有限语料汉蒙统计机器翻译调序方法研究*

陈雷，李淼，张健，曾伟辉
(中国科学院合肥智能机械研究所，安徽 合肥 230031)

摘要：自统计机器翻译技术出现以来，调序一直是语序差异显著的语言对互译系统中的关键问题，基于大规模语料训练的调序方法得到了广泛研究。目前汉蒙双语语料资源十分有限，使得现有的依赖于大规模语料和语言学知识的调序方法难以取得良好效果。本文对已有的相关研究进行了分析，提出了在有限语料条件下的汉蒙统计机器翻译调序方法。该方法依据语言学知识获取对译文语序影响显著的短语类型，研究这些短语类型的调序方案，并融入已有的调序模型实现调序的优化。实验表明该方法在有限语料条件下的效果提升显著。

关键词：统计机器翻译；调序；动词短语；有限语料

中图分类号：TP391　　**文献标识码：**A

Reordering for Chinese-Mongolian SMT Based on Small Parallel Corpus

CHEN Lei, LI Miao, ZHANG Jian, ZENG Weihui
(Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China)

Abstract: The reordering models are significant in reducing the difference of word orders between the language pairs in statistical machine translation. Most reordering approaches have high requirements of the scale of the parallel corpus in statistical machine translation. Chinese minority language resources are very scarce and difficult to achieve substantial growth in a short time. Therefore the current reordering approaches cannot play good effect in the translations between Chinese and minority languages. After analyzing the related studies, the paper proposes a source-side reordering method based on a small parallel corpus. In virtue of the linguistic knowledge, we analyzed both corpus and translations to obtain the verb phrases which affected the word orders of translations evidently. And then we studied the reordering rules of these verb phrases, including manually written rules and automatically extracted rules. Experiments showed that our method can improve the performance of the state-of-the-art phrase translation models.

Keywords: Statistical machine translation；reordering；verb phrase；small parallel corpus

1 引言

在统计机器翻译系统中，互译语言之间的语序差异往往较为显著。为了提升最重的译文质量，调序模型在消除互译语言之间的语序差异方面起到至关重要的作用。

通常来说调序模型分为两大类：一类是将调序知识作为特征，融入对数线性模型[1][2]。

* **收稿日期：**2013-06-01　**定稿日期：**2013-07-15
基金项目：中国科学院信息化专项（XXH12504-1-10）；国家自然科学基金面上项目（61070099）
作者简介：陈雷（1981－），男，博士，助理研究员，研究方向为计算机软件与自然语言处理；李淼（1955-），女，研究员，博士生导师，研究方向为模式识别与人工智能；张健，（1954-），男，研究员，博士生导师，研究方向为计算数学；曾伟辉（1982-），女，硕士，助理研究员，研究方向为数据挖掘。

该模型在寻找所需要的特征时往往存在一些困难。同时，将特征融入训练与解码过程会导致调序模型更加复杂，也更加耗时。另一类调序模型是在前处理过程中将源语言的语序尽可能地调整为与目标语言一致。Visweswariah 等提出了一个基于句法的调序方法[3]，该方法从源语言的解析树上自动抽取重排序规则，并自动生成词对齐。Khalilov 和 Sima'an 提出了一个类似的依据源端解析树的特征来决定重排序的源端重排序系统[4]。国内在汉蒙统计机器翻译调序方法的研究上，王斯日古楞[5]、Liang[6]、Chen[7]等均提出了一些基于规则的方法。这一类调序模型的效果取决于重排序规则及其应用方式，同时还需要依赖高精度的句法分析器。上述两类调序模型不是相互排斥的，一些调序模型既可以作为源端重排序的前处理过程，又可以作为特征函数融入到解码器中[8]。

由以上国内外相关研究现状可知，现有的调序方法面向大规模平行语料行之有效。然而，无论是基于短语还是基于句法，都对平行语料的规模具有较高的要求，且存在一定的局限性：首先，基于语法树的重排序依赖于句法分析，或利用短语结构树分析出句子由哪些短语类型（如：动词短语、名词短语等）组成，或利用依存结构树分析出句子的语法成分（如：主语、宾语等），根据这些句法分析所得信息，采用基于规则的方法实现相应树上的操作，例如交换左右子树等，从而完成对源语言语序的调整。然而一方面目前的句法分析准确度不高；另一方面当重排序规则较为复杂时，容易产生规则的嵌套而影响调序效果。其次，基于词性标注的重排序方法能够在保证较细粒度的前提下尽可能多地利用语言的语法信息进行调序。然而目前自动化词性标注的研究工作尚有不足，获取精准的词性标注仍然需要大量且繁琐的人工校对工作，对语言学专家的依赖性很强。

与汉、英、日、法、德等语言百万句级规模的语料相比，我国少数民族语言的语料资源差距巨大，尤其是汉民平行语料规模还远远不能满足需求，且短时间内难以实现大规模增长。从目前汉蒙统计机器翻译研究现状来看，公开且可用于机器翻译研究与测评的汉蒙双语平行语料仍没有超过 10 万句对。同时，语言学专家数量不能满足大规模语料的分析与处理，蒙文语言学知识相对不足且句法分析准确率较低，导致现有的调序方法在汉蒙统计机器翻译系统中难以取得理想的译文质量。

针对上述问题，本文提出了一种有限语料条件下汉蒙统计机器翻译的调序方法。如上所述，在汉蒙统计机器翻译系统中，第一类调序模型的特征难以获取，因此本文采用第二类调序模型，即源端重排序。首先，借助于语言学知识，在语料与译文两个层面上进行分析，获取对译文语序影响显著的短语类型，研究这些短语类型的调序规则，包括人工撰写规则与自动抽取规则，然后基于规则进行源端重排序。与传统的基于规则的方法不同，本文仅关注对译文语序影响显著的短语类型，借助于已有的语言学知识即可获得，在对大规模平行语料以及语言学知识的需求方面寻找一个平衡点，力求满足现有的实际情况。实验表明本文的方法行之有效，在有限语料条件下能够取得译文质量的显著改善。

2 汉蒙统计机器翻译的相关工作

汉蒙统计机器翻译一直是我国自然语言处理研究领域的重要课题，经历了基于规则、基于实例与基于统计的多个发展过程。2007 年，侯宏旭等给出了用于汉蒙 EBMT 机器翻译的实例搜索以及短语片段划分、匹配、组合的方法[9]，该方法基于词语对齐，利用词语对齐进行词语的匹配，并根据匹配词数和长度计算相似度，选取最好的实例；同时考虑到语料规模的限制，双语词典的词汇覆盖面往往不够，采用双语词典进行词语对齐有召回率不高的缺点，并通过人工对齐工具进行校对。由于汉蒙平行语料的稀缺，直到 2009 年，随着汉蒙统计机器翻译评测的出现，其相关研究才逐步发展起来。杨攀等考虑到汉蒙语言形态信息的差异性以及当前由于缺乏大规模汉蒙平行语料所造成的数据稀疏问题，将形态学方法引入到汉蒙统计机器翻译的研究中[10]，在一定程度上解决了译文的词形选择及语序混乱问题。骆凯等提出了类似的方法，将源语言句法信息和目标语言形态信息引入到汉蒙统计机器翻译的模型构造中，以降低译文的词形错误率，并部分解决了译文的长距离调序的问题，从而提高译文的忠实度[11]。朱海等在汉蒙平行语料的基础上，借助汉蒙对齐词典来构造统计模型，并尝试以混淆网络的形式进行词级别的系统融合，在第五届全国机器翻译研讨会的汉蒙日常用语评测项目中取得了良好的成绩[12]。2010 年，Li 等将蒙古语词素（词干、词缀）作为中间语言，构造了多级的链式机器翻译系统[13]：首先利用统计的方法将蒙古语切分为词素，再构造汉语与蒙古语词素的统计机器翻译系统将汉语翻译为蒙古语词素，然后构造蒙古语词素与蒙古语的统计机器翻译系统将蒙古语词素翻译为蒙古语。该方法通过构造链式机器翻译系统，在第一个统计机器翻译系统中将蒙古语词素作为普通单词对待，其本质上是削减了蒙古语的形态信息，在第二个统计机器翻译系统中利用了蒙古语词素中所包含的语言信息以及蒙古语词素与其表面词形的内在联系，从而提高了最终的译文质量。2011 年，王斯日古楞等针对汉蒙统计机器翻译提出了一种基于人工撰写规则的重排序方案[5]，依据汉蒙语言学知识，给出 12 条调序规则，其中动词短语 7 条，介词短语 3 条，主谓短语 3 条，这些规则较好地反映了汉蒙之间的语序差异，在统计机器翻译系统中取得了良好的效果。Liang 等提出了类似的基于人工撰写规则的源端重排序方案[6]，依据这些规则来匹配源语言短语结构树的子树，并进行左右子树的交换操作，同时利用词性标注信息同步实现短语级别和词级别的调序。在此基础上，Chen 等进一步提出在源端重排序模型中借助源端依存关系信息来平衡汉蒙之间的形态信息差异[7]。上述基于规则的调序模型首先需要对源语言进行句法分析，然而这一过程被认为是这种方法主要的缺点[14]。尤其在蒙汉统计机器翻译中，蒙古语句法分析器的精度偏低，在很大程度上影响了基于规则的调序模型的最终效果。2012 年，斯•劳格劳等基于蒙古语依存树库 MDTB，实现了一种基于词汇依存概率的蒙古语依存句法分析模型[15]，该模型对核心词进行分析的准确率达到了 93.05%。随着句法分析器准确率的提高，基于规则的调序模型的效果也将会随之改善。

如上所述，目前汉蒙统计机器翻译的研究主要是针对语序差异和形态差异的。然而在统

计方法中解决这两个问题对语料规模的依赖性较大，在短时间内难以实现质的突破，因此许多研究都引入了语言学知识，如调序规则、词性标注等信息，取得了一定的成果。本文与上述工作的不同一方面是通过对语料与译文两个层面的分析，仅关注对译文语序影响较大的短语类型并研究其调序方案；另一方面是立足实际情况，充分利用现有的有限语料以及语言学知识来获取更佳的译文质量。

3 有限语料条件下的调序

总的来说，汉语的句子是主-谓-宾结构，蒙古语的句子是主-宾-谓结构，在短语级别与词级别方面，汉蒙语序的差异则更加复杂，其具体表现为词对齐关系存在很多交叉。如图 1 所示，例子中的汉语句子与蒙古语句子（拉丁形式）的词对齐连线存在很多交叉现象。语序的差异问题很大程度上影响了译文的质量。源端重排序的任务就是消除图 1 中这种词对齐连线的交叉现象。

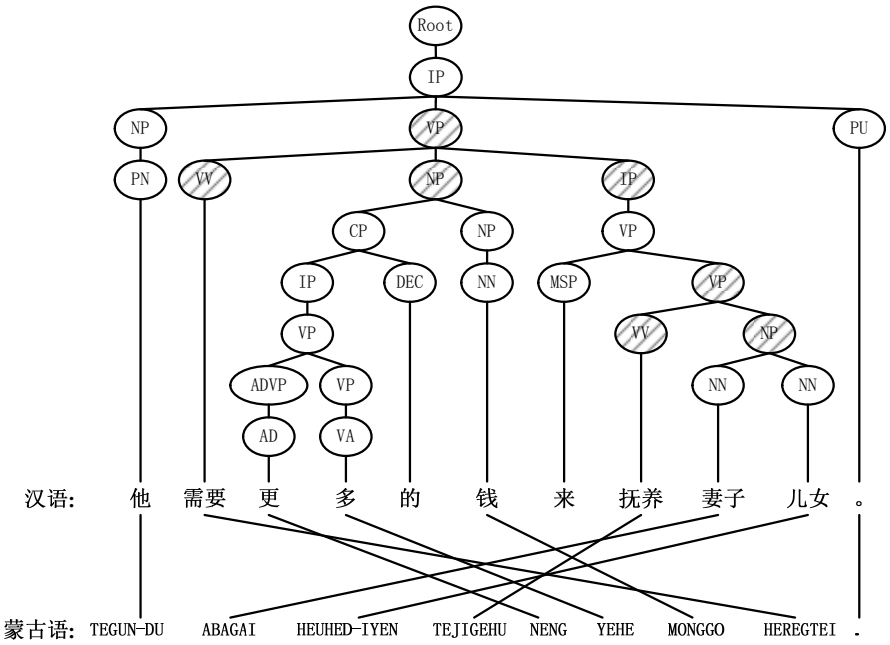

图 1 汉语句子的短语结构树以及与蒙古语句子的词对齐关系

基于短语的统计机器翻译只解决了短距离的局部调序，而处理长距离的调序，正是汉蒙机器翻译语序调整必需的。在基于短语的统计机器翻译系统中，使用隐含长距离调序信息的规则对汉语句子语序进行调整，其中规则的获取是至关重要的。规则可以由人工进行归纳总结，也可以从平行语料库中自动获取。本文分别探讨了基于自动抽取短语结构重排序规则的源端重排序和基于人工编写短语结构重排序规则的源端重排序。

3.1 人工撰写的规则

由于动词或谓语是句子的核心成分，且汉蒙语序的差异主要体现在动词相关的短语上[5][6][7]，因此动词短语的调序在汉蒙统计机器翻译系统中至关重要。本文在汉蒙平行语料与机器翻译系统产生的译文两个层面上分析对译文语序影响显著的动词短语类型。

借助于已有的语言学知识，首先初始化一个包括所有可能显著影响译文语序的动词短语类型的集合 *S*；依据该集合对有限语料进行划分，去除没有对应划分的动词短语类型，得到精简后的集合 *S′*；类似地再用精简后的集合 *S′*对机器翻译系统输出的译文进行划分，再次对集合 *S′*进行精简，得到最终包含所需的动词短语类型的集合 *S″*。

本文基于句法分析所得到的短语结构树[6]来定义针对集合 *S″*中的动词短语的调序规则，其形式为：*VP: x —>x′, w*，其中 *VP* 表示动词短语在短语结构树上对应的节点，*x* 表示 *VP* 的孩子节点序列（按照从左到右的顺序，遵守汉蒙之间的语言规则），*x′*表示对 *x* 进行重排序之后的节点序列，*w* 表示该规则的权值，在人工撰写规则时可由人为指定取值，在自动抽取规则时可从平行语料中训练获得，用于在多条规则产生冲突时进行规则的选取。图 1 给出了一个汉语句子短语结构树，可见每个短语可对应短语结构树上的一棵子树。

表 1 给出了人工撰写动词短语调序规则（不包含权值），其中 *VV* 表示动词，*P* 表示介词，*PP* 表示介词短语，*NP* 表示名词短语，*QP* 表示量词短语。

表1 人工撰写的动词短语调序规则

序号	调序规则
（1）	VP: VP_1 VP_2 —> VP_2 VP_1
（2）	VP: VV PP —> PP VV
（3）	VP: VV NP —> NP VV
（4）	VP: VV QP —> QP VV

在调序时，使用上述规则匹配源语言句子短语结构树的子树进行调序。因此首先需要构造源语言句子的短语结构树，可通过句法分析器获取；其次查找短语结构树中满足如下条件的节点 *n*：标注为 *VP* 且其孩子节点匹配某条规则 *r* 中 *x* 序列；然后根据规则 *r* 的 *x′*序列重新排序节点 *n* 的孩子节点，从而实现源语言句子的重排序。从上述过程可以看出，重排序规则的应用其本质上是短语结构树上的树变换过程。

3.2 自动抽取的规则

除上述人工撰写规则外，本文还研究了如何基于有限语料自动抽取动词短语的重排序规则。给定一个源语言句子 *s*，其短语结构树记为 t_s，t_s 中非叶子节点 *n* 的孩子节点集合记为 C_n，对应于目标端，节点 *n* 的平均位置计算如下：

$$avepos(n)=\frac{1}{C_n}\sum_{\omega\in C_n} pos(\omega)$$

其中 *pos(ω)*表示单词 *ω* 对应于目标端的位置，当单词 *ω* 与目标端的任何单词没有对齐关系时，将无须计算 *pos(ω)*。类似地可以计算短语结构树 t_s 中的每个节点的平均位置，用以调整节点顺序，得到重排序之后的短语结构树，记为 t_r。基于语料中所有句子按照上述过程产生的树对$<t_s, t_r>$，可以抽取所需的重排序规则，并依据最大概率 $P(t_r|t_s)$来选取规则：

$$P(t_r \mid t_s)=\prod_{n\in I(t_s)} P(r(c_n)\mid c_n)$$

其中 $I(t_s)$表示 t_s 的非叶子节点集合，c_n 表示节点 n 的孩子节点序列，$r(c_n)$表示对 c_n 重排序之后的节点序列。$P(r(c_n)|\ c_n)$计算如下：

$$P(r(c_n)\,|\,c_n)=\frac{f(r(c_n))}{f(c_n)}$$

其中 $f(c_n)$是 c_n 在短语结构树 t_s 中出现的频率，$f(r(c_n))$是 $r(c_n)$在短语结构树 t_r 中出现的频率。

给定短语结构树 t_s 上的一个具有 k 个孩子节点的节点 n，其 k 个孩子节点的组合方式共有 $k!$种，本文选择概率最大的组合方式，即选择概率最大的规则，从而获得重排序规则。

利用上述方法，除能够抽出表 1 给出的重排序规则之外，还能够得到大量动词短语相关的重排序规则，如表 2 给出的规则（5），其中 *IP* 表示以屈折成分开头的简单从句。

表 2 人工撰写的动词短语调序规则

序号	调序规则
（5）	VP: VV NP IP —> IP NP VV

应用自动抽取规则的方法与人工撰写规则相同，所不同的是自动抽取规则数量远远超过人工撰写的规则。从本文使用的有限语料中，即可抽出超过 1 千条重排序规则。通过去除错误规则与合并类似规则之后，仍然存在 440 条规则。在应用重排序规则时，容易导致规则选取上的冲突，或造成过度重排序问题。因此定义规则时引入了权值 w 用以缓解此类问题。此外，加入一些语法限制条件也能起到类似的作用[5]。

图 2 给出了在图 1 所示的汉语短语结构树上进行源端重排序之后的结果，标注阴影的节点分别匹配规则（3）和规则（5）。可以看出，针对句中的两个动词短语进行调序，则完全消除了词对齐的交叉现象，意味着重排序之后的汉语句子的语序与蒙古语一致。该例子表明针对动词短语类型的调序在汉蒙统计机器翻译的源端重排序中是行之有效的。

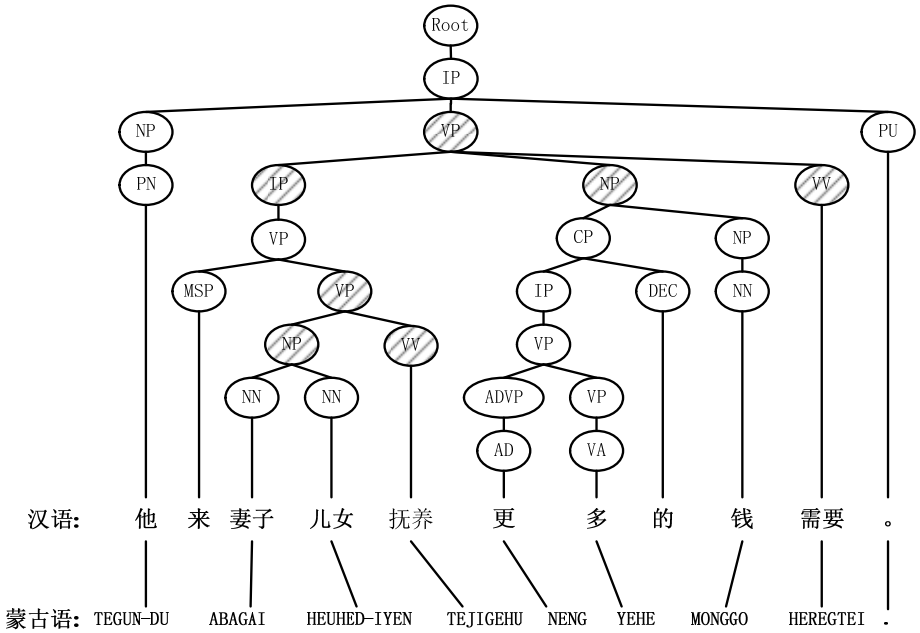

图 2 利用规则（3）和规则（5）对汉语句子进行重排序之后的结果

4 实验

4.1 实验环境与设置

实验软硬件平台为：操作系统 Ubuntu 11.04，处理器 Inter(R) Core(TM)2 Quad CPU Q6700 @ 2.66GHZ，内存 4G。

实验语料为第五届全国机器翻译研讨会（CWMT2009）提供的汉蒙双语平行评测语料，训练集为 67288 句对，开发集为 400 句对，每句汉语对应 4 句由蒙古语言学专家翻译的蒙古语译文，测试集与开发集相同。

在数据处理方面，使用计算所的分词工具 ICTCLAS 2.0 进行汉语分词；使用斯坦福大学的句法分析器 Stanford parser 进行汉语的句法分析，并进行简单的结构映射变换得到短语结构树；在训练时，将训练集、开发集与测试集的传统蒙文转化为拉丁形式；采用开源解码器 Moses[16]进行翻译模型的构建与解码，使用对数线性模型对各种参数特征进行融合，使用的主要特征包括：正反向短语翻译概率、正反向词汇翻译概率、SRILM[17]训练的 3 元语言模型、词长度惩罚、双向 msd 调序模型；使用 GIZA++并采用启发式方法进行词对齐；使用最小错误率训练 MERT[18]来调参。

以标准的基于短语的统计机器翻译系统为参考，本文设置了三组实验：（1）仅使用标准的基于短语的统计机器翻译系统，作为基线系统；（2）使用手动撰写规则进行源端重排序，包括逐个规则的使用与所有规则的同时使用；（3）使用 440 条自动抽取规则进行源端重排序。

4.2 实验结果与分析

上述三组实验的结果如表 3 所示，使用 BLEU 与 NIST 评分来评价实验结果。

表 3 实验结果

实验系统		BLEU	NIST
基线系统		0.2423	5.6578
人工撰写规则	+规则（1）	0.2586	5.7604
	+规则（2）	0.2557	5.7559
	+规则（3）	0.2476	5.7074
	+规则（4）	0.2515	5.7355
	+所有规则	0.2530	5.7516
自动抽取规则		0.2537	5.6926

从表 3 中的实验结果可以看出，无论是人工撰写规则还是自动抽取规则的应用，取得的结果评分均比基线系统显著提高。令人感兴趣的是，取得最佳成绩即提高 1.63 个 BLEU 值的结果是应用人工撰写的规则（1）所获得的，而不是应用所有人工撰写规则，也不是应用数量更多的自动抽取规则。这一结果标明调序规则并不是越多越好。如上所述，数量众多的规则容易导致规则选取上的冲突以及过度重排序问题。

5 结束语

本文提出在有限语料条件下，分析并获取对译文语序影响显著的短语类型，利用这些短语类型的调序规则来调整源语言汉语的语序，实验证明该方法在汉蒙统计机器翻译系统中取得了良好的效果。该方法为现阶段语料资源稀少的其他语言的机器翻译系统调序技术的研究提供了参考。

下一步将研究减少重排序规则选择上的冲突与降低多个规则同时使用造成的过度排序等问题。此外，由于目前本文的方法仍然依赖于句法分析器的准确性，因此需要研究不依赖于句法分析器的调序方案，例如使用序列标注模型（如：条件随机场模型等）来进行特殊短语的识别与匹配问题。

致谢

感谢对本文工作提供帮助的老师和同学。感谢对本文撰写提出中肯建议的评审老师。

参考文献

[1] 薛永增, 李生, 赵铁军, 杨沐昀. 短语统计机器翻译的句法调序模型[J]. 通信学报, 2008, 29(1): 7-14.

[2] 侯宏旭, 刘群, 李锦涛. 一种基于短语的汉蒙统计机器翻译与调序模型[J]. 高技术通讯, 2009, 19(5): 475-479.

[3] K. Visweswariah, J. Navratil, J. Sorensen, V. Chenthamarakshan, and N. Kambhatla. Syntax based reordering with automatically derived rules for improved statistical machine translation[C]. In COLING, 2010, 1119-1127.

[4] [4] M. Khalilov and K. Sima'an. Context-sensitive syntactic source-reordering by statistical transduction[C]. In IJCNLP, 2011, 38-46.

[5] 王斯日古楞, 斯琴图, 那顺乌日图. 汉蒙统计机器翻译中的调序方法研究[J]. 中文信息学报, 2011, 25(4): 88-92.

[6] F. Liang, L. Chen, M. Li, and Nasun-urtu. A rule-based source-side reordering on phrase structure subtrees[C]. In IALP, 2011, 173-176.

[7] L. Chen, M. Li, M. He, and H. Liu. Dependency parsing on source language with reordering information in SMT[C]. In IALP, 2012, 133-136.

[8] N. Yang, M. Li, D. Zhang, and N. Yu. A ranking based approach to word reordering for statistical machine translation[C]. In ACL, 2012, 912-920.

[9] 侯宏旭, 刘群, 那顺乌日图. 基于实例的汉蒙机器翻译[J]. 中文信息学报, 2007, 21(4): 65-72.

[10] 杨攀, 张建, 李淼, 乌达巴拉, 雪艳. 汉蒙统计机器翻译中的形态学方法研究[J]. 中文信息学报, 2009, 23(1): 50-57.

[11] 骆凯, 李淼, 乌达巴拉, 杨攀, 朱海. 汉蒙翻译模型中的依存语法与形态信息应用研究[J]. 中文信息学报, 2009, 23(6): 98-104.

[12] 朱海, 应玉龙, 李文, 李淼, 乌达巴拉. 第五届全国机器翻译研讨会中科院智能所评测技术报告[C]. 第五届全国机器翻译研讨会论文集, 2009.

[13] W. Li, L. Chen, Wudabala and M. Li. A Chained Machine Translation Using Morphemes as Pivot Language[C]. In COLING 2010 workshop: ALR, 2010, 169-177.

[14] K. Visweswariah, R. Rajkumar, A. Gandhe, A. Ramanathan, and J. Navratil. A word reordering model for improved machine translation[C]. In EMNLP, 2011, 486-496.

[15] 斯•劳格劳, 华沙宝, 萨如拉. 基于统计方法的蒙古语依存句法分析模型[J]. 中文信息学报, 2012, 26(3):

27-32.

[16] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation[C]. In ACL, 2007, 177-180.

[17] A. Stolcke. SRILM - an extensible language modeling toolkit[C]. In Proc. Intl. Conf. on Spoken Language Processing, 2002, 901-904.

[18] F. J. Och. Minimum error rate training in statistical machine translation[C]. In ACL, 2003, 160-167.