490 user 5
24.525490 ways eenvoudig
24.525490 ways kiezen
24.525490 ways snel
24.525490 with gebruik
24.525490 you identificatie
24.525490 you verzonden
12.262745 3 als
12.262745 3 telefoonkiezer
12.262745 5 gebruikersschakelaars
12.262745 5 rapporten
12.262745 6 onderhoud
12.262745 at ingesteld
12.262745 at onderdelen
12.262745 at tijdstip
12.262745 before 1
12.262745 before begint
12.262745 before originelen
12.262745 before u
12.262745 before voordat
12.262745 caring onderhoud
12.262745 codes codes
12.262745 codes foutmeldingen
12.262745 copying als
12.262745 copying telefoonkiezer
12.262745 error codes
12.262745 error foutmeldingen
12.262745 fax installatie

12.262745	fax-260e gebruik
12.262745	fax-260e onderdelen
12.262745	features 3
12.262745	features 4
12.262745	features copier
12.262745	features fax-functies
12.262745	features gebruiken
12.262745	for onderhoud
12.262745	from documenten
12.262745	from op
12.262745	guidelines gebruikersinstellingen
12.262745	guidelines instellen
12.262745	guidelines vastleggen
12.262745	identifying documenten
12.262745	index trefwoordenlijst
12.262745	lists rapporten
12.262745	maintenance onderhoud
12.262745	messages codes
12.262745	messages foutmeldingen
12.262745	operating gebruikersinstellingen
12.262745	operating instellen
12.262745	operating vastleggen
12.262745	other documenten
12.262745	other op
12.262745	polling documenten
12.262745	polling op
12.262745	preset op
12.262745	printing rapporten
12.262745	receiving 2
12.262745	registration rapporten
12.262745	reports 5
12.262745	reports afdrukken
12.262745	reports lijsten
12.262745	reports rapporten
12.262745	requesting documenten
12.262745	requesting op
12.262745	send documenten
12.262745	setting gebruikersinstellingen
12.262745	setting instellen
12.262745	setting vastleggen
12.262745	switches gebruikersschakelaars
12.262745	switches rapporten
12.262745	telephone 3
12.262745	telephone als
12.262745	telephone copier
12.262745	telephone gebruik
12.262745	telephone gebruiken
12.262745	telephone telefoonkiezer
12.262745	time op
12.262745	units documenten
12.262745	units op
12.262745	user gebruikersschakelaars
12.262745	user rapporten
12.262745	with als
12.262745	with telefoonkiezer

12.262745	you documenten
12.262745	your installatie
8.175163	2 ontvangen
8.175163	3 de
8.175163	a fax-apparaat
8.175163	a ingesteld
8.175163	a onderdelen
8.175163	a tijdstip
8.175163	a transit
8.175163	a via
8.175163	copying de
8.175163	fax-260e de
8.175163	from ontvangen
8.175163	identifying uw
8.175163	installing uw
8.175163	look de
8.175163	look uw
8.175163	other ontvangen
8.175163	polling ontvangen
8.175163	receiving ontvangen
8.175163	requesting ontvangen
8.175163	send uw
8.175163	telephone de
8.175163	troubleshooting 6
8.175163	units ontvangen
8.175163	using 3
8.175163	using 4
8.175163	using als
8.175163	using copier
8.175163	using fax-functies
8.175163	using gebruik
8.175163	using gebruiken
8.175163	using telefoonkiezer
8.175163	with de
8.175163	you uw
6.131373	3 fax-260e
6.131373	6 oplossen
6.131373	6 problemen
6.131373	at op
6.131373	copying fax-260e
6.131373	fax onderhoud
6.131373	fax-260e als
6.131373	fax-260e fax-260e
6.131373	fax-260e telefoonkiezer
6.131373	features als
6.131373	features telefoonkiezer
6.131373	guidelines gebruikersschakelaars
6.131373	installing fax-260e
6.131373	look fax-260e
6.131373	maintenance oplossen
6.131373	maintenance problemen
6.131373	operating gebruikersschakelaars
6.131373	reports gebruikersschakelaars
6.131373	setting gebruikersschakelaars
6.131373	telephone fax-260e

6.131373	the gebruik
6.131373	troubleshooting oplossen
6.131373	troubleshooting problemen
6.131373	with fax-260e
6.131373	your onderhoud
5.450109	using de
4.905098	2 verzenden
4.905098	confidential verzenden
4.905098	preset verzenden
4.905098	relay verzenden
4.905098	through verzenden
4.905098	time verzenden
4.905098	unit verzenden
4.598529	the als
4.598529	the telefoonkiezer
4.087582	1 deel
4.087582	2 deel
4.087582	3 deel
4.087582	4 deel
4.087582	5 deel
4.087582	6 deel
4.087582	a op
4.087582	at de
4.087582	at uw
4.087582	copying deel
4.087582	fax uw
4.087582	fax-260e uw
4.087582	features de
4.087582	features deel
4.087582	maintenance deel
4.087582	memory deel
4.087582	network deel
4.087582	part 1
4.087582	part 2
4.087582	part 3
4.087582	part 4
4.087582	part 5
4.087582	part 6
4.087582	part begint
4.087582	part copier
4.087582	part deel
4.087582	part fax-functies
4.087582	part gebruiken
4.087582	part u
4.087582	part voordat
4.087582	starting deel
4.087582	switches deel
4.087582	the de
4.087582	troubleshooting onderhoud
4.087582	user deel
4.087582	using fax-260e
4.087582	your uw
3.503641	sending 2
3.503641	sending fax-apparaat
3.503641	sending ingesteld

3.503641	sending originelen
3.503641	sending tijdstip
3.503641	sending transit
3.503641	sending vertrouweijk
3.503641	sending verzenden
3.503641	sending via
3.270065	a verzenden
3.065686	2 en
3.065686	3 en
3.065686	5 en
3.065686	6 en
3.065686	and 2
3.065686	and 3
3.065686	and 4
3.065686	and 5
3.065686	and 6
3.065686	and afdrukken
3.065686	and codes
3.065686	and copier
3.065686	and fax-functies
3.065686	and foutmeldingen
3.065686	and gebruiken
3.065686	and lijsten
3.065686	and rapporten
3.065686	at fax-260e
3.065686	codes en
3.065686	copying en
3.065686	dialling en
3.065686	different en
3.065686	documents 2
3.065686	documents andere
3.065686	documents documenten
3.065686	documents fax-apparaten
3.065686	documents identificatie
3.065686	documents ontvangen
3.065686	documents originelen
3.065686	documents polling
3.065686	documents vertrouweijk
3.065686	documents verzoek
3.065686	documents verzonden
3.065686	error en
3.065686	fax fax-260e
3.065686	fax-260e van
3.065686	features fax-260e
3.065686	from van
3.065686	guidelines van
3.065686	identifying van
3.065686	installing van
3.065686	lists en
3.065686	lists van
3.065686	look van
3.065686	maintenance en
3.065686	messages en
3.065686	of en
3.065686	operating van

3.065686	other van
3.065686	polling van
3.065686	printing en
3.065686	printing van
3.065686	registration en
3.065686	registration van
3.065686	reports en
3.065686	requesting van
3.065686	send van
3.065686	setting van
3.065686	switches en
3.065686	the 3
3.065686	the 4
3.065686	the copier
3.065686	the fax-260e
3.065686	the fax-functies
3.065686	the gebruiken
3.065686	the gebruikersinstellingen
3.065686	the identificatie
3.065686	the instellen
3.065686	the onderdelen
3.065686	the vastleggen
3.065686	the verzonden
3.065686	units van
3.065686	user en
3.065686	ways en
3.065686	with van
3.065686	you van
3.065686	your fax-260e
2.725054	a de
2.725054	a uw
2.725054	using deel
2.682475	and en
2.554739	and deel
2.452549	at verzenden
2.452549	receiving verzenden
2.299265	the van
2.043791	a fax-260e
2.043791	before deel
2.043791	part als
2.043791	part en
2.043791	part gebruikersschakelaars
2.043791	part onderhoud
2.043791	part rapporten
2.043791	part telefoonkiezer
2.043791	receiving deel
2.043791	reports deel
2.043791	telephone deel
2.043791	the uw
1.839412	documents verzenden
1.751821	sending op
1.532843	and als
1.532843	and gebruikersschakelaars
1.532843	and onderhoud
1.532843	and telefoonkiezer

1.532843	at van
1.532843	documents op
1.532843	fax van
1.532843	features en
1.532843	receiving en
1.532843	reports van
1.532843	telephone en
1.532843	telephone van
1.532843	the documenten
1.532843	the gebruikersschakelaars
1.532843	your van
1.362527	part de
1.362527	part ontvangen
1.362527	troubleshooting deel
1.167880	sending ontvangen
1.021895	a van
1.021895	and de
1.021895	and ontvangen
1.021895	documents uw
1.021895	part fax-260e
1.021895	part oplossen
1.021895	part problemen
1.021895	the deel
1.021895	troubleshooting en
1.021895	using en
1.021895	using van
0.875910	sending oplossen
0.875910	sending problemen
0.817516	part verzenden
0.766422	and fax-260e
0.766422	and oplossen
0.766422	and problemen
0.766422	documents oplossen
0.766422	documents problemen
0.766422	documents van
0.613137	and verzenden
0.583940	sending deel
0.510948	documents deel
0.437955	sending en
0.383211	and van
0.383211	documents en
0.383211	the en

APPENDIX G-1 SCORES FOR ALIGNED CHUNKS

Part 1 Before Starting
Deel 1 Voordat u begint
score = 10.071629

Part 2 Sending and Receiving Documents
Deel 2 Verzenden en ontvangen
score = 2.285732

Part 3 Using the Telephone and Copying Features
Deel 3 De FAX-260E gebruiken als telefoonkiezer en copier
score = 4.727727

Part 4 Using the Memory and Network Features
Deel 4 FAX-functies
score = 6.443163

Part 5 Reports and User Switches
Deel 5 Rapporten en gebruikersschakelaars
score = 5.372271

Part 6 Maintenance and Troubleshooting
Deel 6 Onderhoud en problemen oplossen
score = 3.598853

Installing Your FAX
Installatie van uw FAX-260E
score = 4.935864

A Look at the FAX-260E
De onderdelen van uw FAX-260E
score = 4.253443

Identifying the Documents You Send
Identificatie van uw verzonden documenten
score = 5.746231

Before Sending Documents
Originelen
score = 5.087975

Sending Documents
Verzenden
score = 2.538629

Receiving Documents
Ontvangen
score = 5.006244

Different Ways of Dialling
Snel en eenvoudig kiezen
score = 14.582943

Using the Telephone with the FAX-260E
Gebruik van de FAX-260E als telefoonkiezer
score = 5.621435

Sending at a Preset Time
Verzenden op ingesteld tijdstip
score = 7.327703

Sending through a Relay Unit
Verzenden via transit fax-apparaat
score = 10.009936

APPENDIX G-2

Sending Confidential Documents
Vertrouweijk verzenden
score = 4.502135

Polling (Requesting documents from other units)
Polling (op verzoek documenten van andere fax-apparaten ontvangen)
score = 10.322900

Printing Reports and Registration Lists
Afdrukken van rapporten en lijsten
score = 6.270169

Setting the Operating Guidelines
Instellen van gebruikersschakelaars
score = 4.751194

Caring for Your Fax
Onderhoud
score = 8.671070

Error Messages and Codes
Foutmeldingen en codes
score = 6.063523

Troubleshooting
Problemen oplossen
score = 6.131373

Specifications
Technische gegevens
score = 24.525490

Index
Trefwoordenlijst
score = 12.262745

Error Messages and Codes
Foutmeldingen en codes
score = 6.063523

Troubleshooting
Problemen oplossen
score = 6.131373

Index
Trefwoordenlijst
score = 12.262745

Setting the Operating Guidelines
Vastleggen van gebruikersinstellingen
score = 5.986130

Sending Documents
Problemen oplossen
score = 0.819339

说　明　书　附　图

图 1

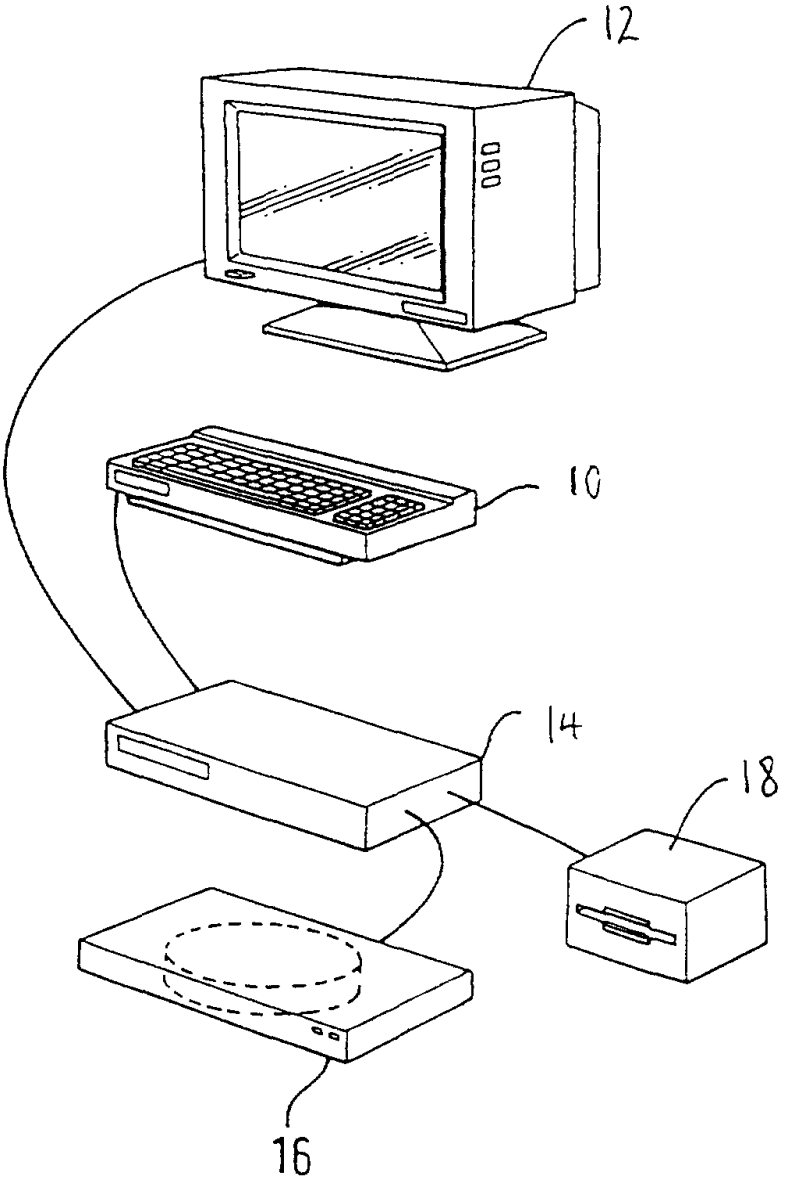
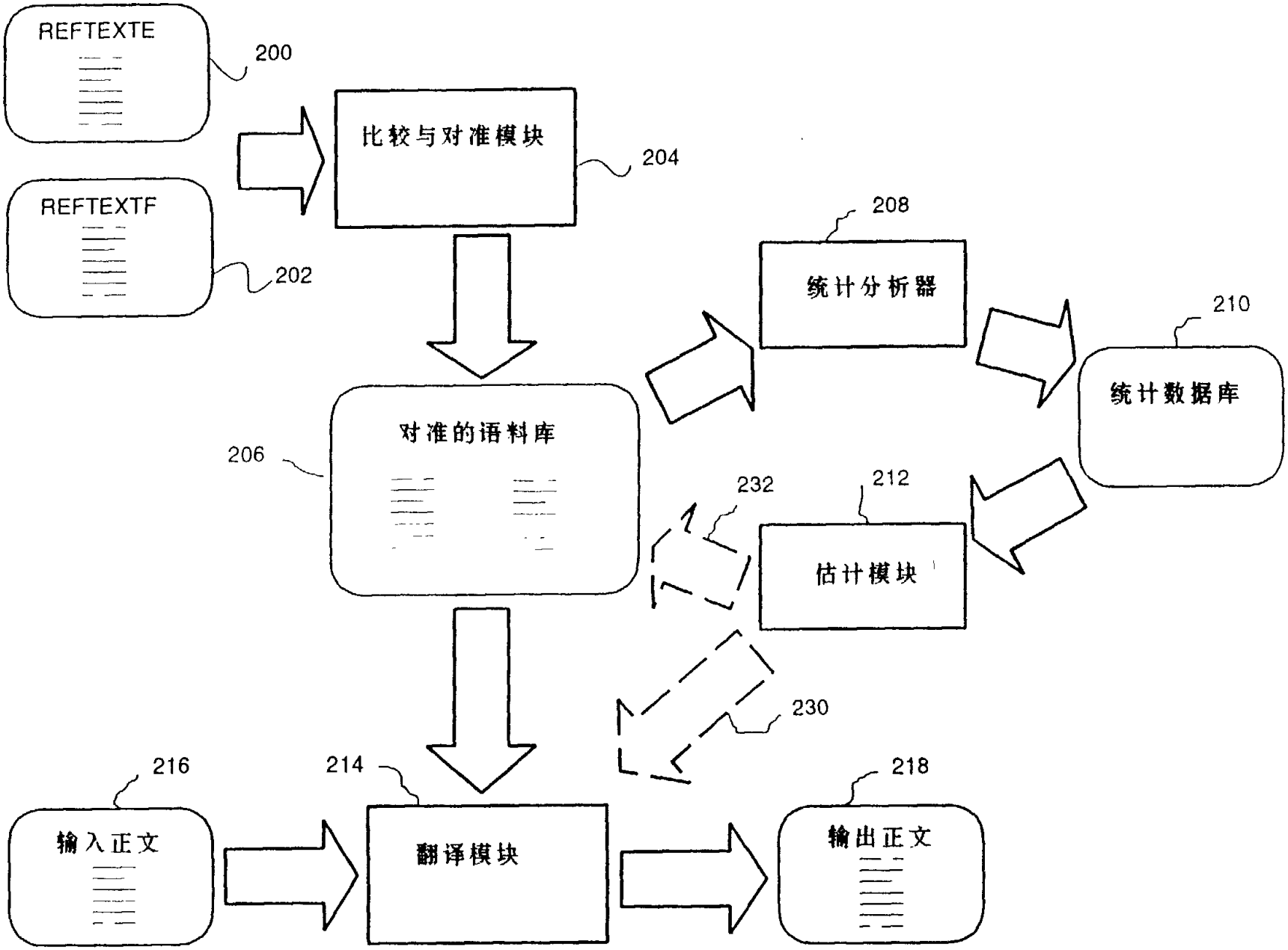


图 2


REFTEXTE
200
REFTEXTF
202
比较与对准模块
204
208
统计分析器
210
统计数据库
对准的语料库
206
232
212
估计模块
230
216
214
218
输入正文
翻译模块
输出正文

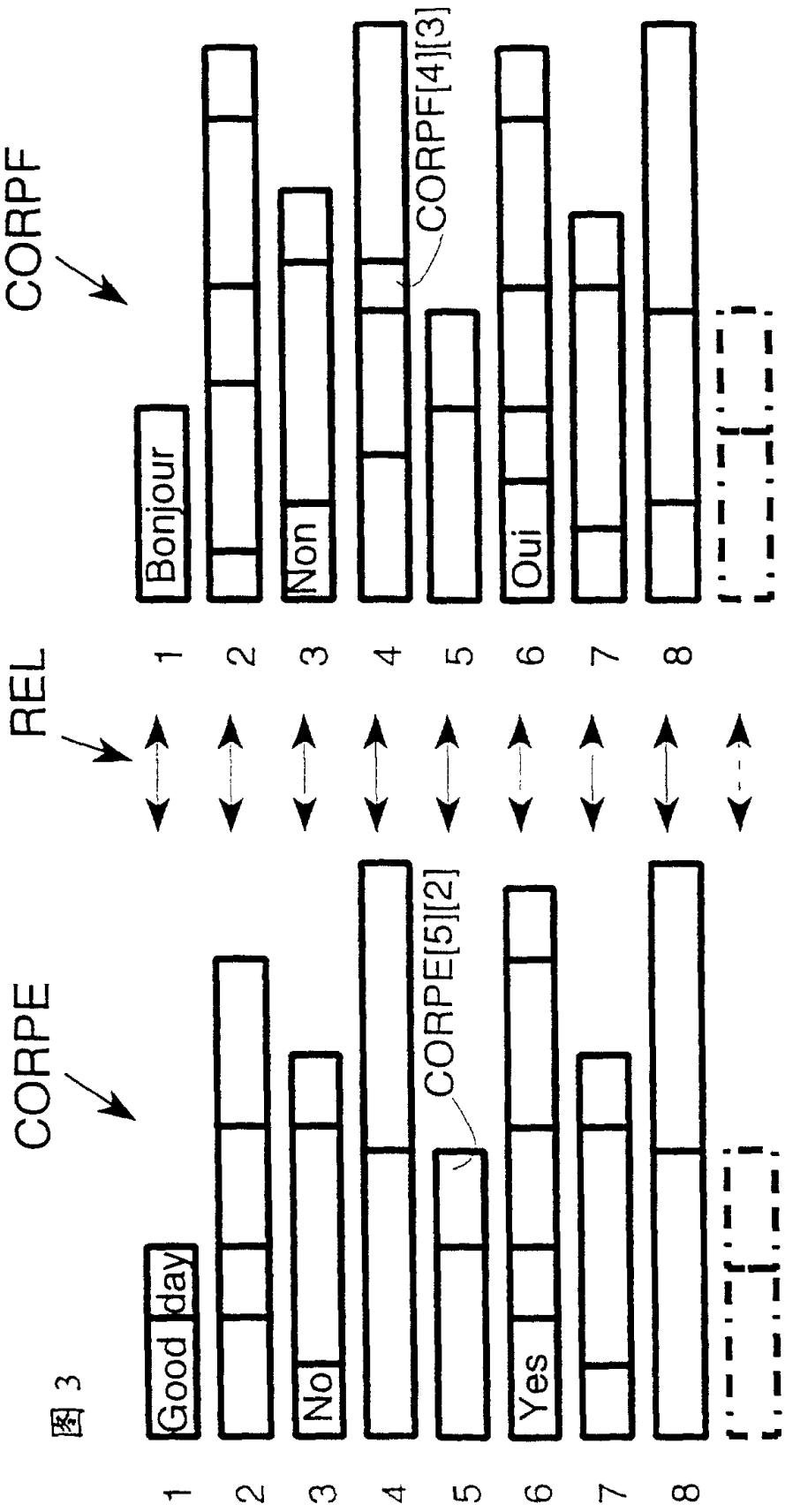

CORPF
CORPF[4][3]
Bonjour
Non
Oui
REL
1
2
3
4
5
6
7
8
CORPE
CORPE[5][2]
Good day
No
Yes

图 3

图 4

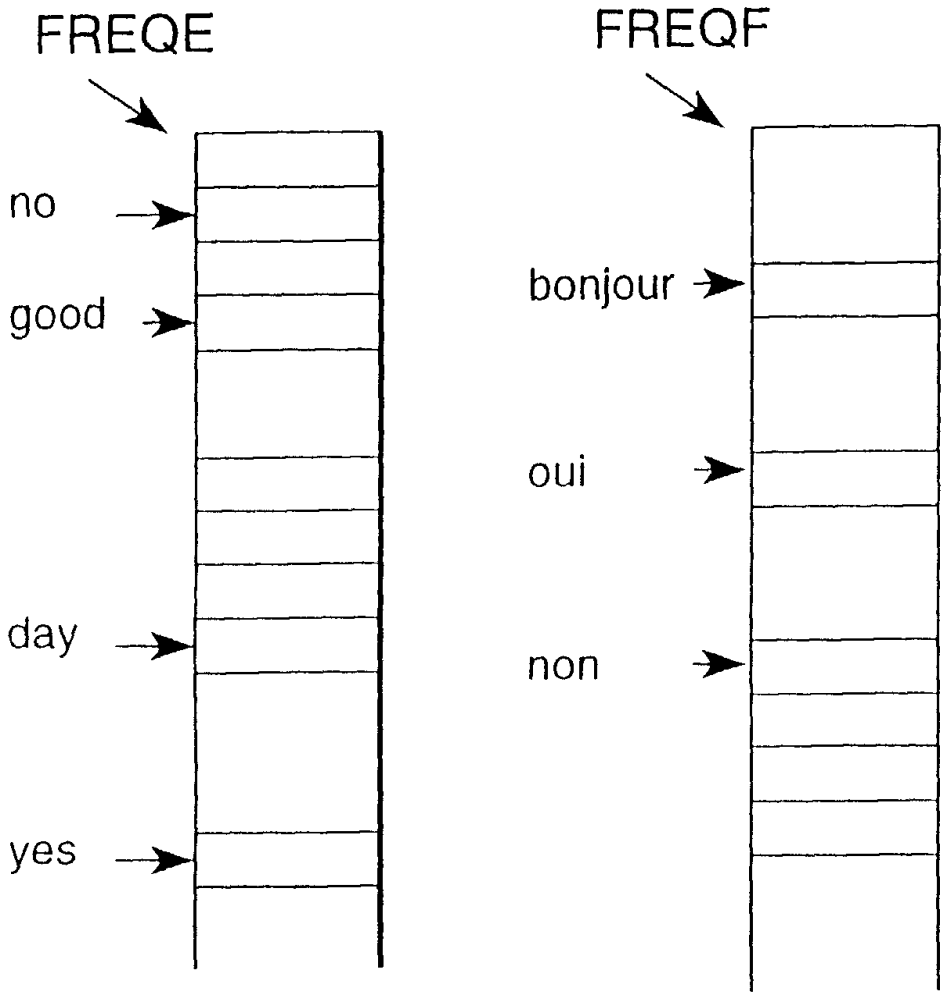

FREQE
no
good
day
yes
FREQF
bonjour
oui
non

图 5

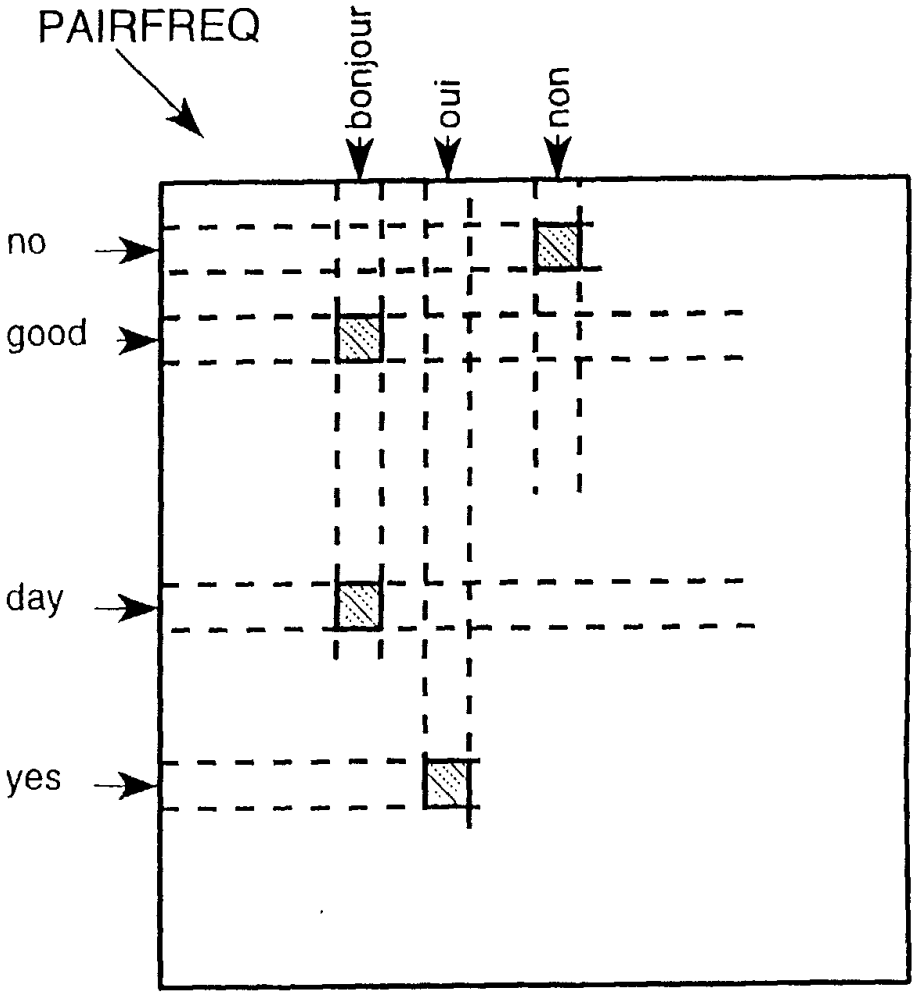

PAIRFREQ
bonjour
oui
non
no
good
day
yes

图 6

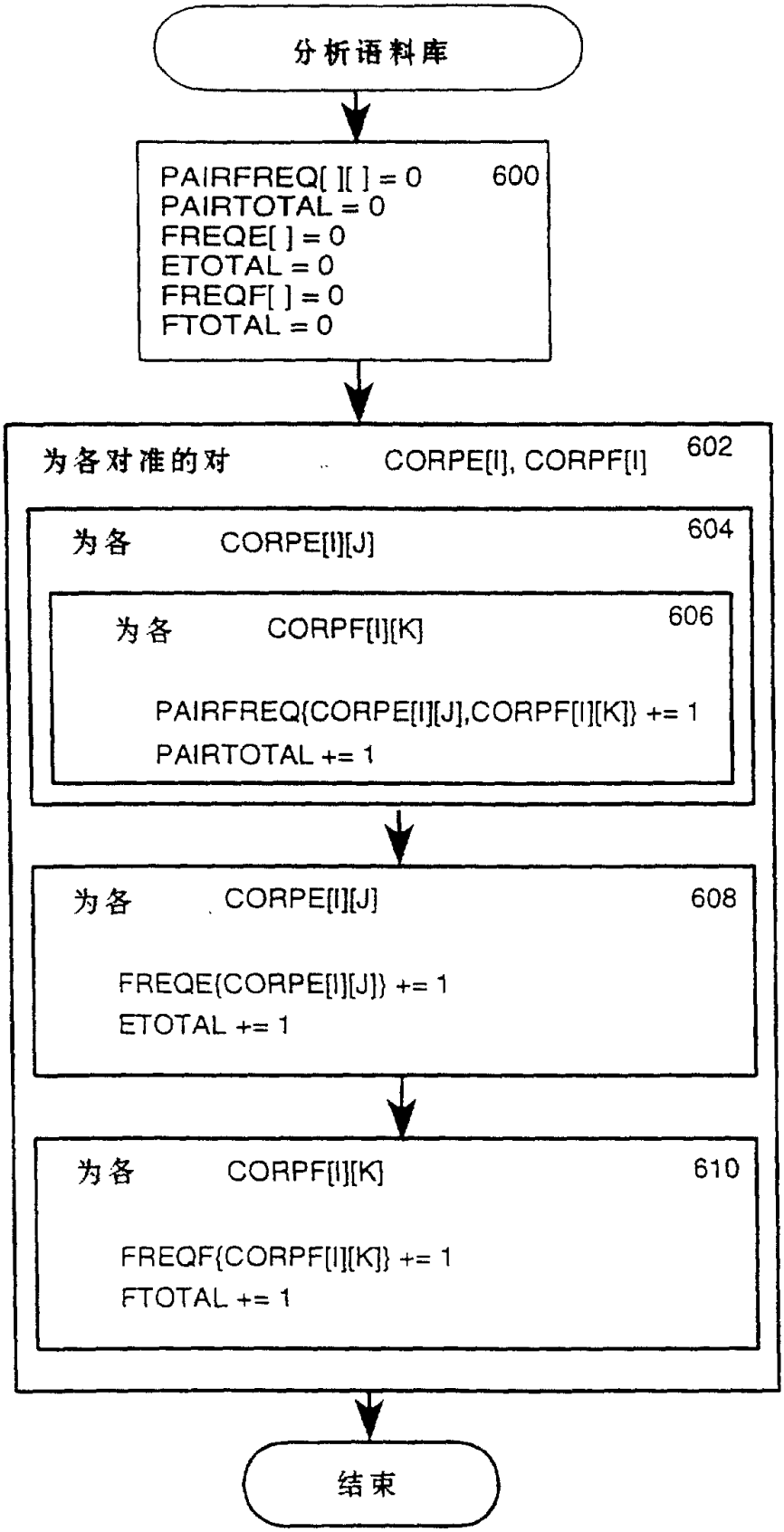

分析语料库
PAIRFREQ[ ][ ] = 0
PAIRTOTAL = 0
FREQE[ ] = 0
ETOTAL = 0
FREQF[ ] = 0
FTOTAL = 0
600
为各对准的对 CORPE[I], CORPF[I]
602
为各 CORPE[I][J]
604
为各 CORPF[I][K]
606
PAIRFREQ{CORPE[I][J],CORPF[I][K]} += 1
PAIRTOTAL += 1
为各 CORPE[I][J]
608
FREQE{CORPE[I][J]} += 1
ETOTAL += 1
为各 CORPF[I][K]
610
FREQF{CORPF[I][K]} += 1
FTOTAL += 1
结束

图 7

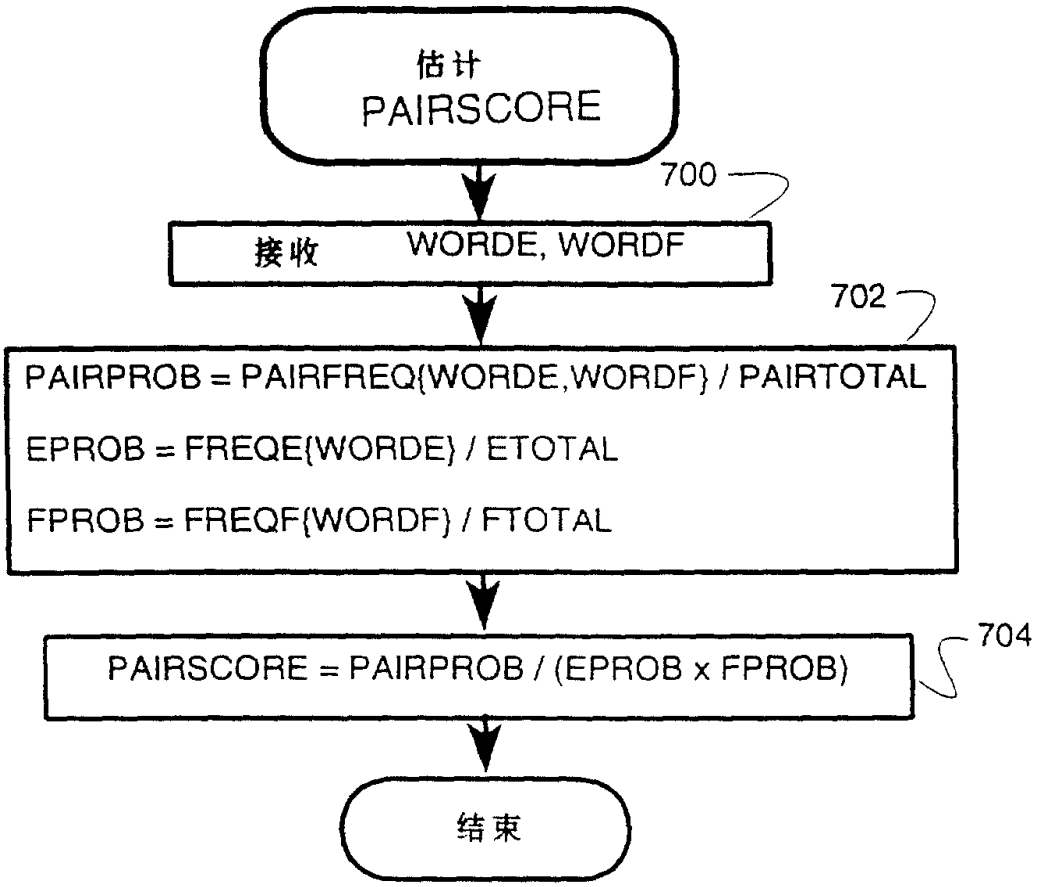

估计
PAIRSCORE
700
接收 WORDE, WORDF
702
PAIRPROB = PAIRFREQ(WORDE,WORDF) / PAIRTOTAL
EPROB = FREQE(WORDE) / ETOTAL
FPROB = FREQF(WORDF) / FTOTAL
704
PAIRSCORE = PAIRPROB / (EPROB x FPROB)
结束

图 8

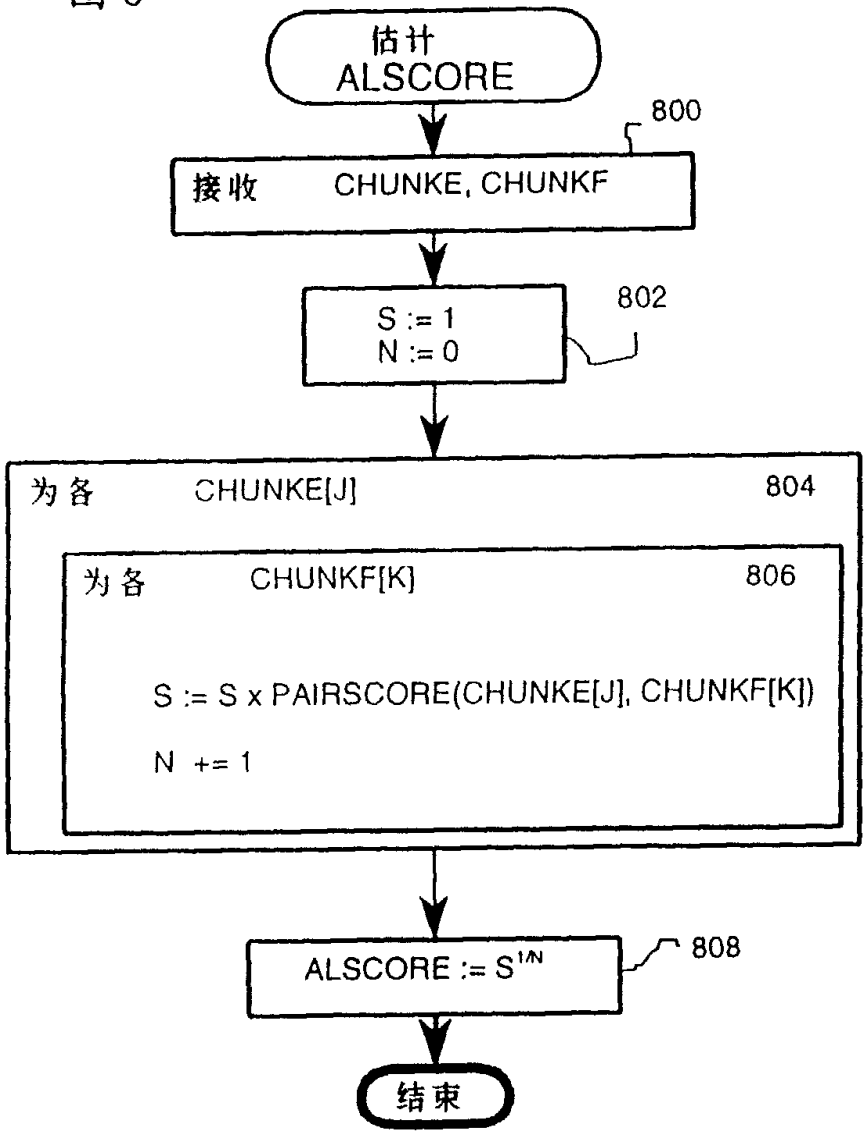

估计
ALSCORE
800
接收 CHUNKE, CHUNKF
S := 1
N := 0
802
为各 CHUNKE[J]
804
为各 CHUNKF[K]
806
S := S x PAIRSCORE(CHUNKE[J], CHUNKF[K])
N += 1
ALSCORE := S^1/N
808
结束

图 9

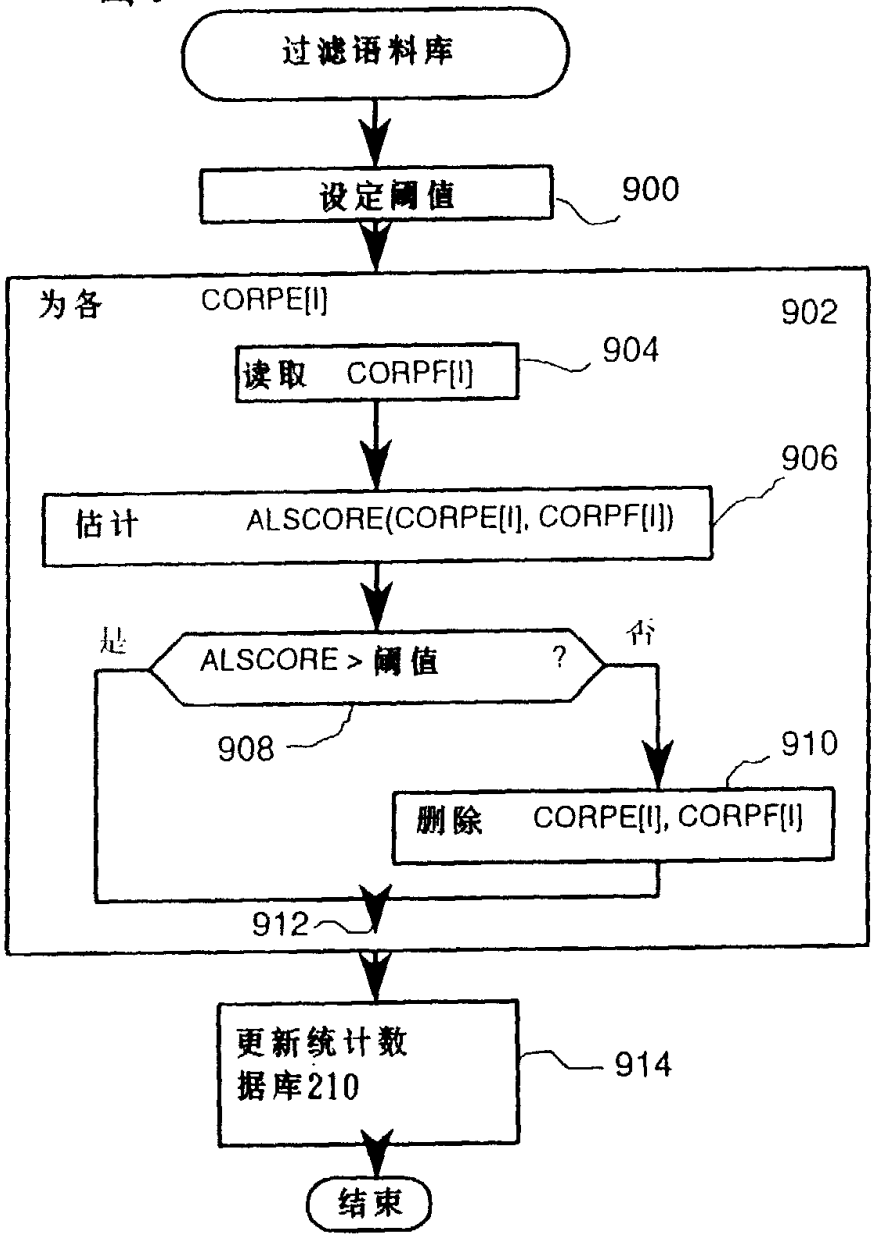

过滤语料库
设定阈值
900
为各 CORPE[I]
902
读取 CORPF[I]
904
估计 ALSCORE(CORPE[I], CORPF[I])
906
是
ALSCORE > 阈值 ?
否
908
910
删除 CORPE[I], CORPF[I]
912
更新统计数据库210
914
结束

图 10

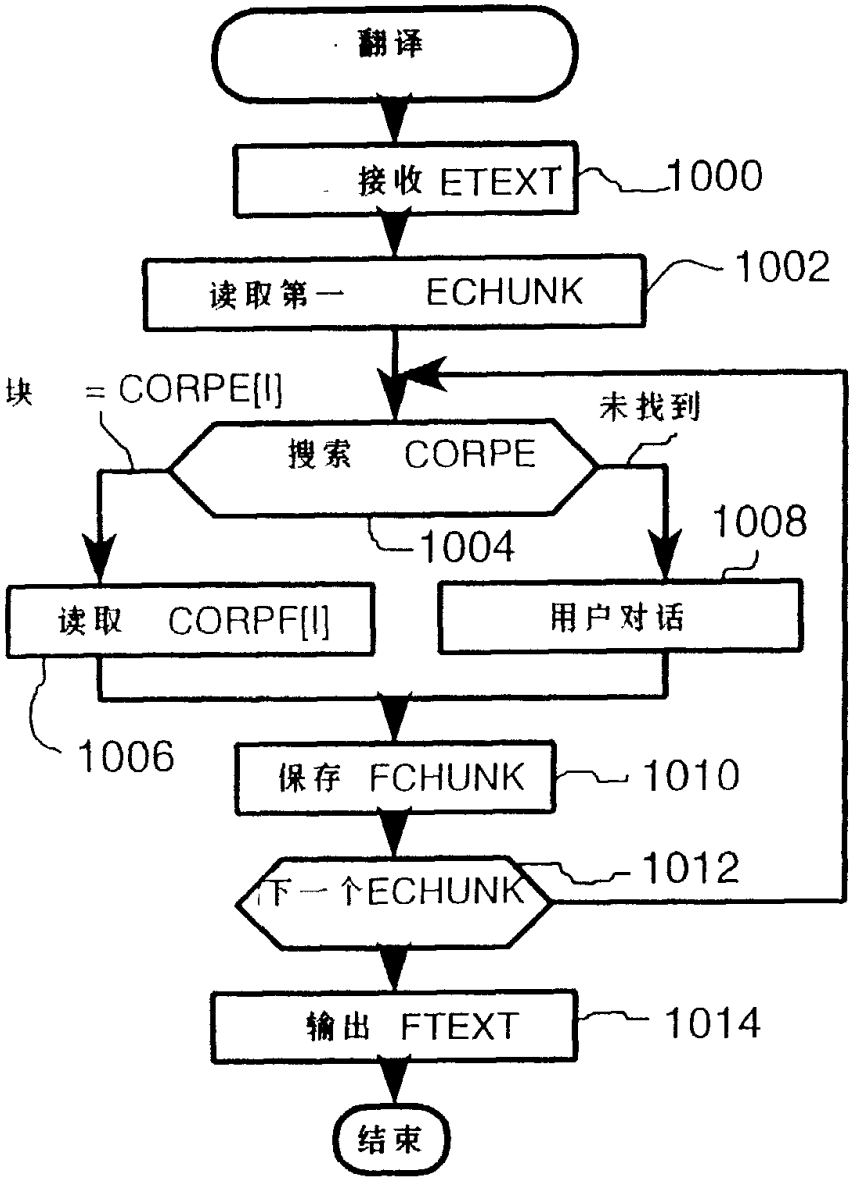
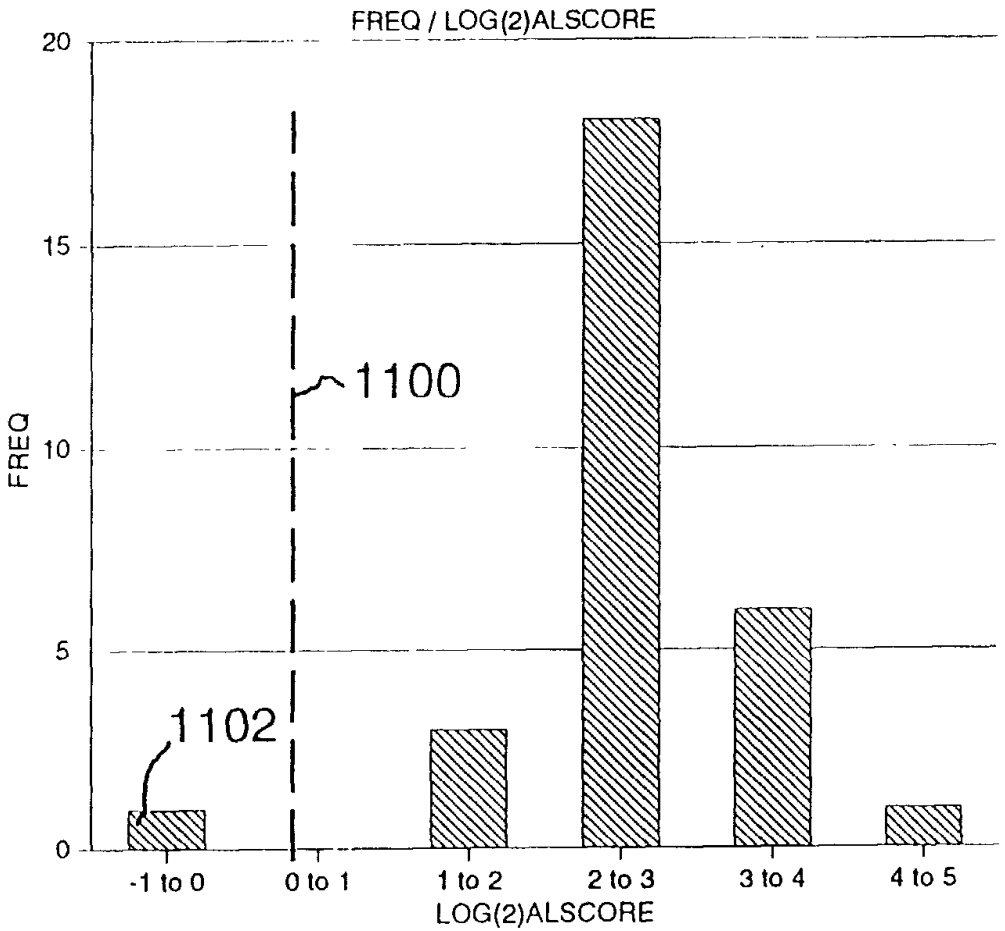

翻译
接收 ETEXT
1000
读取第一 ECHUNK
1002
块 = CORPE[I]
搜索 CORPE
未找到
1004
1008
读取 CORPF[I]
用户对话
1006
保存 FCHUNK
1010
下一个ECHUNK
1012
输出 FTEXT
1014
结束

图 11


FREQ / LOG(2)ALSCORE
FREQ
20
15
10
5
0
1100
1102
-1 to 0
0 to 1
1 to 2
2 to 3
3 to 4
4 to 5
LOG(2)ALSCORE