[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 17/28

[12] 发明专利申请公开说明书

[21] 申请号 01808510.5

[43] 公开日 2003 年 6 月 25 日

[11] 公开号 CN 1426561A

[22] 申请日 2001.4.23 [21] 申请号 01808510.5

[30] 优先权

[32] 2000. 4.24 [33] US [31] 60/199,288

[86] 国际申请 PCT/US01/13090 2001.4.23

[87] 国际公布 WO01/82111 英 2001.11.1

[85] 进入国家阶段日期 2002.10.23

[71] 申请人 微软公司

地址 美国华盛顿

[72] 发明人 徐恩东

[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所

代理人 付建军

权利要求书 8 页 说明书 18 页 附图 12 页

[54] 发明名称 带有跨语言阅读向导的计算机辅助阅读系统和方法

[57] 摘要

某个用户正在以非母语进行阅读，当他需要帮助时，计算机辅助阅读系统为他提供帮助，而无需该用户从文本上转移注意力。 在一个实施例中，阅读系统的实现形式为浏览程序的阅读向导。 这种阅读向导通过某个图形用户界面(UI)而显现，该用户界面允许用户在非母语的文本中选择某个单词、短语、语句或者单词的其它组合，并查看选定文本在该用户自己母语中的译文。 该译文在选定文本附近的窗口或弹出框中出现，以便使分心最小。

ISSN1008-4274

1. 一种阅读系统，包括:

一种用户界面，配置为允许用户选择非母语的文本并查看选定文本在母语中的译文；以及

一种跨语言的阅读向导，包括:

一种分析器，把选定的文本分解为各个翻译单位，

一种单词译文选择器，对若干翻译单位选择候选单词译文，以及

一种译文生成器，把候选单词译文转换为母语中对应的单词和短语，它们可以通过用户界面呈现给用户。

2. 根据权利要求1的阅读系统，其特征在于，该分析器包括一种词态学分析器，从词态学的角度处理各个单词，以获得每个单词的词态学词根。

3. 根据权利要求1的阅读系统，其特征在于，该分析器包括一种语句部分／基本名词短语标识模块，用于给各个单词加上标签标识。

4. 根据权利要求3的阅读系统，其特征在于，该语句部分／基本名词短语标识模块包括一种统计模型。

5. 根据权利要求1的阅读系统，其特征在于，该分析器包括一种短语扩展模块，对各个单词应用短语扩展规则。

6. 根据权利要求1的阅读系统，其特征在于，该译文生成器包括一种词典模块，把候选单词译文转换为对应的单词和短语。

7. 根据权利要求6的阅读系统，其特征在于，该词典模块包括一种单词词典。

8. 根据权利要求6的阅读系统，其特征在于，该词典模块包括一种短语词典。

9. 根据权利要求6的阅读系统，其特征在于，该词典模块包括一种不规则词形词典。

10. 根据权利要求 1 的阅读系统，其特征在于，该译文生成器包

括一种模板模块，它包括一个或多个模板，用于把候选单词译文转换为对应的单词和短语。

11. 根据权利要求 1 的阅读系统，其特征在于，该译文生成器包括一种规则模块，它包含多个规则，用于把非母语单词翻译为母语单词。

12. 根据权利要求 1 的阅读系统，其特征在于，该译文生成器包括一个或多个统计模型。

13. 根据权利要求 1 的阅读系统，实施为一个浏览器。

14. 一种阅读系统，包括:

一种用户界面，配置为允许用户选择英语的文本并查看选定文本的汉语译文；以及

一种跨语言的阅读向导，包括:

一种分析器，把选定的文本分解为各个翻译单位，

一种单词译文选择器，对若干翻译单位选择候选单词译文，以及

一种译文生成器，把候选单词译文转换为汉语中对应的短语，它们可以通过用户界面呈现给用户。

15. 根据权利要求 14 的阅读系统，实施为一个浏览器。

16. 一种阅读系统，包括:

一种用户界面，配置为允许用户选择非母语的文本并查看选定文本的母语译文，该用户界面包括弹出式窗口，用户可在其中查看母语文本；以及

一种跨语言的阅读向导，配置为:

接收用户已经选定的非母语文本，以及

自动地把非母语文本翻译成母语文本。

17. 根据权利要求 16 的阅读系统，其特征在于，该弹出式窗口显示在用户已经选定的文本附近。

18. 根据权利要求 16 的阅读系统，其特征在于，该弹出式窗口是可滚动的，以显示多种译文，并显示在用户已经选定的文本附近。

19. 根据权利要求 16 的阅读系统，实现为一个浏览器。

20. 一种计算机辅助阅读方法，包括:

通过用户界面向用户呈现非母语文本;

接收用户选定的文本;

处理用户选定的文本，以提供已经从非母语翻译成母语的文本;以及

通过用户界面向用户呈现翻译后的文本。

21. 根据权利要求 20 的计算机辅助阅读方法，其特征在于，所述处理包括:

把文本分解为翻译单位；以及

对于一个或多个翻译单位，获得词态学词根。

22. 根据权利要求 20 的计算机辅助阅读方法，其特征在于，所述处理包括:

把文本分解为翻译单位；以及

使用语句部分标签和基本名词短语标识，按特征划分翻译单位。

23. 根据权利要求 20 的计算机辅助阅读方法，其特征在于，所述处理包括:

把文本分解为翻译单位;

使用语句部分标签和基本名词短语标识，按特征划分翻译单位;以及

对按特征划分的翻译单位应用基于规则的短语扩展和模式匹配，以提供树列表。

24. 根据权利要求 23 的计算机辅助阅读方法，其特征在于，所述处理进一步包括根据树列表产生母语的候选单词译文。

25. 根据权利要求 24 的计算机辅助阅读方法，其特征在于，所述处理进一步包括把候选单词译文转换为母语中对应的单词和/或短语。

26. 一种或多种计算机可读的介质，其中具有计算机可读的指令，当某个处理器执行这些指令时，它们指示计算机实现权利要求 20 的方

法。

27. 根据权利要求20的计算机辅助阅读方法，其特征在于，所列举的动作由某个浏览器来实现。

28. 一种阅读系统，包括:

一种或多种计算机可读的介质；以及

记录在该介质上的代码，配置为实现某个浏览器，该浏览器配置为:

通过用户界面向用户呈现英语文本;

接收用户选定的文本;

处理用户选定的文本，以提供已经从英语翻译成汉语的文本；以及

通过用户界面向用户呈现翻译后的文本。

29. 根据权利要求28的阅读系统，其特征在于，该浏览器配置为向用户呈现同一英语文本的多种译文。

30. 根据权利要求28的阅读系统，其特征在于，该浏览器配置为在用户选定的英语文本附近的译文窗口中呈现翻译后的文本。

31. 根据权利要求28的阅读系统，其特征在于，该浏览器配置为在用户选定的英语文本附近的译文窗口中呈现同一英语文本的多种译文，该译文窗口具有下拉的特性，以显示多种译文的至少某一些。

32. 一种计算机辅助阅读方法，包括:

通过用户界面向用户呈现英语文本;

接收用户选定的文本;

处理用户选定的文本，以提供已经从英语翻译成汉语的文本；以及

通过用户界面向用户呈现翻译后的文本。

33. 根据权利要求32的计算机辅助阅读方法，其特征在于，所述处理包括:

把文本分解为翻译单位；以及

对于一个或多个翻译单位，获得词态学词根。

34. 根据权利要求32的计算机辅助阅读方法，其特征在于，所述处理包括:

把文本分解为翻译单位；以及

使用语句部分标签和基本名词短语标识，按特征划分翻译单位。

35. 根据权利要求32的计算机辅助阅读方法，其特征在于，所述处理包括:

把文本分解为翻译单位;

使用语句部分标签和基本名词短语标识，按特征划分翻译单位;以及

对按特征划分的翻译单位应用基于规则的短语扩展和模式匹配，以提供树列表。

36. 根据权利要求35的计算机辅助阅读方法，其特征在于，所述处理进一步包括根据树列表产生汉语的候选单词译文。

37. 根据权利要求36的计算机辅助阅读方法，其特征在于，所述处理进一步包括把候选单词译文转换为汉语中对应的单词和/或短语。

38. 一种或多种计算机可读的介质，其中具有计算机可读的指令，当某个处理器执行这些指令时，它们指示计算机实现权利要求32的方法。

39. 一种计算机辅助阅读方法，包括:

通过用户界面向用户呈现非母语文本，使得用户能够选取至少一个单词;

自动确定是否有对应的短语与选定的一个单词相关联；以及

以母语呈现至少选定单词的一种或多种译文，或者如果有对应的短语与选定的单词相关联，以母语呈现对应短语的至少一种译文。

40. 根据权利要求39的计算机辅助阅读方法，其特征在于，所述呈现包括在译文窗口中呈现译文，该窗口邻近对应的、选定的至少一个单词。

41. 根据权利要求40的计算机辅助阅读方法，其特征在于，所述

译文窗口是可滚动的，以便呈现多种不同的译文。

42. 根据权利要求39的计算机辅助阅读方法，其特征在于，所述呈现包括呈现多种最可能的译文。

43. 根据权利要求42的计算机辅助阅读方法，其特征在于，所述呈现进一步包括根据上下文分选最可能的译文。

44. 根据权利要求39的计算机辅助阅读方法，进一步包括:

接收用户输入，该输入表示用户仅仅需要翻译组成短语一部分的、某个选定的单词；以及

仅仅呈现选定单词的一种或多种译文。

45. 一种或多种计算机可读的介质，其中具有计算机可读的指令，当某个处理器执行这些指令时，它们指示计算机实现权利要求39的方法。

46. 一种阅读系统，包括:

一种或多种计算机可读的介质；以及

记录在该介质上的代码，配置为实现某个浏览器，该浏览器配置为:

使得用户能够选取通过用户界面呈现的至少一个英语单词;

自动确定是否有对应的短语与选定的至少一个英语单词相关联;以及

以汉语呈现选定的至少一个英语单词的一种或多种译文，或者如果有对应的短语与选定的至少一个英语单词相关联，以汉语呈现对应短语的至少一种译文。

47. 一种跨语言的用户界面，包括:

第一区域，配置为显示非母语文本；以及

第二区域，配置为以母语显示文本中至少一部分的译文。

48. 根据权利要求47的跨语言用户界面，其特征在于，第二区域安排在用户选定来翻译的至少某些文本附近。

49. 根据权利要求47的跨语言用户界面，其特征在于，非母语包括英语，母语包括汉语。

50. 根据权利要求47的跨语言用户界面，其特征在于，第二区域包括弹出式窗口。

51. 根据权利要求50的跨语言用户界面，其特征在于，弹出式窗口包括下拉特性，以显示另外的译文。

52. 根据权利要求47的跨语言用户界面，其特征在于，第二区域显示同一文本的多种不同的译文。

53. 一种跨语言的用户界面，包括:

第一区域，其中可以显示文本以便用户选择，该文本以第一语言显示；以及

第二区域，邻近用户选定的文本，该第二区域配置为显示已经翻译成第二语言的文本，翻译后的文本对应于用户已经选定的文本。

54. 根据权利要求53的跨语言用户界面，其特征在于，第一语言包括英语，第二语言包括汉语。

55. 根据权利要求53的跨语言用户界面，其特征在于，第二区域包括弹出式窗口。

56. 根据权利要求53的跨语言用户界面，其特征在于，第二区域包括弹出式窗口，它具有下拉特性，以显示多种译文。

57. 根据权利要求53的跨语言用户界面，其特征在于，第二区域显示同一文本的多种不同的译文。

58. 一种阅读系统，包括:

一种跨语言的阅读向导，包括:

一种分析器，把选定的文本分解为各个翻译单位，该分析器包括包括一种语句部分／基本名词短语标识模块，用于给各个单词加上标签标识，

一种单词译文选择器，对若干翻译单位选择候选单词译文，以及

一种译文生成器，把候选单词译文转换为母语中对应的单词或短语，它们可以通过用户界面呈现给用户。

59. 根据权利要求58的阅读系统，其特征在于，该分析器包括一种词态学分析器，从词态学的角度处理各个单词，以获得每个单词的

词态学词根。

60. 根据权利要求58的阅读系统，其特征在于，该分析器包括一种短语扩展模块，对各个单词应用短语扩展规则。

61. 一种或多种计算机可读的介质，其中具有计算机可读的指令，当一个或多个处理器执行这些指令时，使这一个或多个处理器实现一种跨语言的阅读向导，包括:

一种分析器，把选定的文本分解为各个翻译单位，该分析器包括包括一种语句部分／基本名词短语标识模块，用于给各个单词加上标签标识，

一种单词译文选择器，对若干翻译单位选择候选单词译文，以及

一种译文生成器，把候选单词译文转换为母语中对应的单词或短语，它们可以通过用户界面呈现给用户。

62. 一种或多种计算机可读的介质，其中具有计算机可读的指令，当一个或多个处理器执行这些指令时，使这一个或多个处理器:

通过用户界面向用户呈现非母语文本;

接收用户选定的文本;

处理选定的文本，步骤包括:

把选定的文本分解为翻译单位，

使用语句部分标签和基本名词短语标识，按照特征划分翻译单位，

对按特征划分的翻译单位应用基于规则的短语扩展和模式匹配，以提供树列表，

根据树列表产生母语的候选单词译文，

把候选单词译文转换为母语中对应的单词和／或短语，以提供已经从非母语翻译成母语的文本；以及

通过用户界面向用户呈现翻译后的文本。

带有跨语言阅读向导的计算机辅助阅读系统和方法

相关申请

本申请源于2000年4月24日提交的、序列号为60/199,288的美国临时申请，并要求其优先权，其公开内容在此特别引用作为参考。本申请也涉及2000年4月24日提交的、序列号为09/556,229的美国专利申请，其公开内容在此引用作为参考。

技术领域

本发明涉及机器辅助阅读系统和方法，更具体地说，本发明涉及有助于用户阅读非母语的一种用户界面和基础体系结构。

背景技术

随着因特网的快速发展，遍布全世界的计算机用户正在越来越多地面对非母语写成的文字材料。许多用户对非母语完全不懂。即使用户进行了一些非母语培训，该用户也往往难以阅读和理解非母语。

考虑某个中国用户访问以英语写成的网页或其它电子文档的情况。这个中国用户在校期间可能接受过某些正规的英语培训，但是这种培训往往不足以使他完全地阅读和理解以英语写成的、特定的单词、短语或者语句。这种汉语－英语的情况仅仅是用于说明这一点的一个例子而已。这个问题在跨越其它语言边界时也是存在的。

所以，本发明起源于提供机器辅助阅读系统和方法有关的问题，这些系统和方法帮助计算机用户阅读和理解非母语表达的文字材料。

发明内容

某个用户正在以非母语进行阅读，当他需要帮助时，计算机辅助

阅读系统为他提供帮助，而无需该用户从文本上转移注意力。

在一个实施例中，阅读系统的实现形式为浏览程序的阅读向导。这种阅读向导通过某个图形用户界面（UI）而显现，该用户界面允许用户在非母语的文本中选择某个单词、短语、语句或者单词的其它组合，并查看选定文本在该用户自己母语中的译文。该译文在选定文本附近的窗口或弹出框中出现，以便使分心最小。

一方面，阅读向导的核心包括浅层分析器、统计单词译文选择器和译文生成器。浅层分析器把用户选定的非母语文本中的短语或语句分解成单独的翻译单位（例如短语、单词）。在一个实施例中，浅层分析器将选定文本中的单词划分成片段，并从词态学的角度对其进行分析，以获得每个单词的词态学词根。为了进一步选择译文，浅层分析器采用语句部分（POS）标签和基本名词短语（baseNP）标识，按照特征划分单词和短语。例如，POS 标签和 baseNP 标识可以由某个统计模型来实现。浅层分析器对这些单词应用基于规则的短语扩展和模式匹配，以产生树列表。

对于从非母语文本剖析中得出的翻译单位，统计单词译文选择器选择最优的候选单词译文。单词译文选择器产生所有可能的译文模式，并使用统计翻译和语言模型对该翻译单位进行翻译。输出的是最优的候选译文。

译文生成器将候选单词译文转换成母语中对应的短语。译文生成器部分地使用母语模型，以帮助确定适当的译文。然后，母语单词和短语就通过邻近选定文本的 UI 显示出来。

附图简要说明

图 1 是一个计算机系统的框图，它实现了带有跨语言阅读向导的阅读系统。

图 2 是一个实施例中示范浅层分析器的框图。

图 3 是一张示意图，用于理解一个实施例中进行的处理。

图 4 是一张示意图，用于理解图 3 中的示意图。

图 5 是一张流程图，介绍一个实施例中方法的步骤。

图 6 是一张示意图，用于理解一个实施例中进行的处理。

图 7 是一张流程图，介绍一个实施例中方法的步骤。

图 8 是一个实施例中示范译文生成器的框图。

图 9 至图 13 显示了一个实施例中多种示范用户界面。

具体实施方式

概述

计算机辅助阅读系统帮助用户阅读非母语。为了进行讨论，在一般的浏览程序范围内介绍计算机辅助阅读系统，该程序在通用计算机上运行。不过，计算机辅助阅读系统也可以在许多不同的非浏览环境中实现（例如电子邮件系统、字处理等），并且可以在许多各式各样类型的设备上实行。

下面介绍的实施例，允许更乐于使用母语交流的用户快捷方便地广泛阅读非母语的电子文档，而且其方式促进了集中精力和快速消化有关内容。在要翻译文本的近旁提供带有译文窗口的用户界面，可以增加用户的便利。译文窗口中包含着文本翻译后的译文。将译文窗口放在要翻译文本的近旁，用户的双眼不必移动很远即可确定要翻译的文本。这又减少了用户可察觉的分心，举例来说，假若为了查看要翻译的文本，用户的目光需要扫视一段距离，这种分心就很难避免。

在某些实施例中，光标翻译处理的优点使得用户交互进一步增强。利用鼠标移动来选择文本的位置，用户能够进行快速选择，系统据此自动执行翻译并将翻译后的文本向用户显示。

示范系统体系结构

图 1 显示了示范计算机系统 100，它具有中央处理器（CPU）102、存储器 104 和输入／输出（I/O）接口 106。CPU 102 与存储器 104 和 I/O 接口 106 通信。存储器 104 既代表易失性存储器（例如 RAM），又代表非易失性存储器（例如 ROM、硬盘等）。程序和数据文件可以

存放在存储器 104 中，并在 CUP 102 上执行。

计算机系统 100 通过 I/O 接口 106 连接着一台或多台外围设备。示范外围设备包括鼠标 110、键盘 112(例如字母数字 QWERTY 键盘、电话键盘等)、显示监视器 114、打印机 116、外部存储设备 118 和话筒 120。例如，该计算机系统可以作为通用计算机。所以，计算机系统 100 执行一种计算机操作系统(未显示)，该操作系统在存储器 104 中存放，在 CUP 102 上执行。该操作系统最好是一种支持窗口环境的多任务操作系统。适宜的操作系统的一个实例是微软公司的 Windows 品牌操作系统。

应当注意，也可以使用其它计算机系统配置，比如手持设备、多处理器系统、基于微处理器的或者可编程的消费电子产品、网络 PC、小型计算机、大型计算机等等。此外，尽管图 1 中展示的是一种独立的计算机，语言输入系统也可以在分布式计算环境中实现，通过通信网络(例如 LAN、因特网等)连接的远程处理设备完成其中的任务。在分布式计算环境中，程序模块既可以位于本机的存储设备中，也可以位于远程存储设备中。

示范阅读系统

计算机系统 100 实现了阅读系统 130，它帮助用户阅读非母语。该阅读系统可以在单词、短语或者语句的级别提供帮助。在图 1 中实现的阅读系统是作为浏览程序 132，它在存储器 104 中存放，在 CUP 102 上执行。应当重视和理解，下面介绍的阅读系统也可以在浏览器范围以外的氛围中实现。

阅读系统 130 具有用户界面 134 和跨语言阅读向导 136。UI 134 展现跨语言阅读向导 136。除了阅读系统之外，浏览器程序 132 还可以包括其它组件，但是这类组件对于浏览器程序被视为标准的，因此将不进行详细显示和介绍。

阅读向导 136 包括浅层分析器 140、统计单词译文选择器 142 和译文生成器 144。

示范浅层分析器

浅层分析器 140 把选定的非母语文本中的短语或语句分解成单独的翻译单位（例如短语、单词）。

图 2 稍微详细地显示了一个实施例中的浅层分析器。在任何适当的硬件、软件、固件或者它们的组合中，都可以实现该浅层分析器。在这个展示和介绍的实施例中，浅层分析器是在软件中实现的。

如图所示，浅层分析器 140 包括单词片段模块 200、词态学分析器 202、语句部分（POS）标签／基本名词短语标识模块 204、短语扩展模块 206 和模式或模板匹配模块 208。虽然这些组件显示为独立的组件，但是应当重视和理解，这些组件也可以互相结合或者与其它组件结合。

依据所介绍的实施例，浅层分析器 140 把用户已经选定的文本中的单词划分为片段。它使用单词片段模块来做到这一点。然后浅层分析器使用词态学分析器 202 从词态学的观点处理单词，以获得每个单词的词态学词根。为了发现每个单词的词态学词根，词态学分析器能够对单词应用多种词态学规则。词态学分析器 202 使用的规则可以由熟练于被分析的具体语言的人制定。例如，英语中的一个规则为，以“ed”结尾的单词，或者通过去除“d”或者通过去除“ed”来形成其词态学词根。

为了进一步选择译文，浅层分析器 140 采用语句部分（POS）标签/基本名词短语（baseNP）标识模块 204，按照特征划分单词和短语。例如，POS 标签和 baseNP 标识可以由某个统计模型来实现，下面紧接着的标题为“POS 标签和 baseNP 标识”的一节介绍了它的一个实例。浅层分析器 140 使用短语扩展模块 206，对 POS 标签/基本名词短语标识模块 204 按照特征划分的这些单词，应用基于规则的短语扩展。短语扩展模块的一个目标，是把一个基本名词短语扩展到一个更复杂的名词短语。例如，“baseNP 的 baseNP”是“baseNP”短语的更复杂的名词短语。浅层分析器 140 也使用模式或模板匹配模块

208 来产生树列表。模式或模板匹配模块用于翻译，并认识某些短语翻译是取决于模式的，而不是直接与短语中的单词有关。例如，短语“对 baseNP 感兴趣”包含一种模式（即 “baseNP”），它用于形成一个更复杂的翻译单位，用于翻译。单词 “感兴趣”并不直接涉及用于形成更复杂翻译单位的这种模式。

POS 标签和 baseNP 标识

以下的讨论介绍了用于自动识别英语的 baseNP（名词短语）的一种统计模型，并构成了处理选定的文本以产生树列表的方式之一。所介绍的方法使用两个步骤：N 个最佳语句部分（POS）标签和给予 N 个最佳 POS 序列的 baseNP 标识。所介绍的模型也集成了词法的信息。最后，应用 Viterbi 算法在整个序列中进行全局搜索，它使得在整个过程中能够获得线性的复杂度。

发现简单的和非递归的基本名词短语（baseNP），是许多自然语言处理应用程序（比如部分剖析、信息检索和机器翻译）的一个重要的子任务。baseNP 是简单的名词短语，并不递归地包含其它的名词短语。例如，在以下实例中[…]之内的元素是 baseNP，其中 NNS、IN、VBG 等为语句部分（POS）标签。Marcus et al., *Building a Large Annotated Corpus of English: the Penn Treebank*, Computational Linguistics, 19(2): 313-330, 1993 中介绍了 POS 标签，使其为世人所知。

[Measures/NNS] of/IN [manufacturing/VBG activity/NN] fell/VBD more/RBR than/IN

[the/DT overall/JJ measures/NNS] ./.

统计方法

在本节中介绍两步法统计模型、参数训练和 Viterbi 算法，用于搜索 POS 标签和 baseNP 标识最佳序列。在介绍该算法之前，引入贯穿讨论的某些记号。

让我们把输入语句 E 分别表示为下列单词序列和 POS 序列：

$$E = w_1 \quad w_2 \quad \dots \quad w_{n-1} \quad w_n$$

$$T = t_1 \quad t_2 \quad \dots \quad t_{n-1} \quad t_n$$

其中 n 为语句中的单词数目，t_i 为单词 w_i 的 POS 标签。

给定 E 后，假设 baseNP 标识的结果是一个序列，其中某些单词组合成 baseNP 如下

$$\dots w_{i-1} \quad [w_i \quad w_{i+1} \quad \dots w_j] \quad w_{j+1} \dots$$

对应的标签序列如下：

$$\text{(a) } B = \dots t_{i-1} \quad [t_i \quad t_{i+1} \quad \dots t_j] \quad t_{j+1} \dots \quad = \dots t_{i-1} \quad b_{i,j} \quad t_{j+1} \quad \dots \quad = n_1 \quad n_2 \quad \dots \quad n_m$$

其中 $b_{i,j}$ 对应于某个 baseNP 的标签序列：$[\, t_i\, t_{i+1} \cdots \ t_j \,]$。$b_{i,j}$ 也可以被认为是一种 baseNP 规则。所以，序列 B 既包含 POS 标签，又包含 baseNP 规则。因此 $1 \leqslant m \leqslant n$，$n_i \in$（POS 标签集 ∪ baseNP 规则集）。这是某个语句带有 baseNP 记号的第一表达式。有时我们也使用以下等效形式：

$$\text{(b) } Q = \dots (t_{i-1}, bm_{i-1}) \quad (t_i, bm_i) \quad (t_{i+1}, bm_{i+1}) \dots \quad (t_j, bm_j) \quad (t_{j+1}, bm_{j+1}) \dots \quad = \quad q_1 \quad q_2 \cdots \quad q_n$$

其中每个 POS 标签 t_i 与其相对于 baseNP 的位置信息 bm_i 相关联。位置信息为{F,I,E,O,S}之一。F、E 和 I 分别表示该单词在某个 baseNP 的左边界、右边界，或者某个 baseNP 内部的另一个位置。O 表示该单词在某个 baseNP 之外。S 标注某个单一单词的 baseNP。

例如，上面给定实例的两种表达式如下：

(a) B= [NNS] IN [VBG NN] VBD RBR IN [DT JJ NNS]

(b) Q=(NNS S) (IN O) (VBG F) (NN E) (VBD O) (RBR O) (IN O) (DT F) (JJ I) (NNS E) (. O)

集成的'两步法过程

所介绍方法的原理如下。最可能的 baseNP 序列 B^*可以一般地表示如下:

$$B^* = \underset{B}{\operatorname{argmax}}(p(B|E))$$

我们把整个过程分成两步，也就是:

$$B^* \approx \underset{B}{\operatorname{argmax}}(P(T|E) \times P(B|T,E)) \qquad (1)$$

为了减少搜索空间和计算复杂度，我们仅仅考虑 N 个最佳 POS 标签，也就是

$$T(N-best) = \underset{T=T_1,\ldots,T_N}{\operatorname{argmax}}(P(T|E)) \qquad (2)$$

所以，我们有:

$$B^* \approx \underset{B,T=T_1,\ldots,T_N}{\operatorname{argmax}}(P(T|E) \times P(B|T,E)) \qquad (3)$$

相应地，算法也由两步组成: 使用（2）式确定 N 个最佳 POS 标签，然后使用（3）式从这些 POS 序列确定最佳 baseNP 序列。这两步集成在一起，而不是像在其它方法中那样分开。现在让我们更密切地查看这两步。

确定 N 个最佳 POS 序列

在第一步中算法的目标是在搜索空间（POS 网格）之内搜索 N 个最佳 POS 序列。根据贝叶斯法则，我们有

$$P(T|E)=\frac{P(E|T)\times P(T)}{P(E)}$$

由于 $P(E)$不影响 $P(T|E)$的最小化过程，（2）式变为

$$T(N-best)=\underset{T=T_1,\ldots,T_N}{\operatorname{argmax}}(P(T|E))=\underset{T=T_1,\ldots,T_N}{\operatorname{argmax}}(P(E|T)\times P(T)) \qquad (4)$$

我们现在假设，E 中的单词是独立的。因此

$$P(E|T)\approx\prod_{i=1}^{n}P(w_i|t_i) \qquad (5)$$

然后我们使用三字模型作为 $P(T)$的一种近似，也就是：

$$P(T)\approx\prod_{i=1}^{n}P(t_i|t_{i-2},t_{i-1}) \qquad (6)$$

最后我们有

$$T(N-best)=\underset{T=T_1,\ldots,T_N}{\operatorname{argmax}}(P(T|E))=\underset{T=T_1,\ldots,T_N}{\operatorname{argmax}}(\prod_{i=1}^{n}P(w_i|t_i)\times P(t_i|t_{i-2},t_{i-1})) \qquad (7)$$

在搜索 N 个最佳的 Viterbi 算法中，$P(w_i|t_i)$被称为词法生成（或输出）概率，$P(t_i|t_{i-2},t_{i-1})$被称为隐藏马尔可夫模型中的跃进概率。Viterbi, *Error Bounds for Convolution Codes and Asymptotically Optimum Decoding Algorithm*, IEEE Transactions on Information Theory

IT-13(2):pp.260-269, April, 1967 中介绍了 Viterbi 算法。

确定 baseNP

如上所述，给定 N 个最佳 POS 序列后，第二步的目标是搜索最佳 baseNP 序列。

把 E、T 和 B 视为随机变量，根据贝叶斯法则，我们有

$$P(B|T,E)=\frac{P(B|T)\times P(E|B,T)}{P(E|T)}$$

$$P(B|T)=\frac{P(T|B)\times P(B)}{P(T)}$$

由于 ，我们有，

$$P(B|T,E)=\frac{P(E|B,T)\times P(T|B)\times P(B)}{P(E|T)\times P(T)} \qquad (8)$$

因为我们对给定语句 E 中每个可能的 POS 序列搜索最佳 baseNP 序列， $P(E|T)\times P(T)=P(E\cap T)$=常数。不仅如此，从 B 的定义，在每个搜索过程期间，我们有 $P(T|B)=\prod_{i=1}^{n}P(t_i,...,t_j|b_{i,j})=1$。所以，（3）式变为

$$B^*=\underset{B,T=T_1,...,T_N}{\arg\max}(P(T|E)\times P(B|T,E))$$

$$=\underset{B,T=T_1,...,T_N}{\arg\max}(P(T|E)\times P(E|B,T)\times P(B)) \qquad (9)$$

使用独立假设，我们有

$$P(E|B,T)\approx\prod_{i=1}^{n}P(w_i|t_i,bm_i). \qquad (10)$$

利用 $P(B)$的三字近似，我们有:

$$P(B)\approx\prod_{i=1}^{m}P(n_i\mid n_{i-2},n_{i-1}) \qquad (11)$$

最后，我们得到

$$B^{*}=\arg\max_{B,T=T_1,..T_N}(P(T\mid E)\times\prod_{i=1}^{n}P(w_i\mid bm_i,t_i)\times\prod_{i=1,m}P(n_i\mid n_{i-2},n_{i-1})) \qquad (12)$$

总而言之，在第一步中，在 POS 标签过程中应用搜索 N 个最佳的 Viterbi 算法，对于计算的每个 POS 序列确定路径概率 f_t 如下: $f_t=\prod_{i=1,n}p(w_i\mid t_i)\times p(t_i\mid t_{i-2},t_{i-1})$。在第二步中，对于每个可能的 POS 标签结果，再一次应用 Viterbi 算法搜索最佳 baseNP 序列。在这一步所找到的每个 baseNP 序列还与路径概率 $f_b=\prod_{i=1}^{n}p(w_i\mid t_i,bm_i)\times\prod_{i=1,m}p(n_i\mid n_{i-2},n_{i-1})$ 相关。某个 baseNP 序列的总概率由$f_t^{\alpha}\times f_b$确定，其中 α 为均衡系数（在我们的实验中 α = 2.4）。对于给定的语句 E，当我们确定最佳 baseNP 序列时，我们也确定了 E 的最佳 POS 序列，它对应于 E 的最佳 baseNP。

作为这个过程如何起作用的一个实例，考虑以下文本： “stock was down 9.1 points yesterday morning.”在第一步中，该句的 N 个最佳 POS 标签结果之一是: T=NN VBD RB CD NNS NN NN。

对于这个 POS 序列，第二步将试图确定 baseNP，如图 3 所示。虚线中路径的细节在图 4 中给出。在第二步中计算出它的概率如下（Φ 为虚变量）:

$$\begin{aligned}P(B\mid T,E)&=p(stock\mid NN,S)\times p(was\mid VBD,O)\times p(down\mid RB,O)\times p(NUMBER\mid CD,B)\\&\times p(points\mid NNS,E)\times p(yesterday\mid NN,B)\times p(morning\mid NN,E)\times p(.\mid .,O)\\&\times p([NN]\mid\Phi,\Phi)\times p(VBD\mid\Phi,[NN])\times p(RB\mid[NN],VBD)\times p([CD\quad NNS]\mid VBD,RB)\\&\times p([NN\quad NN]\mid RB,[CD\quad NNS])\times p(.\mid[CD\quad NNS],[NN\quad NN])\end{aligned}$$

统计参数训练

在这项工作中，训练和试验的数据是取自 Penn Treebank 的 25 个部分。我们把整个 Penn Treebank 数据分成两部分，一部分用于训练，另一部分用于试验。

在我们的统计模型中，我们计算以下四种概率：(1)$P(t_i|t_{i-2},t_{i-1})$，(2)$P(w_i|t_i)$，(3)$P(n_i|n_{i-2}n_{i-1})$和 (4)$P(w_i|t_i,bm_i)$。第一种和第三种参数分别是 T 和 B 的三字概率。第二种和第四种是词法生成概率。利用下列公式可以从有 POS 标签的数据计算概率（1）和（2）：

$$p(t_i|t_{i-2},t_{i-1})=\frac{count(t_{i-2}t_{i-1}t_i)}{\sum_j count(t_{i-2}t_{i-1}t_j)} \qquad (13)$$

$$p(w_i|t_i)=\frac{count(w_i\ with\ \ tag\ \ t_i)}{count(t_i)} \qquad (14)$$

由于训练集中的每个语句既有 POS 标签，又有 baseNP 边界标签，它可以转换为两个序列，如上节中介绍的 B（a）和 Q（b）。使用这些序列，利用分别类似于（13）和（14）式的计算公式，可以计算出参数（3）和（4）。

在训练三字模型（3）之前，应当从训练全集提取所有可能的 baseNP 规则。例如，提取的 baseNP 规则包括以下三个序列。

(1) *DT CD CD NNPS*
(2) *RB JJ NNS NNS*
(3) *NN NN POS NN*

… …

在 Penn Treebank 中有超过 6000 种 baseNP 规则。训练三字模型（3）时，我们以两种方式对待这些 baseNP 规则。首先，给每个 baseNP 规则分配一个唯一标识（UID）。这表明，该算法考虑了每个 baseNP

规则对应的结构。其次，给所有这些规则分配相同标识（SID）。在这种情况下，这些规则组合成同一类。然而，baseNP 规则的标识仍然与分配给 POS 标签的标识不同。

为了使参数平滑，使用了 Katz, *Estimation of Probabilities from Sparse Data for Language Model Component of Speech Recognize*, IEEE Transaction on Acoustics, Speech, and Signal Processing, Volume ASSP-35, pp. 400-401, March 1987 中介绍的一种方法。为了预测参数（1）和（3）的概率，建立了一个三字模型。在识别 baseNP 期间遇到未知单词的情况下，以下列方式计算参数（2）和（4）：

$$p(w_i \mid bm_i, t_i) = \frac{count(bm_i, t_i)}{\max_j (count(bm_j, t_i))^2} \qquad (15)$$

$$p(w_i \mid t_i) = \frac{count(t_i)}{\max_j (count(t_j))^2} \qquad (16)$$

这里，bm_j 表示附着在 t_i 上所有可能的 baseNP 标号，t_j 是为未知单词 w_i 猜测的一个 POS 标签。

图 5 是一张流程图，介绍一个实施例中方法的步骤。这些步骤可以在任何适当的硬件、软件、固件或者它们的组合中实现。在展示的实例中，这些步骤是在软件中实现的。在上述跨语言的阅读向导 136——它形成了浏览器程序 132 的一部分（见图 1）——中，可以发现这种软件的一个具体实施例。更确切地说，将要介绍的方法可以由某个浅层分析器来实现，比如图 2 中显示和介绍的分析器。

步骤 500 接收选定的文本。实现这个步骤与某个用户选择文本中要翻译的一个部分有关。通常，用户通过使用一种输入设备比如鼠标等来选择文本。步骤 502 把选定文本中的单词划分成片段。将会受到本领域的技术人员重视的任何适当的划分处理都可以使用。步骤 504 获得每个单词的词态学词根。在展示和介绍的实施例中，这一步是由词态学分析器实现的，比如图 2 中显示的词态学分析器。在展示的实

例中，词态学分析器配置为处理以英语写成的单词。不过应当重视和理解，任何适当的语言都可以提供一个基础，在其上可以建立一个词态学分析器。

步骤 506 使用语句部分（POS）标签和基本名词短语标识来划分单词的特征。任何适当的技术都可以使用。在上面的“POS 标签和 baseNP 标识”一节中，详细介绍了一种示范技术。步骤 508 对按照特征划分的单词应用基于规则的短语扩展和模式匹配，以产生树列表。在上面的实例中，这一步是使用短语扩展模块 206 和模式或模板匹配模块 208 来实现的。步骤 510 输出树列表，以便进一步处理。

作为树列表的一个实例，考虑图 6。其中已经以上面介绍的方法，处理了语句“The Natural Language Computing Group at Microsoft Research China is exploring research in advanced natural language technologies”。确切地说，树列表展示了语句中已经使用上面介绍的 POS 标签和 baseNP 技术，经过分段、词态学处理和特征划分的各个单词。例如，考虑元素 600。其中，单词“Natural”已经从语句中和从父元素“natural language”中划分出来。元素 600 也已经按特征划分为带有 POS 标签“JJ”。树上的其它元素也经过类似的处理。

示范单词译文选择器

单词译文选择器 142 接收树列表并产生所有可能的译文模式。选择器 142 使用统计翻译和语言模型对剖析后的翻译单位进行翻译，得出母语文本中最优的候选单词译文。输出的是最优的候选译文。

图 7 是一张流程图，介绍一个实施例中方法的步骤。该方法可以在任何适当的硬件、软件、固件或者它们的组合中实现。在展示和介绍的实施例中，该方法是在软件中实现的。这种软件的一个实施例可以包括单词译文选择器 142（见图 1）。

按照上面介绍的处理产生的树列表在步骤 700 接收。步骤 702 从树列表产生译文模式。在一个实施例中，产生所有可能的译文模式。例如，对于英语到汉语的翻译，英语名词短语“NP1 of NP2”可能

有两种可能的译文：（1）T(NP1)+T(NP2) 和（2）T(NP2)+T(NP1)。在短语译文中，翻译的短语是一棵句法树，在一个实施例中，所有可能的译文顺序都受到考虑。步骤 704 使用翻译模型和语言模型来翻译剖析后的翻译单位。翻译单位可以包括单词和短语。然后步骤 706 输出最优的 N 个候选单词译文。可以使用统计模型来选择最优的 N 个候选单词译文。

示范译文生成器

译文生成器 144 把最优的 N 个候选单词译文转换为母语中对应的短语。然后母语的单词和短语通过 UI 出现在选定的文本附近。

图 8 稍微详细地显示了一个实施例中的译文生成器 144。为了翻译最优的候选单词，译文生成器可以利用许多不同的资源。例如，译文生成器可以包括它在翻译过程中使用的词典模块 800。词典模块 800 可以包括单词词典、短语词典、不规则词形词典或者通常在自然语言翻译处理中使用的任何其它词典，这对本领域的技术人员是显而易见的。本领域的技术人员将会理解这类词典的操作和功能，为简便起见此处不再更详细地介绍。

译文生成器 144 可以包括模板模块 802，它包含用于翻译处理中的多个模板。任何适当的模板都可以使用。例如，可以使用所谓的大短语模板来协助翻译过程。在自然语言翻译中使用的模板，其操作已知，此处不再更详细地介绍。

译文生成器 144 可以包括规则模块 804，它包含用于使翻译过程便利的多个规则。规则可以是手工制定的规则，制定它们的人员熟悉被翻译的特定语言。可以制定规则来涉及属于翻译、剖析、翻译模式中的统计错误。本领域的技术人员将会理解基于规则翻译的原理。

译文生成器 144 可以包括用于翻译过程中的一个或多个统计模型 806。这些可以使用的统计模型可以变化范围很广，尤其是对与所需翻译有关的、给定数目的可能的非母语和母语。这些统计模型可以基于上面介绍的 POS 和 baseNP 统计参数。在需要从英语翻译成汉语的一

个特定实施例中，可以使用以下模型：汉语三字语言模型和汉语交互信息模型。当然，也可以使用其它模型。

上面介绍的模块和模型可以分别使用，也可以按多种相互组合使用。

在处理过程中，至此用户已经选择了非母语文本的一部分，要翻译成母语。选定的文本已经按照上面的介绍进行了处理。在下面紧接着的讨论中，要介绍的方法和系统以对用户方便和高效的方式把翻译后的文本呈现给用户。

阅读向导的用户界面

剩余的讨论是针对阅读向导中用户界面 134 的特性。具体地说，阅读向导的用户界面 134 允许用户选择不能确定如何阅读和解释的、以非母语写成的文本。选定的内容可以是各个单词、短语或者语句。

图 9 至图 13 显示了阅读向导的若干示范用户界面，以图形 UI（GUI）实现，作为浏览器程序或者其它计算机辅助阅读系统的一部分呈现给用户。展示的实例显示了设计为帮助中国用户阅读英语文本的阅读系统。英语文本显示在窗口中。用户可以选择英语文本的某些部分。为了响应用户的选择，阅读向导把选定的内容翻译成汉语文本，并把汉语文本呈现在弹出式译文窗口或滚动框中。

图 9 显示了用户界面 900，它包括 “非母语”文本中已经突出显示的一部分。突出显示的文本显示在用户界面的第一区域中。用户界面的第二区域的形式为译文窗口 902，配置为以母语显示至少某些文本内容翻译后的部分。在这个实例中，突出显示的文本包括短语 “ research in advanced natural language technologies ”。在这个实例中，用户已经突出显示了单词 “ advanced ”，阅读系统已经自动确定该单词组成被突出显示短语的一部分。然后阅读系统在译文窗口 902 中自动显示被突出显示短语的最佳译文。通过自动确定包含用户选定单词的短语然后提供该短语的至少一种译文，不仅向读者提供了该单词的译文，而且提供了该单词在其中使用的、翻译后的上下文。这样

做的优点在于，它给予读者更多翻译后的信息，这又有助于他们理解正在阅读的材料。

注意，译文窗口 902 位于突出显示文本的至少一个部分的附近。以这种方式定位译文窗口，为了看见翻译后的文本，用户就不必使其注意力从突出显示的文本转移很远。这样做的优点在于，它不会令用户的阅读过程减慢到不期望的程度。也要注意，译文窗口包含一个下拉箭头 904，它可以用于显示选定文本的其它翻译结果。作为一个实例，考虑图 10。其中译文窗口 902 已经下拉，以显示突出显示短语的所有译文。

图 11 显示了用户界面 1100，它带有译文窗口 1102。注意，阅读系统自动探测出单词 “generated”不在某个短语中，并且只翻译单词 “generated”。阅读系统也能够在译文窗口 1102 中提供多种最可能的译文。例如，显示了三种可能的示范译文。在展示的实例中，显示的译文是与上下文有关的，并按照上下文来排序。所以，在这个实例中，阅读系统仅仅可以显示该单词最佳的 n 个译文，而不是该单词所有可能的译文。图 12 显示了用户界面 1100，其中单词 “generated”的所有可能的译文都在译文窗口 1102 中呈现给用户。

图 13 显示了用户界面 1300，它带有译文窗口 1302，展示了所介绍实施例的一种特性。具体地说，可以给用户一种选择，他们是需要翻译包含选定单词的整个短语，还是只需要翻译选定的单词。在这个实例中，用户已经使其鼠标放置在选择单词 “advanced”进行翻译的位置。由于单词 “advanced”组成某个较长短语的一部分，阅读系统会自动翻译包含着选定单词的短语，然后向用户呈现各种选择，如上所述。不过在这种情况下，用户已经指示阅读系统，他们仅仅需要翻译选定的单词。他们能够以任何适当的方式做到这一点，比如举例来说，当选定单词时按下 “Ctrl”键。

结论

上面介绍的若干实施例帮助用户阅读非母语，并能够允许更乐于

使用母语交流的用户快捷方便地广泛阅读非母语的电子文档，而且其方式促进了集中精力和快速消化有关内容。提供的用户界面在要翻译文本的近旁具有译文窗口，包含着文本翻译后的译文，可以增加用户的便利。将译文窗口放在要翻译文本的近旁，用户的双眼不必移动很远即可确定要翻译的文本。这又减少了用户可察觉的分心，举例来说，假若为了查看要翻译的文本，用户的目光需要扫视一段距离，这种分心就很难避免。在某些实施例中，光标翻译处理的优点使得用户交互进一步增强。利用鼠标移动来选择文本的位置，用户能够进行快速选择，系统据此自动执行翻译并将翻译后的文本向用户显示。

尽管是以结构特性和／或词态学步骤的特定语言来介绍了本发明，还是应当理解，在附带的权利要求书中规定的本发明不必局限于所介绍的特定特性或步骤。相反，公开这些特定的特性和步骤，是作为实现要求权利的本发明的优选形式。

图1

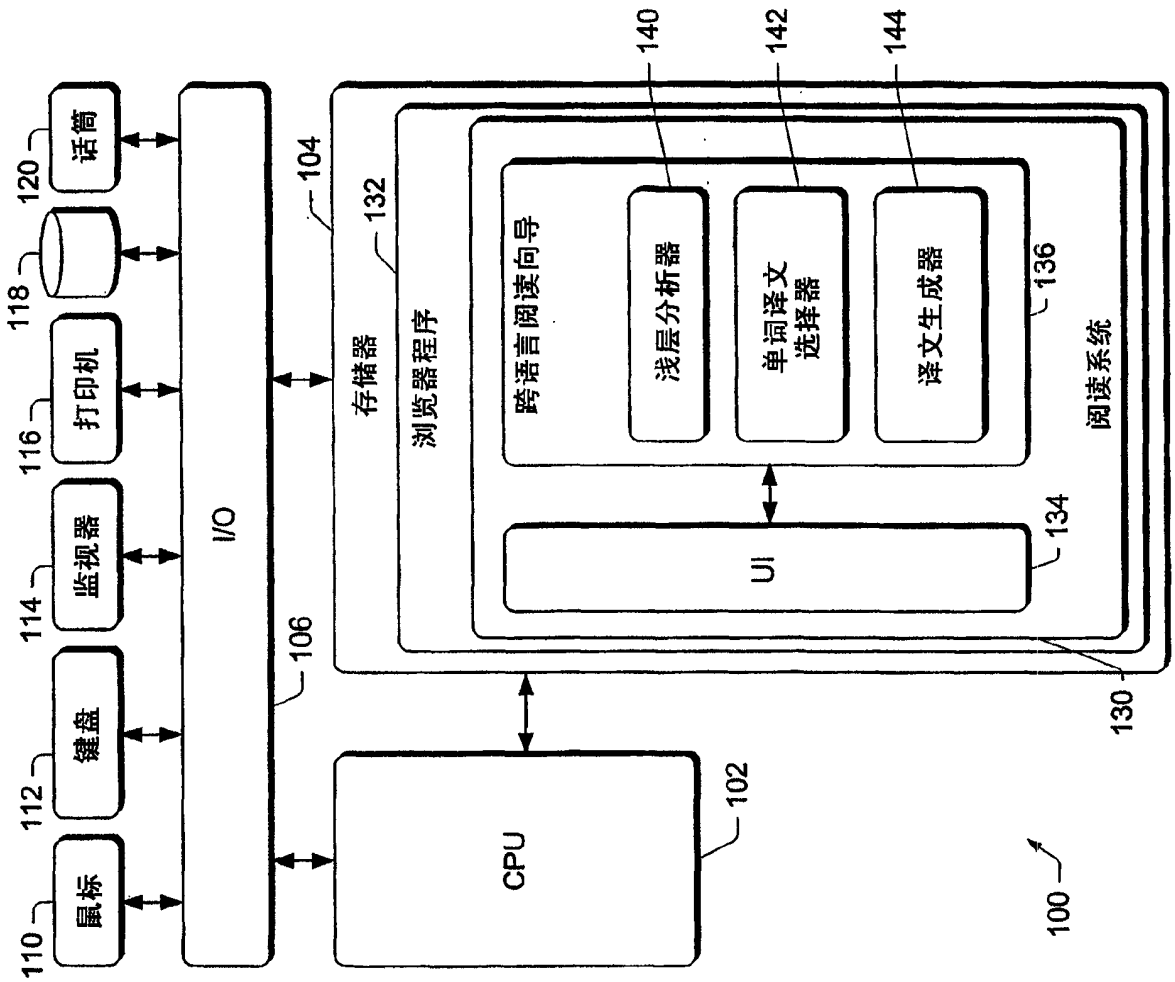
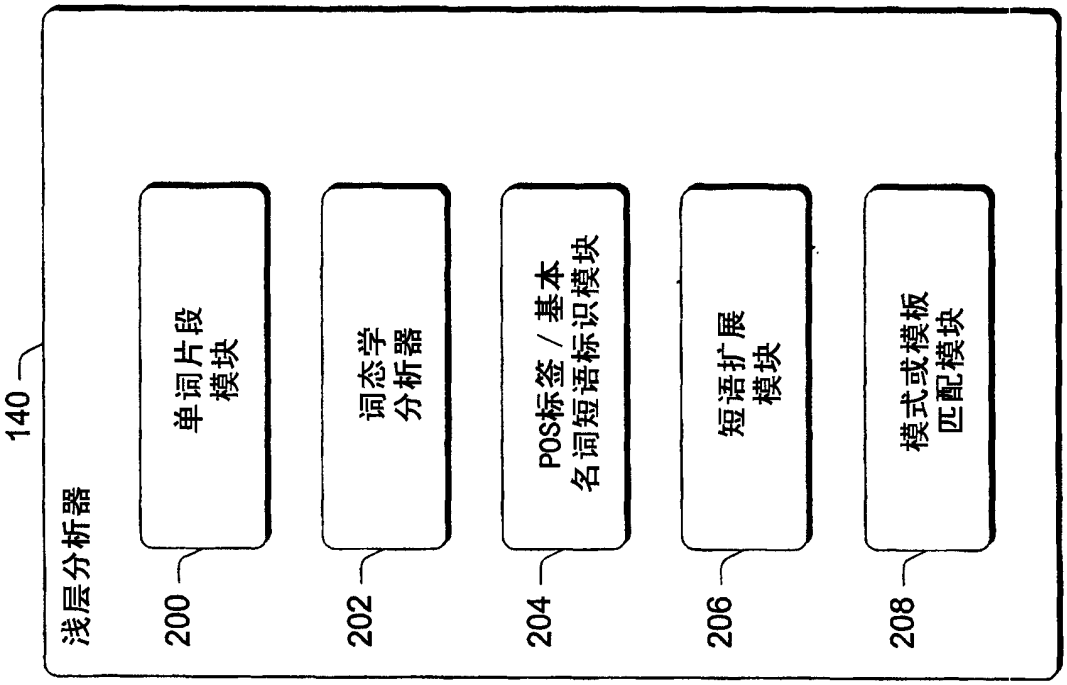

话筒
120
118
打印机
116
监视器
114
键盘
112
鼠标
110
I/O
106
104
存储器
132
浏览器程序
跨语言阅读向导
浅层分析器
单词译文选择器
译文生成器
140
142
144
136
UI
134
阅读系统
130
CPU
102
100

图2


140
浅层分析器
200
单词片段模块
202
词态学分析器
204
POS标签 / 基本名词短语标识模块
206
短语扩展模块
208
模式或模板匹配模块

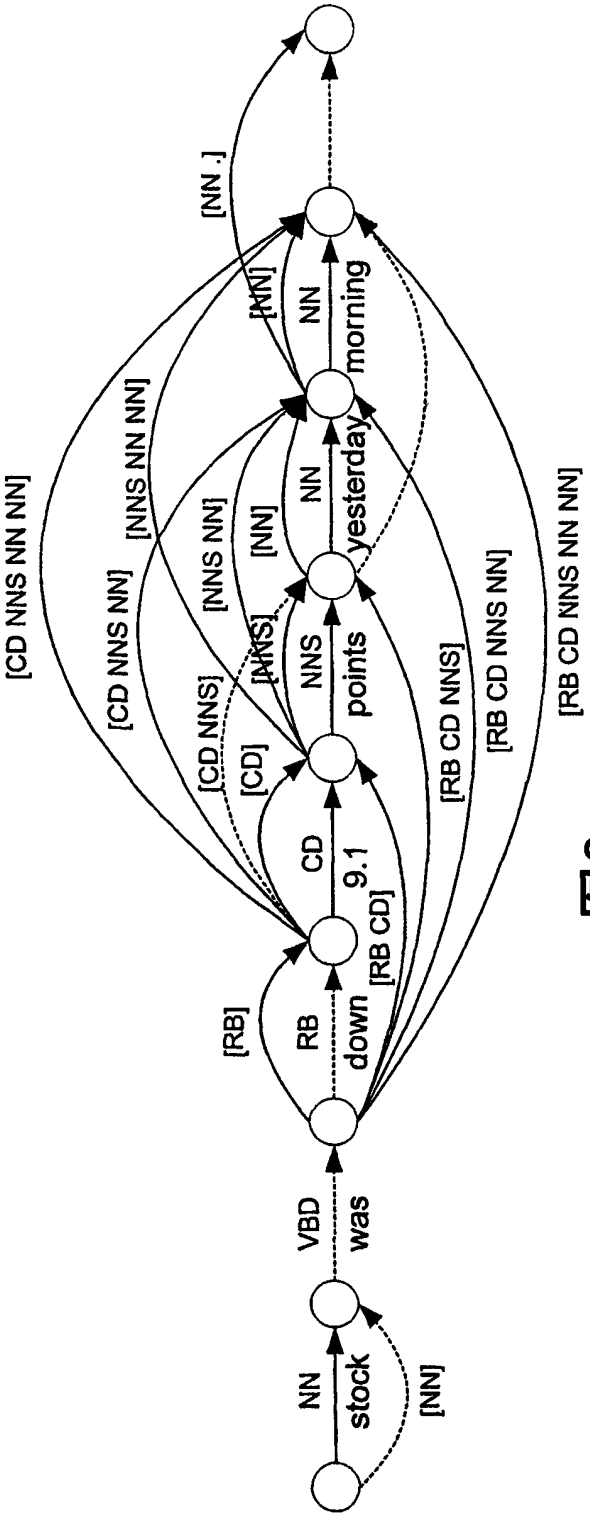

[CD NNS NN NN]
[CD NNS NN]
[NNS NN NN]
[CD NNS]
[NNS NN]
[CD]
[NNS]
[NN]
[RB]
[RB CD]
[RB CD NNS]
[RB CD NNS NN]
[RB CD NNS NN NN]
[NN .]
[NN]
NN
stock
VBD
was
RB
down
CD
9.1
NNS
points
NN
yesterday
NN
morning
[NN]

图3

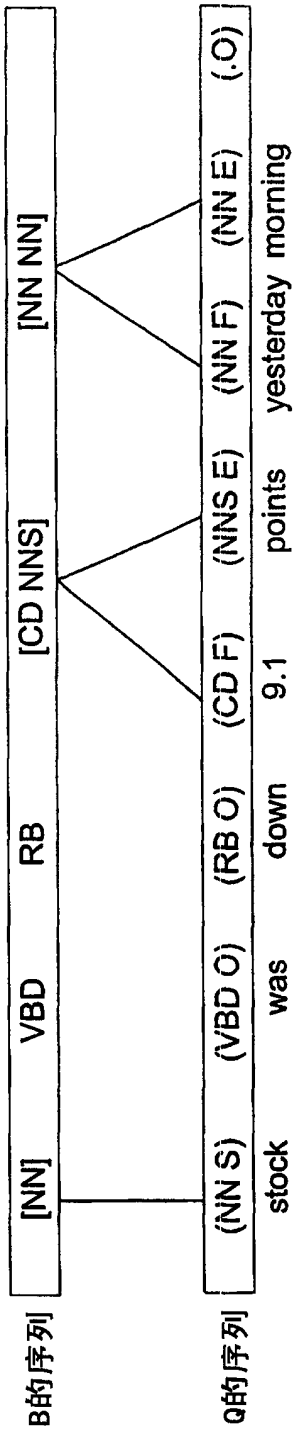

B的序列
[NN]
VBD
RB
[CD NNS]
[NN NN]
Q的序列
(NN S)
(VBD O)
(RB O)
(CD F)
(NNS E)
(NN F)
(NN E)
(.O)
stock
was
down
9.1
points
yesterday
morning

图4

图5

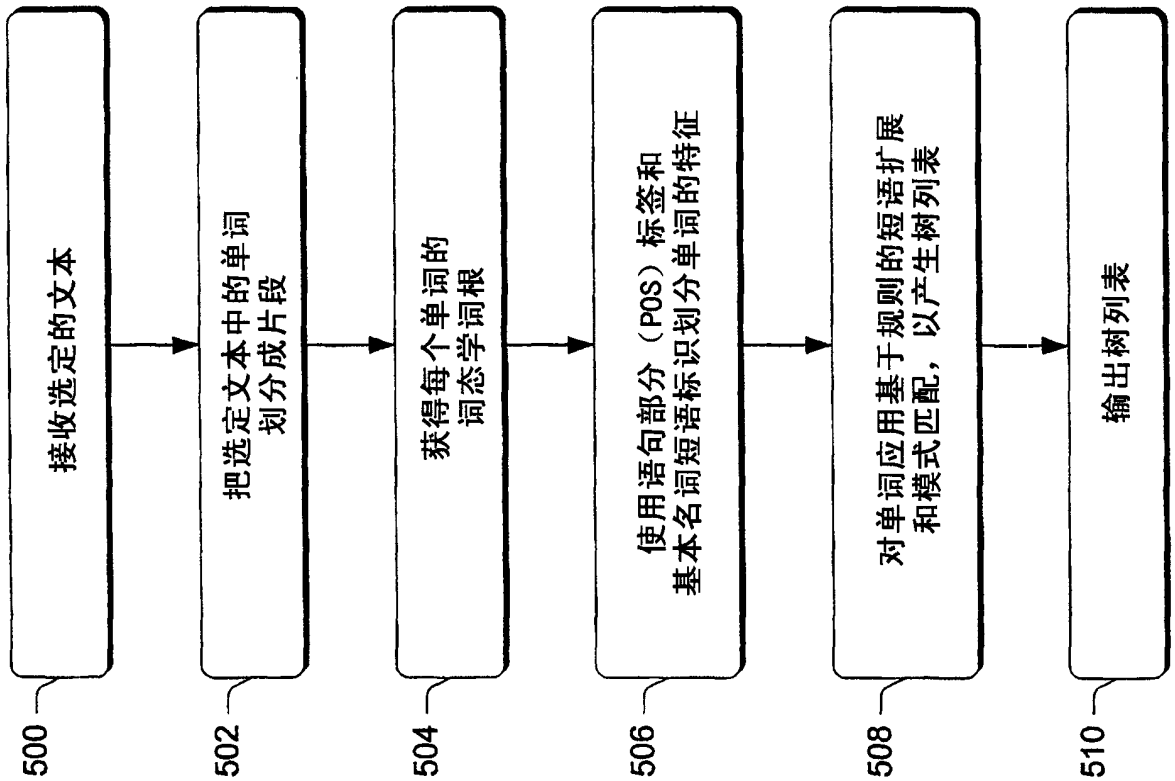

接收选定的文本
把选定文本中的单词划分成片段
获得每个单词的词态学词根
使用语句部分（POS）标签和基本名词短语标识划分单词的特征
对单词应用基于规则的短语扩展和模式匹配，以产生树列表
输出树列表
500
502
504
506
508
510

图6

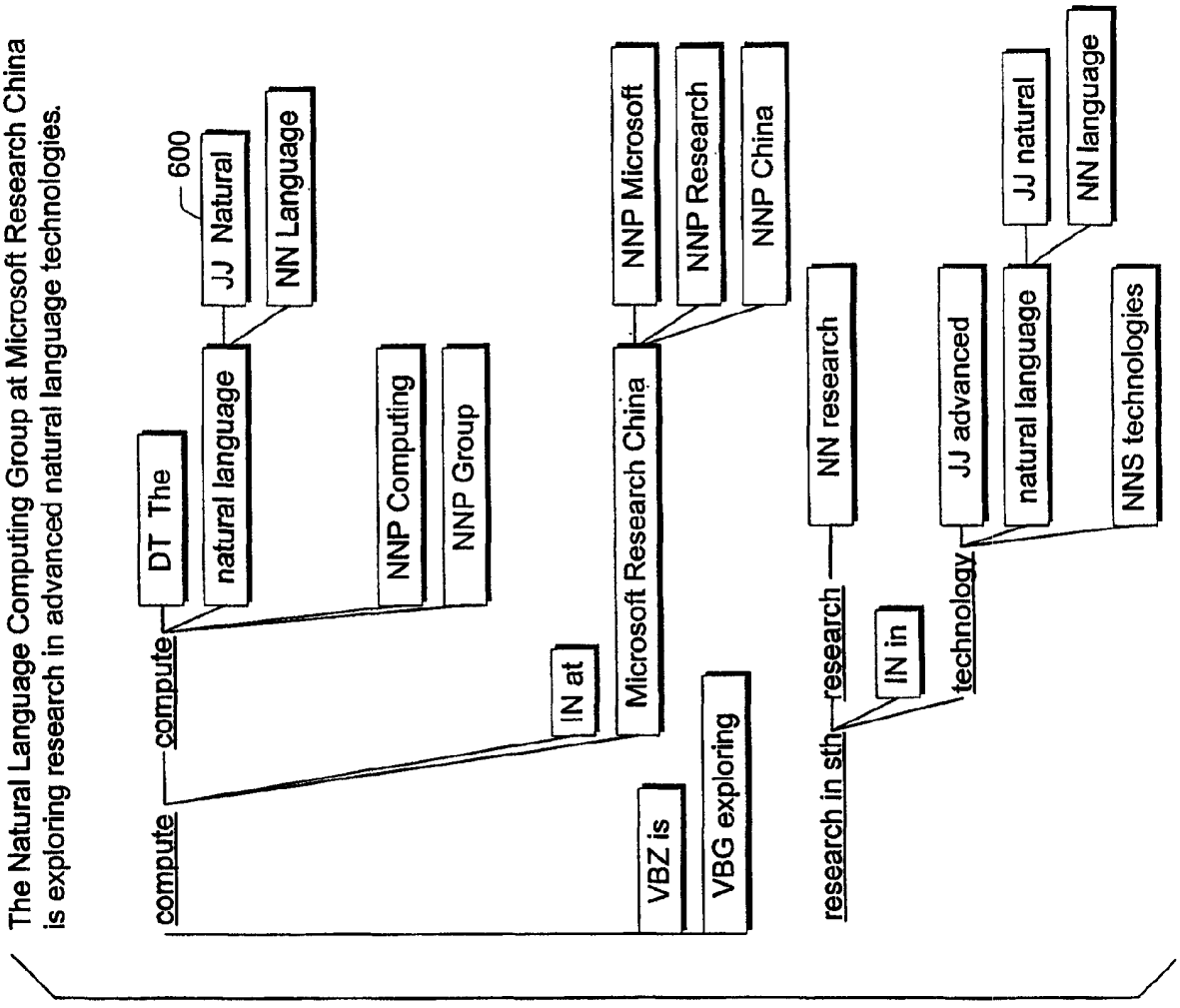

The Natural Language Computing Group at Microsoft Research China is exploring research in advanced natural language technologies.
compute
compute
DT The
600
JJ Natural
natural language
NN Language
NNP Computing
NNP Group
IN at
Microsoft Research China
NNP Microsoft
NNP Research
NNP China
VBZ is
VBG exploring
research in sth
research
NN research
IN in
technology
JJ advanced
natural language
JJ natural
NN language
NNS technologies

图7

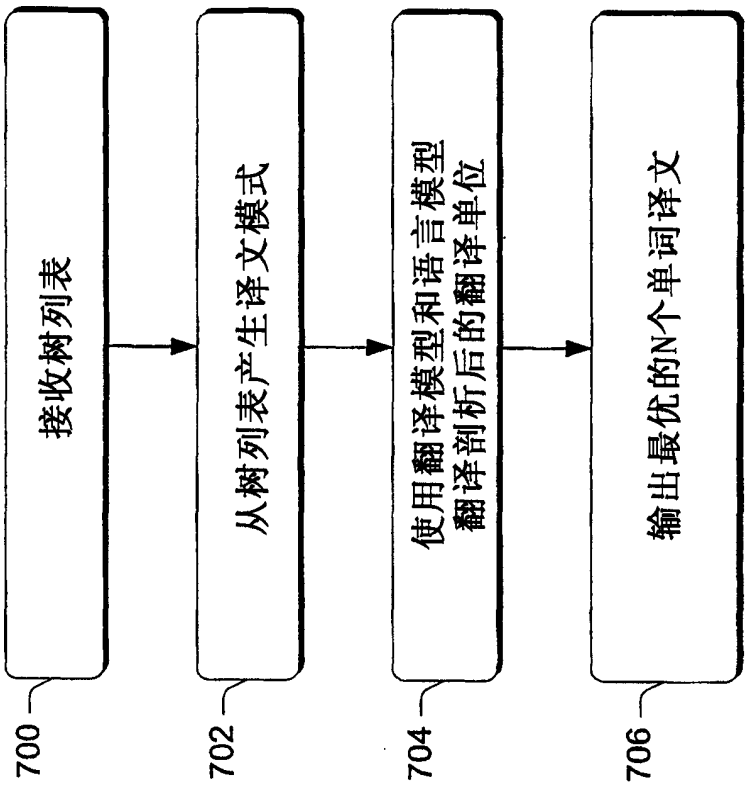
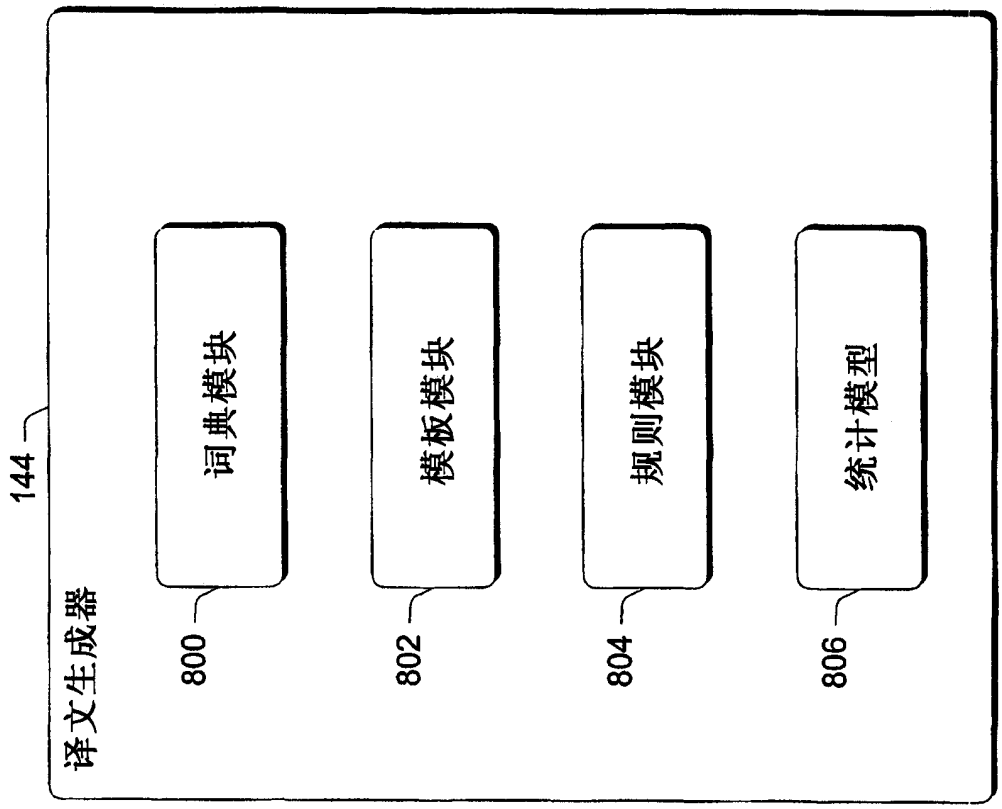

接收树列表
从树列表产生译文模式
使用翻译模型和语言模型翻译剖析后的翻译单位
输出最优的N个单词译文
700
702
704
706

图8



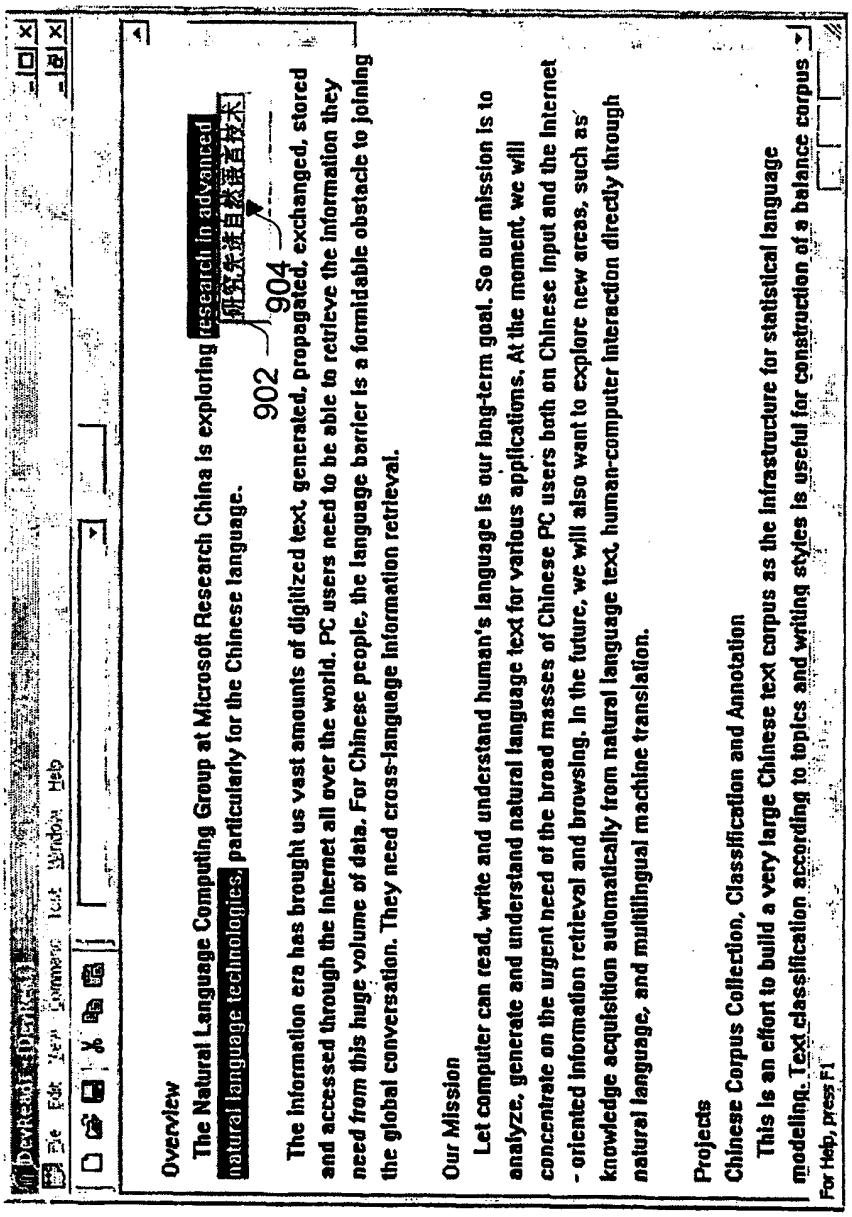

900
Overview
The Natural Language Computing Group at Microsoft Research China is exploring natural language technologies, particularly for the Chinese language.
902
research in advanced
904
The information era has brought us vast amounts of digitized text, generated, propagated, exchanged, stored and accessed through the Internet all over the world. PC users need to be able to retrieve the information they need from this huge volume of data. For Chinese people, the language barrier is a formidable obstacle to joining the global conversation. They need cross-language information retrieval.
Our Mission
Let computer can read, write and understand human's language is our long-term goal. So our mission is to analyze, generate and understand natural language text for various applications. At the moment, we will concentrate on the urgent need of the broad masses of Chinese PC users both on Chinese input and the Internet - oriented information retrieval and browsing. In the future, we will also want to explore new areas, such as knowledge acquisition automatically from natural language text, human-computer interaction directly through natural language, and multilingual machine translation.
Projects
Chinese Corpus Collection, Classification and Annotation
This is an effort to build a very large Chinese text corpus as the infrastructure for statistical language modeling. Text classification according to topics and writing styles is useful for construction of a balance corpus
For Help, press F1

图9

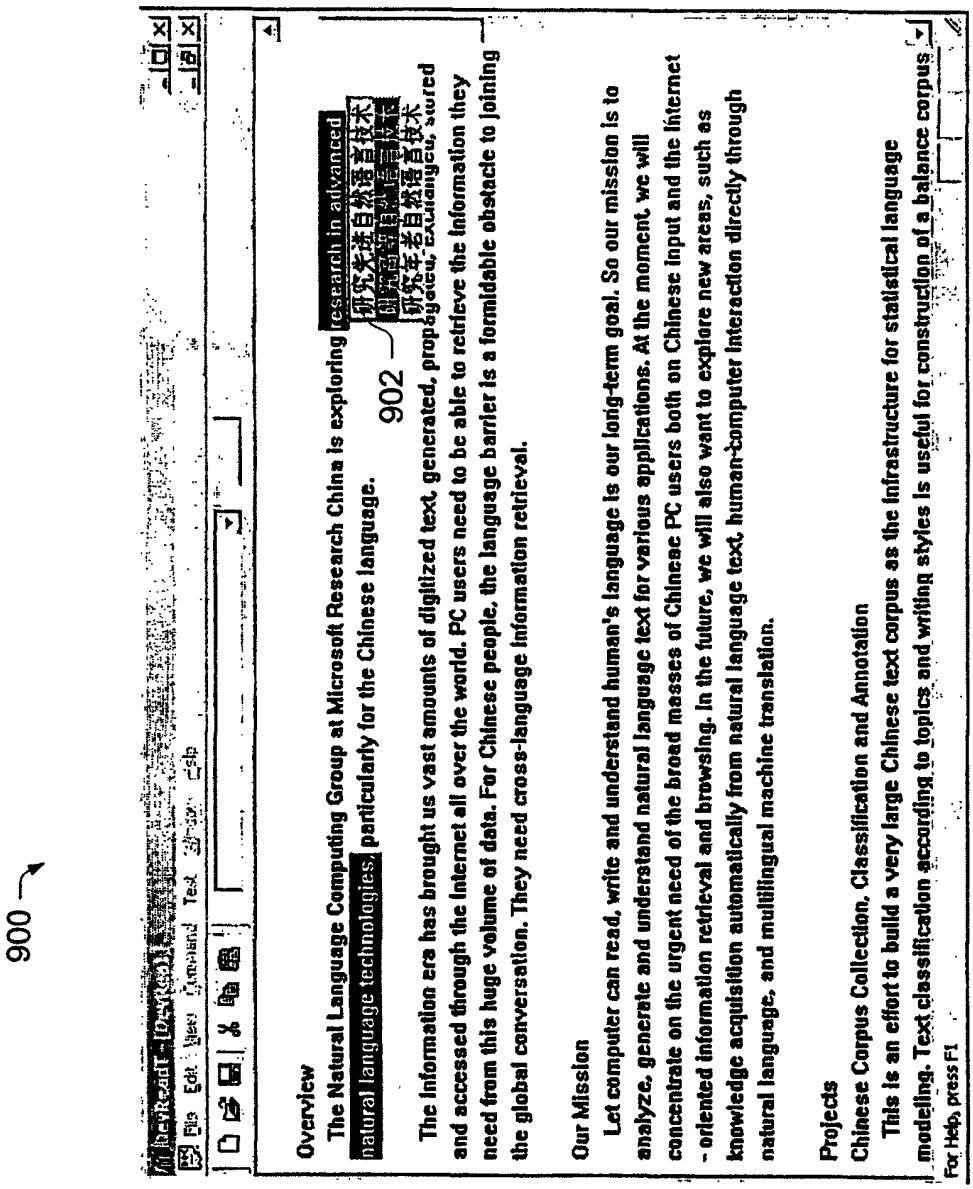

900
Overview
The Natural Language Computing Group at Microsoft Research China is exploring natural language technologies, particularly for the Chinese language.
902
研究者将自然语言技术
The information era has brought us vast amounts of digitized text, generated, propagated, exchanged, stored and accessed through the Internet all over the world. PC users need to be able to retrieve the information they need from this huge volume of data. For Chinese people, the language barrier is a formidable obstacle to joining the global conversation. They need cross-language information retrieval.
Our Mission
Let computer can read, write and understand human's language is our long-term goal. So our mission is to analyze, generate and understand natural language text for various applications. At the moment, we will concentrate on the urgent need of the broad masses of Chinese PC users both on Chinese input and the Internet - oriented information retrieval and browsing. In the future, we will also want to explore new areas, such as knowledge acquisition automatically from natural language text, human-computer interaction directly through natural language, and multilingual machine translation.
Projects
Chinese Corpus Collection, Classification and Annotation
This is an effort to build a very large Chinese text corpus as the infrastructure for statistical language modeling. Text classification according to topics and writing styles is useful for construction of a balance corpus
For Help, press F1

图10

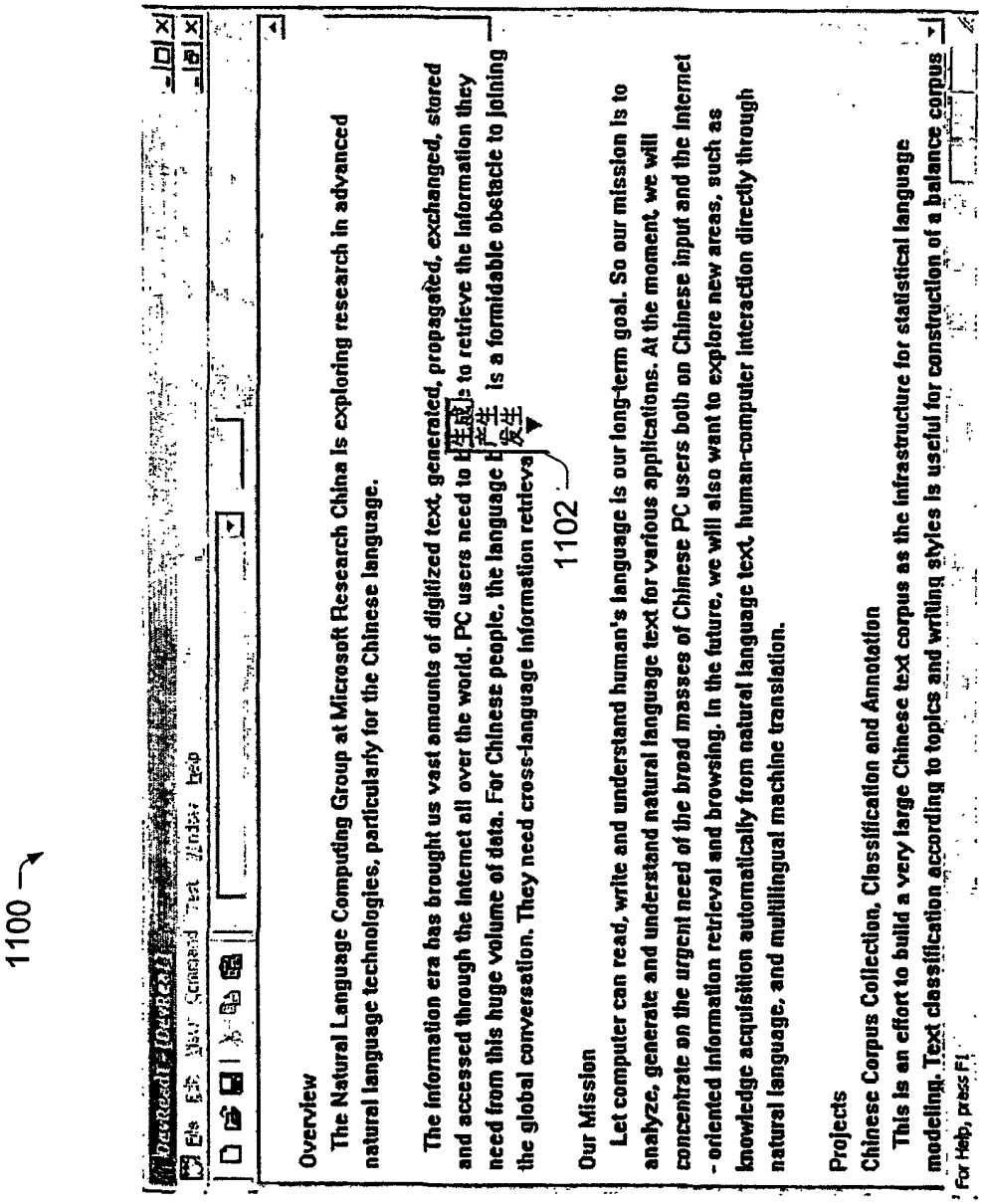

1100
Overview
The Natural Language Computing Group at Microsoft Research China is exploring research in advanced natural language technologies, particularly for the Chinese language.
The information era has brought us vast amounts of digitized text, generated, propagated, exchanged, stored and accessed through the Internet all over the world. PC users need to retrieve the information they need from this huge volume of data. For Chinese people, the language is a formidable obstacle to joining the global conversation. They need cross-language information retrieval
1102
Our Mission
Let computer can read, write and understand human's language is our long-term goal. So our mission is to analyze, generate and understand natural language text for various applications. At the moment, we will concentrate on the urgent need of the broad masses of Chinese PC users both on Chinese input and the Internet - oriented information retrieval and browsing. In the future, we will also want to explore new areas, such as knowledge acquisition automatically from natural language text, human-computer interaction directly through natural language, and multilingual machine translation.
Projects
Chinese Corpus Collection, Classification and Annotation
This is an effort to build a very large Chinese text corpus as the infrastructure for statistical language modeling. Text classification according to topics and writing styles is useful for construction of a balance corpus

图11

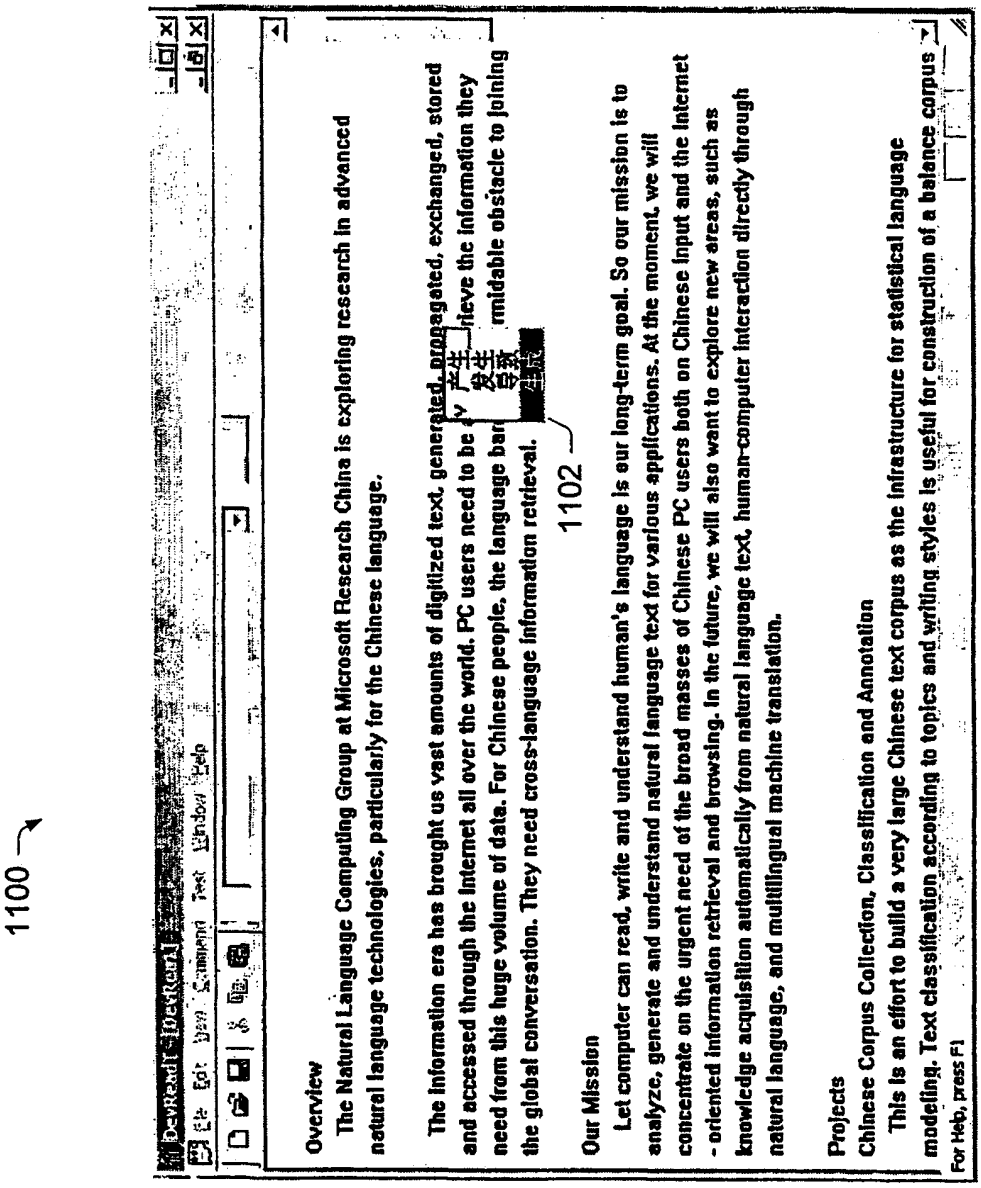

1100
DevStudio - [DevStudio]
File Edit View Command Tool Window Help
Overview
The Natural Language Computing Group at Microsoft Research China is exploring research in advanced natural language technologies, particularly for the Chinese language.
The information era has brought us vast amounts of digitized text, generated, propagated, exchanged, stored and accessed through the Internet all over the world. PC users need to be able to achieve the information they need from this huge volume of data. For Chinese people, the language barrier is a formidable obstacle to joining the global conversation. They need cross-language information retrieval.
1102
Our Mission
Let computer can read, write and understand human's language is our long-term goal. So our mission is to analyze, generate and understand natural language text for various applications. At the moment, we will concentrate on the urgent need of the broad masses of Chinese PC users both on Chinese input and the Internet - oriented information retrieval and browsing. In the future, we will also want to explore new areas, such as knowledge acquisition automatically from natural language text, human-computer interaction directly through natural language, and multilingual machine translation.
Projects
Chinese Corpus Collection, Classification and Annotation
This is an effort to build a very large Chinese text corpus as the infrastructure for statistical language modeling. Text classification according to topics and writing styles is useful for construction of a balance corpus
For Help, press F1

图12

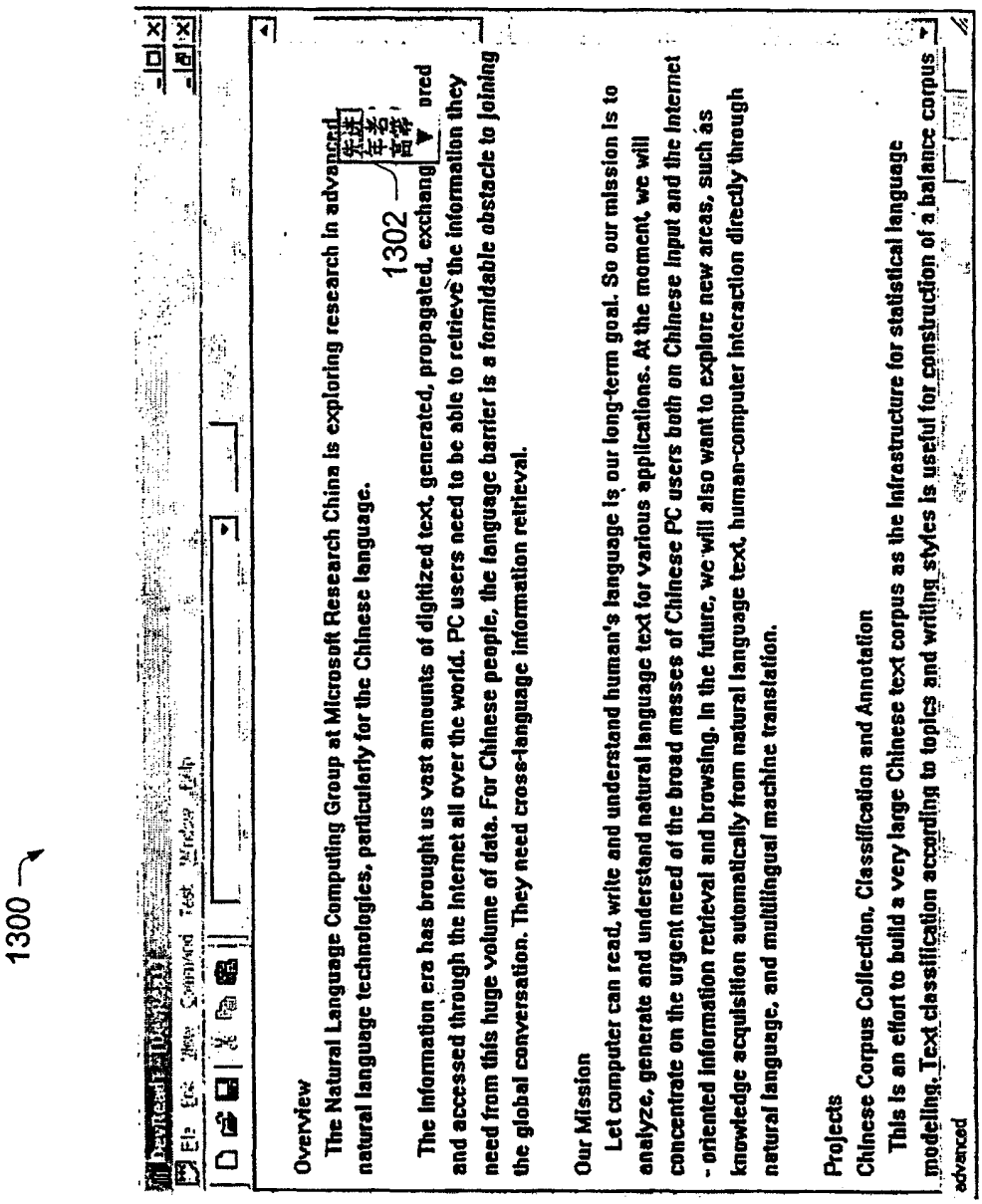

1300
1302
Overview
The Natural Language Computing Group at Microsoft Research China is exploring research in advanced natural language technologies, particularly for the Chinese language.
The information era has brought us vast amounts of digitized text, generated, propagated, exchanged and accessed through the Internet all over the world. PC users need to be able to retrieve the information they need from this huge volume of data. For Chinese people, the language barrier is a formidable obstacle to joining the global conversation. They need cross-language information retrieval.
Our Mission
Let computer can read, write and understand human's language is our long-term goal. So our mission is to analyze, generate and understand natural language text for various applications. At the moment, we will concentrate on the urgent need of the broad masses of Chinese PC users both on Chinese input and the Internet - oriented information retrieval and browsing. In the future, we will also want to explore new areas, such as knowledge acquisition automatically from natural language text, human-computer interaction directly through natural language, and multilingual machine translation.
Projects
Chinese Corpus Collection, Classification and Annotation
This is an effort to build a very large Chinese text corpus as the infrastructure for statistical language modeling. Text classification according to topics and writing styles is useful for construction of a balance corpus.

图13