[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 03807027.8

[43] 公开日 2005 年 7 月 27 日

[11] 公开号 CN 1647068A

[22] 申请日 2003.3.28 [21] 申请号 03807027.8

[30] 优先权

[32] 2002. 3.28 [33] US [31] 60/368,851

[86] 国际申请 PCT/US2003/009749 2003.3.28

[87] 国际公布 WO2003/083709 英 2003.10.9

[85] 进入国家阶段日期 2004.9.27

[71] 申请人 南加利福尼亚大学

地址 美国加利福尼亚州

[72] 发明人 P·克伊赫恩 K·克奈特

[74] 专利代理机构 上海专利商标事务所有限公司

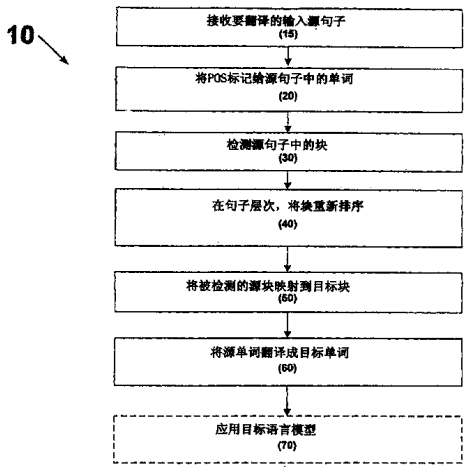
代理人 李家麟

权利要求书 2 页 说明书 6 页 附图 4 页

[54] 发明名称 统计机器翻译

[57] 摘要

一种方法包括检测第一语言的源串中的句法块，将句法标签分配给源串中被检测的句法块，将源串中被检测的句法块映射到第二语言的目标串中的句法块，所述映射基于所分配的句法标签，以及将源串翻译成第二语言的可能翻译。



ISSN1008-4274

1. 一种方法，其特征在于，包括：

检测第一语言的源串中的句法块；

将句法标签分配给源串中被检测的句法块；

将源串中被检测的句法块映射到第二语言的目标串中的句法块，所述映射基于所分配的句法标签；以及

将源串翻译成第二语言的可能的翻译。

2. 如权利要求 1 所述的方法，其特征在于，分配句法标签的步骤包括根据标记给源串中的单词的词性标识符进行分配。

3. 如权利要求 1 所述的方法，其特征在于，进一步包括限定源串中被检测的句法块和目标串中的块之间的连接。

4. 如权利要求 3 所述的方法，其特征在于，限定连接的步骤包括根据块映射表确定连接，该块映射表使用根据句法块标签预先限定的连接。

5. 如权利要求 3 所述的方法，其特征在于，限定连接的步骤包括限定来自源串的被检测块与目标串中至少两个非相邻块之间的连接。

6. 如权利要求 3 所述的方法，其特征在于，限定连接的步骤包括限定来自源串的至少两个被检测的块到目标串中的单个块的连接。

7. 如权利要求 1 所述的方法，其特征在于，翻译包括纳入与源串中的单个单词相对应的目标串中的至少两个单词。

8. 如权利要求 1 所述的方法，其特征在于，翻译包括翻译短语。

9. 如权利要求 1 所述的方法，其特征在于，进一步包括：

将语言模型应用于源串，该语言模型基于目标串的语言。

10. 如权利要求 1 所述的方法，其特征在于，进一步包括：

确定所述映射的概率。

11. 如权利要求 1 所述的方法，其特征在于，翻译包括在目标串中插入至少一个 NULL 单词。

12. 一种包括含机器可执行指令的机器可读介质的制品，该指令用于使得机器：

检测第一语言的源串中的句法块；

将句法标签分配给源串中的被检测句法块；

使源串中被检测的句法块与第二语言串中的句法块对准，所述映射基于所分配的句法标签；以及

将源串的每个单词翻译成与第二语言的可能翻译相对应的第二单词。

13. 如权利要求 12 所述的制品，其特征在于，用于使得机器分配句法标签的指令包括用于根据标记给源串中的单词的词性标识符使得机器分配句法标签的指令。

14. 如权利要求 12 所述的制品，其特征在于，进一步包括指令，它用于使得机器限定源串中被检测的句法块与目标串中的块之间的连接。

15. 如权利要求 14 所述的制品，其特征在于，用于使得机器限定连接的指令包括用于使得机器根据块映射表确定连接的指令，该块映射表使用基于句法块标签的预先限定的连接。

16. 如权利要求 14 所述的制品，其特征在于，用于使得机器限定连接的指令包括用于使得机器限定源串的被检测块与目标串中的至少两个非相邻块之间的连接的指令。

17. 如权利要求 3 所述的制品，其特征在于，用于使得机器限定连接的指令包括用于使得机器限定源串的至少两个被检测块到目标串中的单个块的连接的指令。

18. 如权利要求 12 所述的制品，其特征在于，用于使得机器翻译的指令包括用于使得机器包括与源串中的单个单词相对应的目标串中的至少两个单词的指令。

19. 如权利要求 12 所述的制品，其特征在于，用于使得机器翻译的指令包括用于使得机器翻译短语的指令。

20. 如权利要求 12 所述的制品，其特征在于，进一步包括指令，它们用于使得机器：

将语言模型应用于源串，该语言模型基于目标串的语言。

21. 如权利要求 12 所述的制品，其特征在于，进一步包括指令，它们用于使得机器：

确定所述映射的概率。

22. 如权利要求 12 所述的制品，其特征在于，用于使得机器翻译的指令包括用于使得机器在目标串中插入至少一个 NULL 单词的指令。

统计机器翻译

相关申请对照

本申请要求 2002 年 3 月 28 日提交的美国临时申请序号 No.60/368851 的优先权，在此全文并入以供参考。

发明起因

本申请中描述的研究和开发在准许号 N66001-00-1-8914 下由 DARPA-ITO 支持。美国政府可以具有所要求的本发明的某些权利。

背景

机器翻译(MT)是自动翻译，例如使用计算机系统从第一语言(“源”语言)翻译成另一种语言(“目标”语言)。执行 MT 过程的系统被称为将源语言“解码”成目标语言。从最终用户的观点来看，MT 过程是相对直接的。MT 过程接收作为输入的源句子(或者单词的“串”)并在处理该输入句子后输出目标语言的翻译句子。

一种类型的 MT 过程被称作统计 MT 解码器。常规统计 MT 解码器可以包括语言模型(LM)和翻译模型(TM)。

概述

根据本发明的一个方面，一种方法包括在第一语言的第一串中检测句法块，将句法标签分配给第一串中被检测的句法块，将第一串中被检测的句法块与第二语言串中的句法块对准，所述对准基于被分配的句法标签，以及将来自第一串的每个单词翻译成对应于第二语言中可能的翻译的第二单词。

还可以包括一个或多个以下特点。根据标记给至少两个单词的词性标识符来分组来自第一串的这至少两个单词。限定源串中被检测的句法块和第二串中的块之间的连接。根据块映射表确定连接，该块映射表使用基于句法块标签的预先限定的连接。限定第一串的被检测块与目标串中的至少两个非相邻块之间的连接。限定源串的至少两个被检测块到目标串中的单个块的连接。

附图概述

图 1 是语言统计翻译(LST)过程的流程图。

图 2 示出实例性的源和目标句子。

图 3 示出与图 2 的句子相对应的句子层次块重新排序表。

图 4 示出与图 2-3 相对应的块映射对准表。

图 5 示出与图 1-4 相对应的单词翻译。

图 6 示出 LST 过程的第二实施例。

具体实施方式

这里描述的统计 MT 系统可以被模拟成三个分开的部分：(1)将概率 P(e)分配给单词的任何目标串的语言模型(LM)，(2)将概率 P(f|e)分配给目标和源串的任何对的翻译模型(TM)，以及(3)根据 LM 和 TM 的被分配概率确定翻译的解码器。

常规 MT 系统可以通过进行一连串基于单词的判定将源句子翻译成目标句子。基于单词的判定可以包括翻译判定，其中每个源单词都被翻译成目标单词。也可以为每个被翻译单词执行映射(“对准”)判定，例如根据被翻译源单词的被确定的丰度将多个源单词映射到单个目标单词。也可以执行重新排列(“变形”)判定，例如将源句子的单词序列重新排列成相应的目标句子中的被翻译的单词。翻译、映射和变形判定基于在翻译过程期间确定的权重概率。

某些源句子引起翻译的挑战，它们不能通过常规基于单词的 MT 系统得到良好的处理。例如，翻译挑战包括短语的翻译、出于句法原因重新组织句子以及将非相邻单词翻译成目标句子中的单个单词或短语。

图 1 描述了一种语言统计翻译模型(LST)过程 10，它包括接收要翻译源句子(15)，为源句子中的每个源单词分配“词性”(POS)标记(20)，以及检测源句子中包含的句法“块”(30)。LST 过程 10 还包括动作(40)、(50)和(60)，它们部分基于被分配的 POS 标记和/或被检测的句法块。过程 10 中 POS 标记和/或句法块的使用允许改善源到目标句子的翻译，以及部分改善前述翻译挑战的串翻译。

POS 标记涉及表示单词类型的识别符号，例如“VVFIN”符号可以被标记为限定动词。可以用于过程 10 中的一组实例性 POS 标记被称作“Penn Treebank Tag set”，并描述于 Mitchell P. Marcus, Beatrice Santorini 和 Mary Ann Marcinkiewicz: “Building a Large Annotated Corpus of English: The Penn Treebank”, in Computational Linguistics, 卷 19, 号 2(1993 年 6 月), pp. 313—330(Special Issue on Using Large Corpora)，在此全文并入以供参考。

分块涉及非递归(non-recursive)的动词、名词、介词或句子中的其它短语的

分组。分块可以包括检测源句子中的分组和目标句子中块组合的输出。在 Abney, S.(1991)“Parsing by chunks(通过块分析)”Robert Berwick, Steven Abney 和 Carol Tenny: Principle-based Parsing(基于原理的分析).Kluwer Academic Publishers 中讨论了分块的概念。

仍旧参考图 1，LST 过程 10 包括接收要翻译的输入源句子(15)，用 POS 标记来标记源句子中的每个单词(20)，检测每个源句子中的句法块(例如，短语)(30)，句子层次块的重新排序(40)，将被检测的源块映射到目标句子中的目标块(50)，以及将每个单词从源翻译到目标句子(60)。可以采用可选的目标语言模型(70)进一步改进通过动作(60)产生的单词翻译。

图 2 示出实例性的源句子 100，其中每个单词都具有动作(20)中生成的相关 POS 标记 110-116，以及动作(30)中生成的被检测句法块 101-105。被检测的块 101-105 还分别包括句法标签，例如“N,V,N,V 和!”。句法标签涉及用于被检测块的句子的句法部分，例如，“N”可以表示基本名词短语，“V”可以表示动词复合，“P”可以表示基本介词短语，“A”可以表示形容词，“F”可以表示功能词，而“!”可以表示标点。

句子层次块的重新排序(40)限定每个源块 101-106 和将包含于目标句子 150 中的相应目标块 130-134 之间的连接 120-125。在许多情况下，相对于源块重新排序目标块。该重新排序可以基于限定被检测句法块和目标句子中相应的句法块之间可能连接的模板。连接可以是单值或多值的(例如，一对一、多对多、或者一对多等等)。图 3 示出块连接表 160，它表示源块 101-105 与目标块 130-134 之间的连接 120-125，与图 2 中示出的那些相对应。

图 4 示出块映射表 170、180、190 和 200，它们表示通过过程 10 的活动(50)产生的块映射，如应用于实例性句子 100 的那样。块映射涉及每个源块到目标块的对准并可以按照源块中的单词和目标块中的单词的 POS 标记参考。例如，如表 170 所示，源 POS 标记 110(“ART”)和 111(“NN”)被对准到目标 POS 标记 140(“DT”)和 141(“NNP”)。块映射可以将多个块(“复合块”)映射到单个块或其它复合块。例如，如表 190 所示，源块 103 被对准到包含目标块 130 和 131 的复合块。来自源句子 110 的非相邻块可以被组合成单个块，例如，如表 180 所示，将块 102 和 104 组合成目标块 132。

如前所述，可以用被分配的句法块标签来“标注”每个复合块。该标注可以允许改善句子层次的块重新排序，因为句法标签可以识别它们在句子中的句法作用。

随后，过程 10 将来自源语言句子的源单词翻译成目标语言句子的单词(60)。

可以部分根据分配给相应源单词的词性(通过块映射选择)来确定单词翻译，例如限制与分配的 POS 标记相对应的单词的选择。图 5 描述了来自过程 10 的活动(60)的执行，例如描述了与图 1-4 所示的实例相对应的单词翻译。

在实施例中，代替通过单个单词翻译生成目标语言单词，可以通过准确的短语查找翻译复合块。更详细地，如果确定整个源块是已知短语，整个源块就可以被翻译为已知短语。例如，如图 2 所示，如果源块 103“der Agrarausshuss”中包含的单词是已知短语，则可以将其直接翻译为目标块 130-131“the sub-committee for agriculture”中的单词。准确的短语查找允许使用惯用短语的翻译，这是基于单词的翻译所不容易翻译的。

过程 10 可以包括可选的目标语言模型(70)，它被执行来提供对目标句子的附加的流畅性改善。

过程 10 的数学公式化

可以数学地模拟 LST 过程 10 的操作，例如基于一组概率判定来模拟。以下过程 10 的数学模型包括按照噪声信道模型(noisy channel model)的公式化。更详细地，这意味着代替直接估计 p(e|f)(例如，用于输入串 f 的最佳翻译 e)，将贝斯法则应用于使 p(f|e) x p(e)最大化。因此，这将模型分成两个部分：翻译部分 p(f|e)和语言模型 p(e)。对于语言部分，可以使用三字母组语言模型。

翻译部分被分解成句子层次重新排序(SLR)、块映射(CM)和单词翻译(W)，并用以下的概率等式模拟：

$$P(f|e)=p(SLR|e)\times\Pi_i p(CM_i|e,SLR)\times\Pi_j p(W_{ij}|CM_i,SLR,e)$$

由于 POS 标记和分块是确定性的，e 不仅表示目标串的单词，还表示它们的 POS 和分组为块。可以使用模板执行句子层次块重新排序(SLR)和块内的单词重新排序(CM)，例如使用表示来自图 3 和 4 所示的表的信息的模板。可以使用逐字翻译表来完成单词翻译(W)。

由于稀少的数据，直接应用以上三个概率等式是有问题的。因此，可以如下地简化三个附条件的概率分配：

p(SLR)可以仅以每个目标块标签序列为条件；

$p(CM_i)$可以仅以有关源和目标块标签，以及目标 POS 标记为条件；

$p(W_{ij})$可以仅以有关目标 POS 标记和单词为条件。

块映射中的每个单词对准以单词翻译概率为因素。未对准的源单词以概率 $p(f_k|ZFERT,f_{posk})$为因素。未对准的目标单词以概率 $p(NULL|e_k,f_{posk})$为因素。

代替将块映射分解成单词翻译，可以执行直接短语查找，它是通过以下等式模拟的：

$p(W_{i1}, \ldots, W_{in} | CM_i, SLR, e)$

可以使用所谓的相似文集(parallel corpus)方法确定用于单词对准的参数，在该方法中，源语言串中的文本(第一文集)被对准到目标语言串中的被翻译文本(第二文集)。这些对准建立了源串中的源单词和目标串之间的对应。相似文集的两侧也可以被 POS 标记或被分块。

可以使用相似文集方法确定块映射，例如如果源块和目标块包含相互对准的源单词和目标单词，则可以连接这两个块。没有包含对准单词的块可以根据一组规则被附着到其它块，例如如果未对准，副词被附加到以下的动词块，或者如果未对准，逗号被附着到以下的功能词，等等。

随后可以在任何块对准上执行传递闭包(transitive closure)，例如使用以下的规则组：如果块 f_i 与 e_x 对准，f_j 与 e_x 对准，且块 f_i 与 e_y 对准，则块 f_j 就被认为与 e_y 对准，即使它们没有包含任何相互对准的单词。传递闭包确保源句子和目标句子中复合块之间的一对一映射。

根据以上公式对应相似文集允许对单词翻译(包括 $p(f_k|ZFERT, f_{posk})$ 和 $p(NULL|e_k, f_{posk})$)、复合块映射以及句子层次重新排序的要收集的统计。随后，通过最大可能性估计收集附条件的概率分配。由于用于准确的短语查找的数据是高度有噪声的，可以使概率平滑。

在实施例中，模型的翻译部分(例如，“解码”)可以以两个步骤执行：第一，生成用于每个句子层次块重新排序的句子层次模板(SLT)。第二，从左向右每次一个单词地构成目标翻译。对于每个给定的源块序列，为最高的 n 个 SLT 重复以上内容。最终，选择具有总的最好分数的翻译作为系统输出。

对于给定句子层次模板(SLT)的目标句子的构建可以通过使用动态编程的 Viterbi 查找实现。在这种情况中，按需要选择块映射模板。随后，使用逐字翻译表和语言模型填充单词空位。在每个复合块的末端，丢弃关于使用哪个块映射模板的信息。在某些实施中，目标串的构建可以包括 NULL 单词的插入。

但是，对于每个部分翻译(或假设)，维持以下信息：

-创建的最近的两个单词(语言模型需要)；

-如果未完成，当前块映射模板；

-当前分数(‘分数’涉及部分翻译判定、块映射判定等的组合的概率的乘积)；

-到最佳路径的向后指针；

-最后块的位置；

-块内创建的最后单词的位置；

-“堆叠的块映射模板”

堆叠的块映射模板涉及当分离的复合块被填充到目标翻译中时所需的信息：例如，如果 SLT 要求创建“V+P”块，其中在“V”和“P”之间具有附加内容。在这种情况中，关于所选择的块映射模板的信息必须维持于“V”和“P”之间，直到它被完全填充。

目标句子中任何给定位置处假设空间的复合性可以表示为 $O(V^2C^{1+s})$，其中 V 是词汇大小，C 是可应用的块映射模板的数量，且 s 是堆叠的块映射模板的数量。

可以通过将翻译限制于目标语言中的邻接复合块来简化模型，它消除了对堆叠的块映射模板的需要。在任何给定位置处，这将复合性等式简化为 $O(V^2C)$。关于句子长度，这还确保解码具有线性的复合性。

图 6 示出 LST 过程 100 的实施例，它根据以上讨论的等式和公式模拟。在该实例中，LST 过程 100 包括环(135、140、150、160 和 170)，对于 n 个不同句子层次模板，该环重复 n 次。

已描述了大量实施例。然而，将理解，可以进行各种修改而不背离本发明的精神和范围。例如，翻译成多个目标单词的源单词会引起块映射错误。通过添加丰度特点或者进一步预先处理复合名词可以避免或减少这种类型的错误。作为另一个实例，通过使用概率单词翻译方法(例如，“T-Table”翻译方法)可以执行单词翻译。作为另一个实例，没有足够的统计来可靠地估计句子层次模板(SLT)。因此，可以使用其它估计，例如从句层次模板，或者使用将句子层次块翻译步骤分解成大量块分段和翻译判定的方法。

因此，其它实施例也在以下权利要求书的范围内。

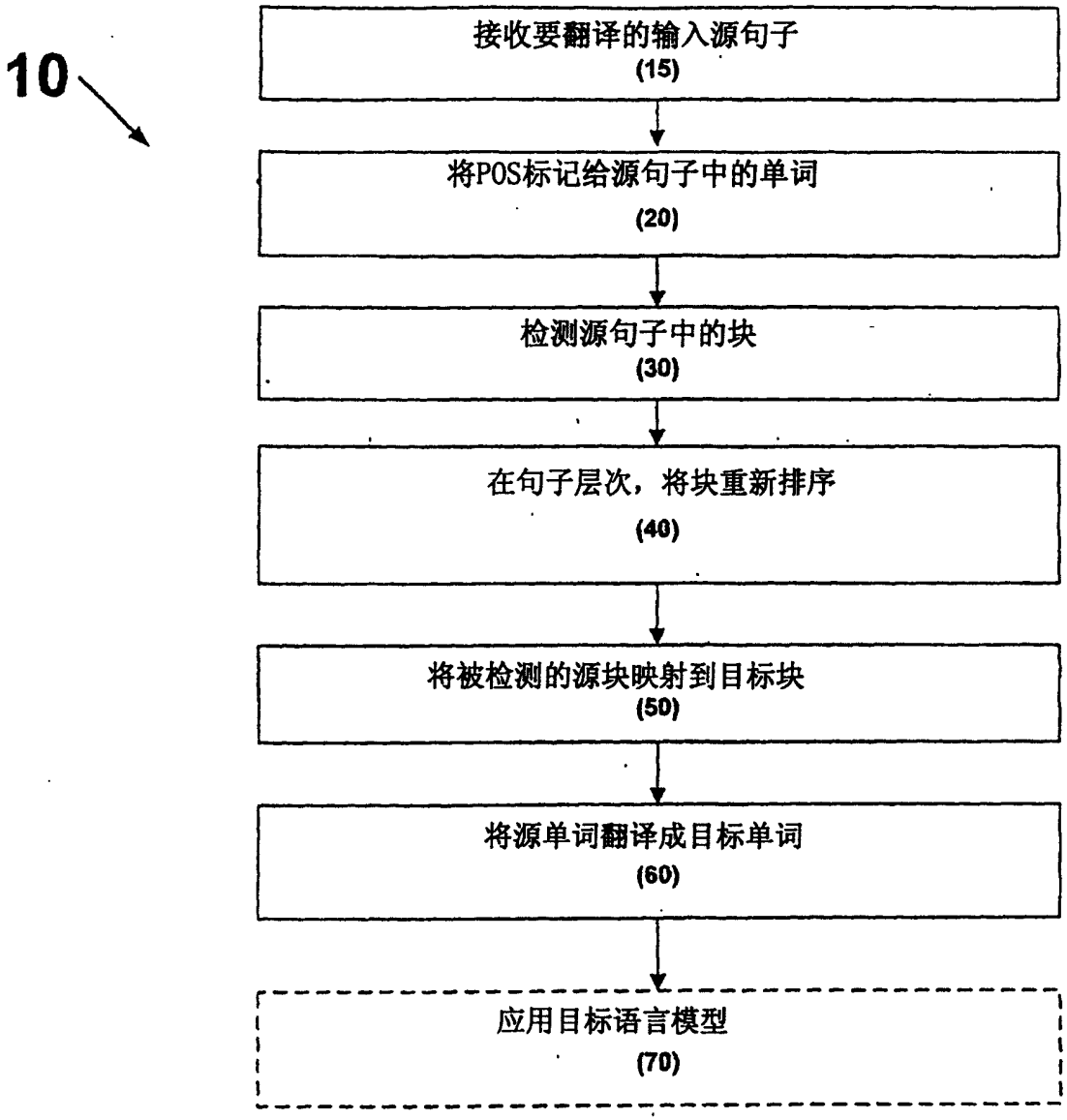

10
接收要翻译的输入源句子
(15)
将POS标记给源句子中的单词
(20)
检测源句子中的块
(30)
在句子层次，将块重新排序
(40)
将被检测的源块映射到目标块
(50)
将源单词翻译成目标单词
(60)
应用目标语言模型
(70)

图 1

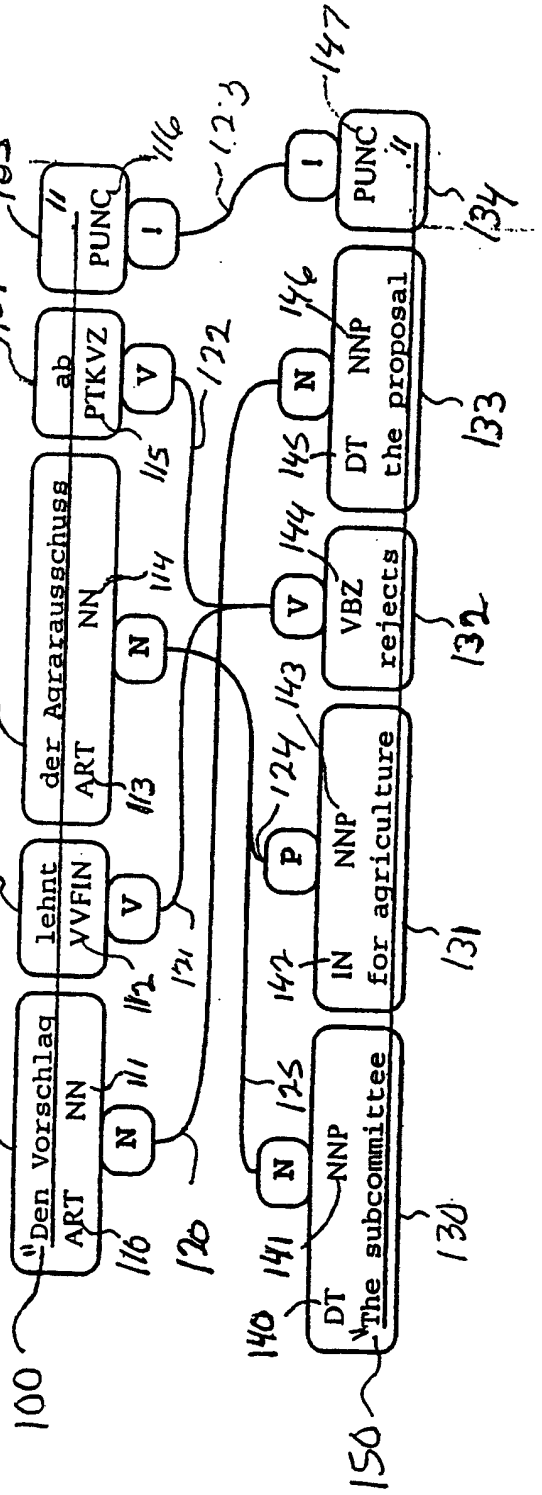

"Den Vorschlag
ART NN
der Agrarausschuss
ART NN
lehnt
VVFIN
ab
PTKVZ
"
PUNC
.
PUNC
the proposal
DT NNP
rejects
VBZ
for agriculture
IN NNP
The subcommittee
DT NNP
100
101
102
103
104
105
111
112
113
114
115
116
120
121
122
123
124
125
130
131
132
133
134
140
141
142
143
144
145
146
147
150

图 2

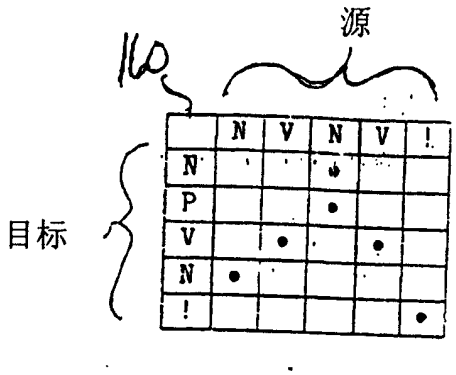


图 3

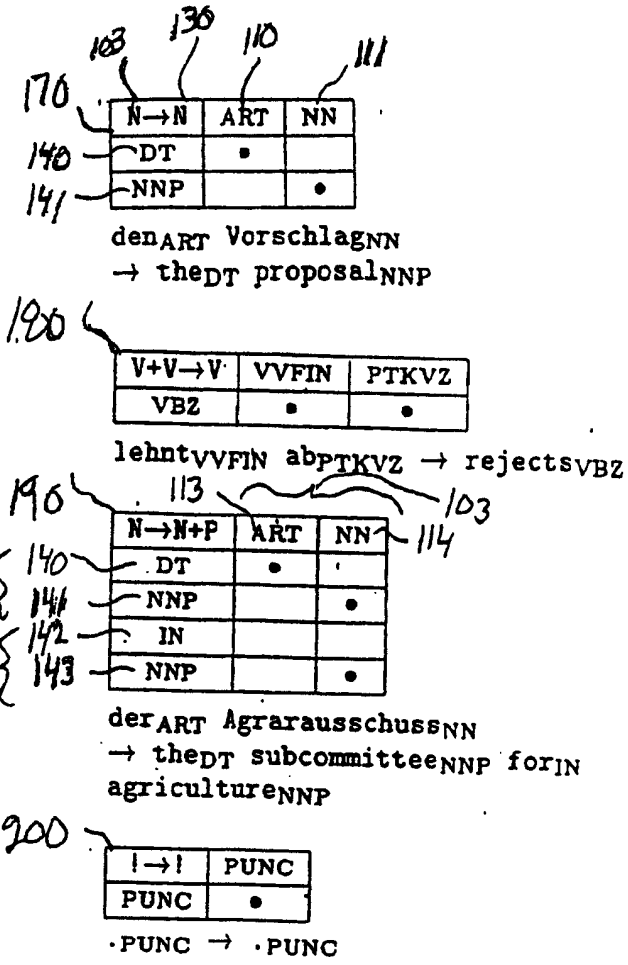


图 4

der	→	DT	the
Agrarausschuss	→	NNP	subcommittee
NULL	→	IN	for
Agrarausschuss	→	NNP	agriculture
lehnt, ab	→	VBZ	rejects
den	→	DT	the
vorschlag	→	NNP	proposal

图 5

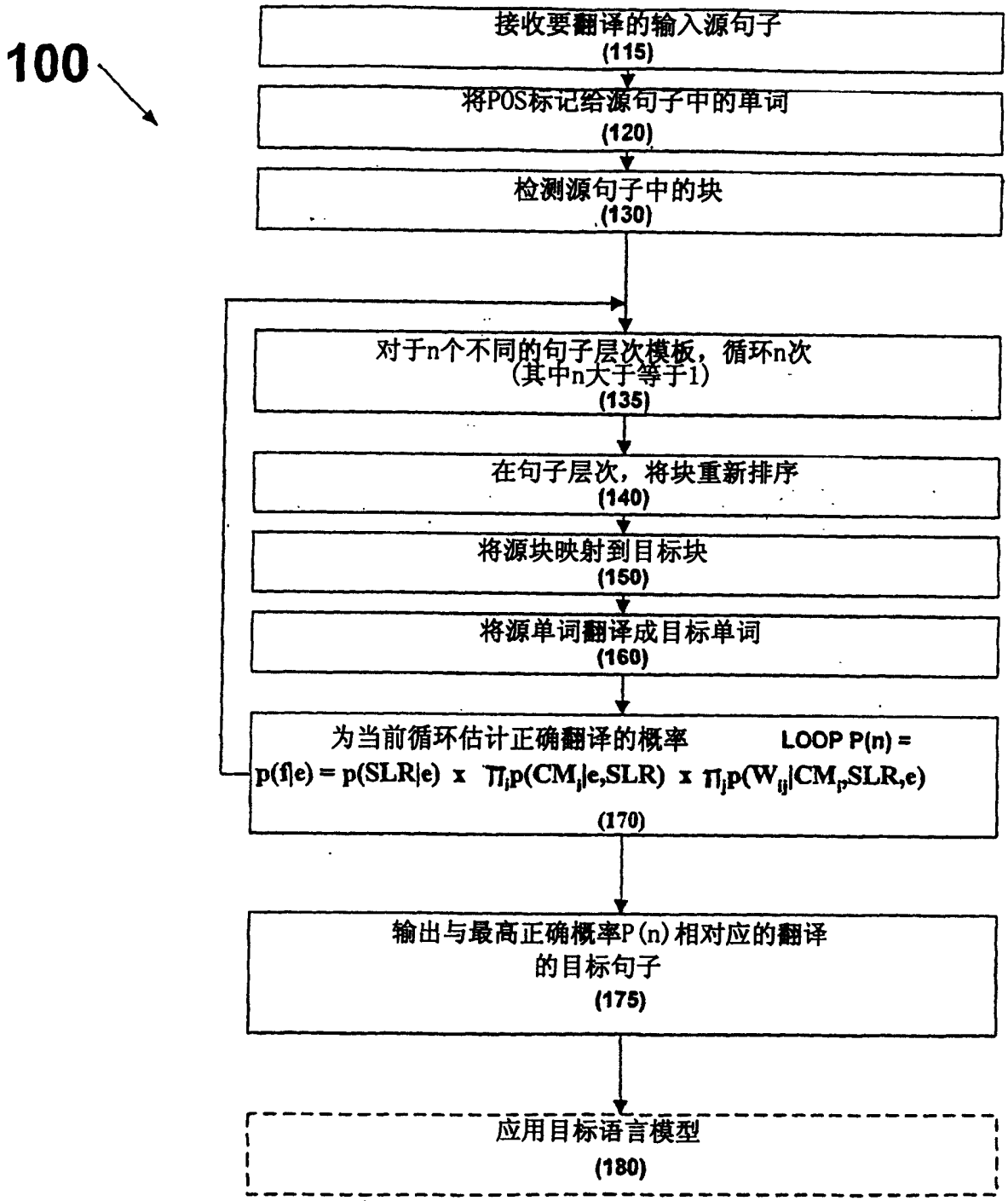

100
接收要翻译的输入源句子
(115)
将POS标记给源句子中的单词
(120)
检测源句子中的块
(130)
对于n个不同的句子层次模板，循环n次
(其中n大于等于1)
(135)
在句子层次，将块重新排序
(140)
将源块映射到目标块
(150)
将源单词翻译成目标单词
(160)
为当前循环估计正确翻译的概率 LOOP P(n) =
p(f|e) = p(SLR|e) x ∏ip(CMi|e,SLR) x ∏jp(Wij|CMi,SLR,e)
(170)
输出与最高正确概率P(n)相对应的翻译
的目标句子
(175)
应用目标语言模型
(180)

图 6