一种基于词素媒介的汉蒙统计机器翻译方法

杨振新[1,2]，李淼[2]，陈雷[2]，卫林钰[1,2]，陈晟[1]，孙凯[1]

（1. 中国科学技术大学 自动化系，合肥，230026；
2. 中国科学院 合肥智能机械研究所，合肥，230031）

摘要：汉蒙形态差异性以及平行语料库规模小制约了汉蒙统计机器翻译性能的提升。本文将蒙古语形态信息引入汉蒙统计机器翻译中，通过将蒙古语切分成词素的形式，构造汉语词和蒙古语词素以及蒙古语词素和蒙古语的映射关系，弥补汉蒙形态结构上的非对称，并将词素作为中间语言，通过训练汉语—蒙古语词素以及蒙古语词素—蒙古语统计机器翻译系统，构建出新的短语翻译表和调序模型，并采用多路径解码及多特征的方式融入汉蒙统计机器翻译。实验结果表明，将基于词素媒介构建出的短语翻译表和调序模型引入现有统计机器翻译方法使得译文在 BLEU 值上比基线系统有了明显提高，一定程度上消解了数据稀疏和形态差异对汉蒙统计机器翻译的影响。该方法是一种通用的方法，通过词素和短语两个层面信息的结合，实现了两种语言在形态结构上的对称，不仅适用于汉蒙统计机器翻译，还适用于形态非对称且低资源的语言对。

关键词：中间语言；词素；统计机器翻译；短语翻译表；调序模型

A Morpheme-based Approach for Chinese-Mongolian SMT

YANG Zhenxin[1,2], LI Miao[2], CHEN Lei[2], WEI Linyu[1,2], CHEN Sheng[1], SUN Kai[1]

(1. Department of Automation, University of Science and Technology of China, Hefei 230026, China;
2. Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China)

Abstract: The morphological difference between Chinese and Mongolian and the scarcity of parallel corpus are the main problems in Chinese-Mongolian statistical machine translation (SMT). In this paper, we propose a method that adopts morpheme, which is the basic syntactic or semantic units, as a pivot language to induce new translation knowledge. First, we segment Mongolian word into morphemes, achieving a balance in the morphology of the language pair. Then we treat Mongolian morpheme as pivot language and construct two new SMT systems, which are Chinese-Morpheme SMT system and Morpheme-Mongolian SMT system. A new translation knowledge including phrase translation table and reordering model is induced via these two SMT systems. Finally, we use multiple decoding paths and multiple features method to incorporate the new translation knowledge into our baseline system. Experimental results demonstrate our method can improve the translation quality significantly. The method in this paper is a universal method, besides the Chinese-Mongolian SMT, the method can also be adopted to other morphological low-resource language pairs.

Keywords: pivot language; morpheme; statistical machine translation; phrase translation table; reordering model

1 引言

我国作为一个历史悠久的多民族国家，民族语言之间的相互翻译对促进民族间的文化交流、经济发展具有重要意义。汉语和少数民族语言之间的差异性以及平行语料库规模小使得汉民统计机器翻译面临挑战。

对于汉蒙统计机器翻译来说，汉语和蒙古语在形态方面差异极大。汉语属于没有形态变化的孤立语，蒙古语是形态丰富的黏着语，蒙古语的词由词干和词缀组成。根据所表达的不同意思，蒙古语的词干后面可以层层缀接不同的词缀，因此，在汉蒙统计机器翻译中，需要大规模语料才能覆盖复杂多变的蒙古语。

目前，汉蒙双语语料库需要依靠语言学专家人工构造，费时费力，在短时间里内无法得以大量扩充。汉蒙双语语料资源稀缺，使得以统计为基础的机器翻译面临严重的数据稀疏问题，加之汉蒙在形态方面差异较大，进一步制约了汉蒙统计机器翻译性能提升[1]。

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（61572462，61502445）
作者简介：杨振新（1990—），男，博士研究生，研究方向为统计机器翻译；李淼（1955—），女，研究员，研究方向为自然语言处理

两种语言在形态方面的差异使得统计机器翻译面临严重挑战[2]。相关研究证实融入形态信息对于提高统计机器翻译译文质量有很大帮助。Al-Haj 和 Lavie[3]在阿拉伯语到英语的翻译中对阿拉伯语进行形态切分，有效提高了阿拉伯语—英语机器翻译质量；Kholy 和 habash[4]比较了三种不同的方法将形态信息融入英语—阿拉伯语统计机器翻译中；Luong 等[5]在翻译模型训练和解码过程中采用混合的词素级、词级策略，提高了英语—芬兰语的翻译质量；Goldwater 和 McClosky[6]在捷克语—英语的翻译中对捷克语进行形态分析；Singh 和 Habash[7]在希伯来语—英语翻译中使用形态信息改善未登录词的翻译；Salameh 等[8]针对形态丰富的目标语言翻译，将形态信息融入词格解码过程，显著提高译文质量。

在汉蒙机器翻译方面，杨攀等[1]将蒙古语形态信息引入统计机器翻译因子化模型中，在一定程度上消除了汉蒙形态差异以及译文选词混乱等问题；骆凯等[9]将汉语依存句法信息及蒙古语形态信息融入因子化模型，并采用 LOP 对模型参数进行调整；但因子化模型翻译解码时间较长，且受生成模型影响。Li 等[10]分两步完成汉蒙机器翻译，首先将汉语翻译成蒙古语词素，再将蒙古语词素翻译成蒙古语，但两次机器翻译也会产生相关误差。

与上述工作不同的是，本文将蒙古语形态信息引入统计机器翻译，将蒙古语词素视为中间语言，训练汉语—蒙古语词素和蒙古语词素—蒙古语统计机器翻译系统，通过基于蒙古语词素为媒介的统计机器翻译方法构建出有用的翻译知识，并将其融入汉语—蒙古语统计机器翻译系统中，以此消解汉蒙形态差异及数据稀疏对统计机器翻译的影响。

2 汉蒙形态差异

汉语属于孤立语，词语几乎没有形态变化，同时也没有表示语法意义的附加成分。然而，蒙古语属于黏着语，有着丰富的形态，在构词和构形上与汉语不同。蒙古语的构词、构形都是通过在词干后缀接不同的词尾而实现的，并且根据需求还可以层层缀接，这使得蒙古语形态丰富且复杂。在英语或汉语中必须用词表达的意义，在蒙古语中用构形词缀表示就可以。表 1 是蒙古语形态变化丰富的例子。

表 1 蒙古语形态学示例

词干	词缀	蒙古语	汉语
		SVRVGCI	学生
	D	SVRVGCID	学生们
SVRVGCI	D -VN	SVRVGCID-VN	学生们的
	-YIN	SVRVGCI-YIN	学生的
	-TAI	SVRVGCI-TAI	与学生一起

根据《蒙古语语法信息词典》中的统计，蒙古语有超过 3 万词干，297 个构形词缀，由此派生出来的蒙古语词理论上是呈指数级增长的，这需要大规模语料才能覆盖蒙古语可能的表面词形[1]。相对于汉语而言，蒙古语属于低资源语言，汉蒙双语平行语料稀缺，加之蒙古语形态变化丰富，语法、句法表达能力强，在形态非对称的汉蒙机器翻译系统中，语料稀疏问题更加严重，统计翻译模型的建模能力受到了很大的挑战。蒙古语形态信息中蕴含了丰富的知识，从直观上看，充分利用蒙古语形态信息，对于消解统计机器翻译中面临的数据稀疏问题有很大帮助。

需要说明的是，形态学中的词素包含“词根”、“词干”、“构词词缀”和“构形词缀”等多个概念。但由于蒙古语自动词法分析技术目前只能做到词干、构形词缀的自动识别和标注，因此，本文中的词素包括词干和构形词缀。

3 基于词素的短语翻译表

3.1 短语翻译表

短语翻译表作为基于短语的统计机器翻译中重要的组成部分[11]，由以下四部分组成：正向短语翻译概率、正向词汇化加权、反向短语翻译概率、反向词汇化加权。

反向短语翻译概率的计算方式如下：

$$\phi(\bar{f}\,|\,\bar{e})=\frac{count(\bar{f},\bar{e})}{\sum_{\bar{f}'}count(\bar{f}',\bar{e})} \quad (1)$$

其中，$\bar{f}$ 和 $\bar{e}$ 分别为源语言短语和目标语言短语，$count(\bar{f},\bar{e})$ 是短语对 $\bar{f}$ 和 $\bar{e}$ 同时出现的次数，$\sum_{\bar{f}'}count(\bar{f}',\bar{e})$ 是语料中所有与 $\bar{e}$ 对齐的短语次数。

词汇化加权作为一种有效的平滑方式，可以反映短语对的可靠性，通过将短语分解为词，可获得更多的统计数据，反向词汇化加权计算方式如下：

$$p_w(\bar{f}\,|\,\bar{e})=\prod_{i=1}^{n}\frac{1}{|\{j\,|\,(i,j)\in a\}|}\sum_{\forall(i,j)\in a}w(f_i\,|\,e_j) \quad (2)$$

其中，a 是源语言短语 $\bar{f}$ 和目标语言短语 $\bar{e}$ 的词对齐结果，对于可能存在的空对齐情况，在目标端词首插入字符 NULL，并将源语言短语内部所有空对齐的词对齐到 NULL。$\bar{f}$ 中词的位置 $i=1,...,n$，$\bar{e}$ 中词的位置 $j=0,1,...,m$。$w(f\,|\,e)$ 是源语言词和目标语言词的词汇翻译概率，可以通过最大似然估计计算。

统计机器翻译对数线性框架允许加入任意特征，正向短语翻译概率 $\phi(\bar{e}\,|\,\bar{f})$ 和正向词汇化加权 $p_w(\bar{e}\,|\,\bar{f})$ 同样可以有效提高统计机器翻译译文质量，其计算方式与公式(1)(2)类似。

3.2 中间语言策略的短语翻译表构建

汉蒙形态差异大对统计机器翻译建模造成极大困难，本文将蒙古语进行形态切分，将蒙古语词表示成词素形式，即蒙古语词表示为“词干+ +词缀”。通过将蒙古语词素作为中间语言，训练汉语—蒙古语词素翻译系统和蒙古语词素—蒙古语翻译系统，并构建出新的短语翻译表。

本文用 $\bar{f}$ 表示源语言短语，$\bar{m}$ 表示词素中间语言短语，$\bar{e}$ 表示目标语言短语，则通过两个统计机器翻译系统构建出的短语翻译表如下：

$$\phi_m(\bar{f}\,|\,\bar{e})=\sum_{\bar{m}}\phi(\bar{f}\,|\,\bar{m})\phi(\bar{m}\,|\,\bar{e}) \quad (3)$$

$$\phi_m(\bar{e}\,|\,\bar{f})=\sum_{\bar{m}}\phi(\bar{e}\,|\,\bar{m})\phi(\bar{m}\,|\,\bar{f}) \quad (4)$$

$$p_m(\bar{f}\,|\,\bar{e})=\sum_{\bar{m}}p(\bar{f}\,|\,\bar{m})p(\bar{m}\,|\,\bar{e}) \quad (5)$$

$$p_m(\bar{e}\,|\,\bar{f})=\sum_{\bar{m}}p(\bar{e}\,|\,\bar{m})p(\bar{m}\,|\,\bar{f}) \quad (6)$$

本文将构建出的短语翻译表添加到原始基线系统中，通过多路径解码策略和基线短语翻译表相结合，提升翻译效果。

3.3 算法

根据源语言—词素以及词素—目标语言短语翻译表 PT1 和 PT2，构建出新的基于词素中间语言的短语翻译表 PT_new，算法具体步骤如下：

第 1 步：依次读取 PT1 的每一行，构造两个哈希表，hash_pt1 的键为 PT1 的短语对，值为四个概率；hash_key1 的键为汉语短语，值为源语言所对应的所有词素短语；

第 2 步：依次读取 PT2 的每一行，构造两个哈希表，hash_pt2 的键为 PT2 的短语对，值为四个概率；hash_key2 的键为词素语言，值为源语言所对应的所有蒙古语短语；

第 3 步：循环读取 hash_pt1 的键、值；

第 4 步：根据 hash_pt1 的键，通过 hash_key1 和 hash_key2 找到所有汉语短语对应的蒙古语短语；

第 5 步：依次遍历每一个对应的蒙古语短语，根据汉语短语查找所有的词素短语，通过公式(3)(4)(5)(6)计算相关概率，并将汉语、蒙古语组成的短语对及相关概率存入 PT_new 中；

第 6 步：若处理完 hash_pt1 所有的键，则返回 PT_new，否则转向第 3 步。

4 基于词素的调序模型

4.1 调序模型

词汇化调序模型[12]可以显著提高机器翻译质量，在统计机器翻译中广泛使用。在基于词的词汇化调序模型中，当前短语对与目标语言前方词的位置关系称为前向关系，与后方词的位置关系称为后向关系。根据当前短语对与前后上下文的位置关系，使用三种调序方向，单调（Monotone）、交换（Swap）、非连续（Discontinuous）。因此，

考虑前后向关系，基于词的调序模型一共 6 种特征。

调序方向的识别如图 1 所示。

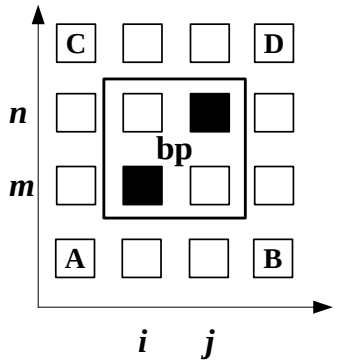

图 1 调序方向识别示意图

在词对齐矩阵中，对于每一个抽取出的短语对，识别其与前后词之前的位置关系。(i, j)是源语言短语位置，(m, n)是目标语言短语位置。对于抽取出的短语对 bp，前向关系的调序方向识别如下：

- 单调调序：如果（i-1, m-1）存在词对齐，且（j+1, m-1）没有对齐点；
- 交换调序：如果（i-1, m-1）没有词对齐，且（j+1, m-1）存在对齐点；
- 非连续调序：单调调序和交换调序以外的情况。

后向关系的调序方向识别如下：

- 单调调序：如果（j+1, n+1）存在词对齐，且（i-1, n+1）没有对齐点；
- 交换调序：如果（j+1, n+1）没有词对齐，且（i-1, n+1）存在对齐点；
- 非连续调序：单调调序和交换调序以外情况。

bp 是由源语言短语 $\overline{f}$ 和目标语言短语 $\overline{e}$ 组成，在给定当前翻译使用的短语对 bp 时，引入一个调序模型 p_r 来预测调序的方向类型 o：

$$p_r(o \mid bp) = \frac{Count(o, bp)}{\sum_{o'} Count(o', bp)} \quad (7)$$

为了避免零概率出现对机器翻译解码造成的干扰，本文将 $Count(o, bp)$ 加 0.5 平滑。考虑前后向调序关系，对于每个给定短语对，本文根据公式(7)计算 6 个不同概率。

4.2 中间语言策略的调序模型构建

同第 3 节所述，本文将蒙古语词素看作中间语言，训练汉语—蒙古语词素机器翻译系统和蒙古语词素—蒙古语机器翻译系统，通过两个系统的调序模型建立新的调序模型。本文将源语言短语 $\overline{f}$ 和词素中间语言短语 $\overline{m}$ 组成的短语对称为 bp_{m1}，将 $\overline{m}$ 和目标语言短语 $\overline{e}$ 组成的短语对称为 bp_{m2}，则构建出的调序模型表示如下：

$$p_{rm}(o \mid bp) = \sum_{\overline{m}} p(o \mid bp_{m1}) p(o \mid bp_{m2}) \quad (8)$$

本文将构建出的调序模型作为特征融入统计机器翻译对数线性框架。中间语言策略的调序模型构建算法与 3.3 节类似。

5 实验

在汉蒙统计机器翻译中，翻译的任务是给定汉语句子 $f = f_1 \cdots f_n$，搜索使得条件概率 $p(e \mid f)$ 最大的蒙古语句子 $e = e_1 \cdots e_m$ 作为译文的输出。在对数线性模型框架下，最优的蒙古语翻译可以定义为：

$$\begin{aligned} \hat{e} &= \arg\max_{e} \{ p(e \mid f) \} \\ &= \arg\max_{e} \{ \sum_{j=1}^{J} \lambda_j h_j(f, e) \} \end{aligned} \quad (9)$$

其中，$h(f, e)$ 统计机器翻译所采用特征，λ 是特征参数。对数线性模型允许添加任意多的特征，每个特征对应一个参数，参数的调节采用最小错误率算法。

基于词素媒介的汉蒙统计机器翻译系统如图 2 所示。

5.1 实验数据

实验采用的训练集是第五届全国机器翻译研讨会提供的汉蒙日常用语训练语料，开发集为 500 句，测试集也为 500 句。开发集和测试集中每句汉语都有 4 句由语言学专家独立翻译的蒙古语译文。蒙古语语料均进行了传统蒙文到拉丁蒙文的转换。实验数据信息如表 2 所示，其中 500*4 指的是 500 句源语言句子，每 1 句源语言对应 4 句目标语言参考译文。

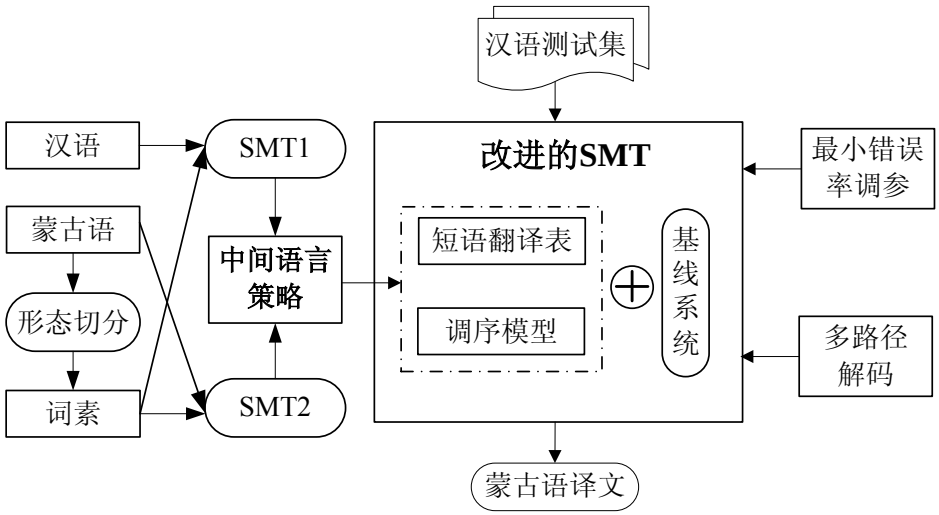

图 2 基于词素的统计机器翻译框图

表 2 实验数据

数据集		汉语	蒙古语
训练集	句子数	67288	67288
	词数	849916	822167
开发集	句子数	500	500*4
	词数	4330	12614
测试集	句子数	500	500*4
	词数	4456	12894

5.2 实验设置

本文首先采用 HMM 方法[19]将蒙古语词切分成词素形式，不考虑词干还原现象。构造出词素中间语言所需的三种不同形式语料：汉语、蒙古语词素、蒙古语。

本文使用开源工具 GIZA++[13]并采用 grow-diag-final-and[11]启发式策略产生双语词对齐。但是在蒙古语词素—蒙古语机器翻译系统中，本文根据蒙古语词干、词缀的规律生成了双语词对齐，并没有使用 GIZA++。使用 SRILM[14]训练 3 元语言模型，并采用改进的 KN 平滑算法[15]。汉语使用 ICTCLAS[16]进行中文分词。采用最小错误率算法[17]对参数进行调整。短语抽取时最大短语长度设为 7。

本文在统计机器翻译系统中采取多路径解码策略，对于基线短语表和基于中间语言策略的短语表，设置两条独立的翻译路径，同时进行解码，择优选择最佳译文。

对于两个调序模型，本文以特征的方式融入对数线性框架下，需要注意的是，由于采用多个短语翻译表，在机器翻译解码过程中，翻译候选可能在调序模型中找不到相应调序概率。针对这种情况，本文采用默认概率的方法，即分别对两个调序模型求出相应的调序概率平均值，如果某个翻译候选无法在调序模型中找到调序概率，则将调序概率平均值作为翻译选项的调序概率。

5.3 实验结果与分析

为了验证本文提出方法的有效性，本文设计了四组系统。

1）系统 A：基线系统，使用的是汉蒙机器翻译领域研究最广泛的基于短语的统计机器翻译系统；

2）系统 B：将词素中间语言构建出的短语翻译表以多路径解码策略融入基线系统中；

3）系统 C：针对系统 B 中构建出的短语翻译表可能无法在基线调序模型中找到相关调序概率情况，设置默认概率；

4）系统 D：在系统 C 的基础上，将词素中间语言的调序模型以特征的方式融入统计机器翻译对数线性模型，对于在调序模型中找不到概率信息的翻译选项，采用默认概率。

本文采用 BLEU[18]对译文进行打分，所有实验均重复三次取平均值，以此消解调参

过程对实验结果的影响。实验结果如表 3 所示。

表 3 实验结果

系统	BLEU(%)
A	20.10
B	20.53
C	20.96
D	21.07

通过表 3 可以看出，本文所提出的基于词素中间语言策略的统计机器翻译方法可以有效地提高机器翻译译文质量。系统 B、C、D 都比基线系统有提升，系统 C 通过采用默认调序概率的方法比系统 B 又有所提高，加入基于词素中间语言的调序模型虽然取得了最好的效果，但是和系统 C 相比并没有太多提高，我们分析的原因可能是双调序模型产生了特征冗余的现象，在开发集上调得的参数不能有效地发挥双调序模型的优势。

为了从直观上理解本文方法的有效性，本文将表现最好的系统 D 和基线系统翻译出的译文进行比较，分析两个系统译文的差异性。翻译结果如表 4 所示。

表 4 翻译结果

例 1：
源语言：你 晚上 干什么 ？
系统 A: TA 0R0I YAGV HIDEG BVI？
系统 D: TA 0R0I YAGV HIHU BVI？
ref0: CI 0R0I YAGV HIHU BVI？
ref1: TA 0R0 YAGV HIHU BVI？
ref2: TA 0R0I YAGV HIHU YVM BVI？
ref3: TA 0R0I DAGAN YAGV HIHU-BER BAYIN_A？
例 2：
源语言：这 是 一 枚 硬币 。
系统 A: ENE B0L NIGE HATAGV .
系统 D: ENE B0L NIGE HATAGV J0G0S .
ref0: ENE B0L NIGEN TEMURLIg J0G0S .
ref1: ENE NI TEMUR J0G0S BAYIN_A .
ref2: TEMURLIG J0G0S BAYIN_A .
ref3: TEMUR J0G0S BAYIN_A .

表 4 是不同系统的机器翻译结果，源语言是汉语，系统 A 和系统 D 输出结果是机器翻译出来的蒙古语，ref0、ref1、ref2、ref3 是语言学家对源语言进行翻译的结果。对于表 4 的两个例子，系统 D 的结果明显优于系统 A。

例 1 是译文时态选择的例子，系统 A 和系统 D 中不同的地方在于 HIDEG 和 HIHU。HI 是动词的词干，表示“干什么”。其后面缀加不同的词缀表示不同的时态，DEG 是表示经常体的形动词词缀，它充当述语时表示动作的经常性或习惯性，多用在现在时。而 HU 是表示将来的词缀，它充当述语时表示绝对的或相对的将来时。因此，系统 D 对于译文时态的选择有帮助。

例 2 是词汇精确翻译的例子，系统 A 翻译结果不完整，系统 A 译文结果没有“币”，而 J0G0S 在蒙古语中表示“币”的意思。因此，系统 D 可以使译文的词汇翻译更加准确。

6 结论

本文针对汉蒙统计机器翻译面临的数据稀疏和形态差异大的问题，将蒙古语进行形态切分，一定程度上消解了汉蒙形态结构不一致，同时将蒙古语词素看作中间语言，训练汉语—蒙古语词素、蒙古语词素—蒙古语翻译系统，并以此构建出两类新的翻译知识：短语翻译表和调序模型，改善了数据稀疏对汉蒙机器翻译系统的影响。另外，本文采用了多路径解码和多特征方法将以词素为媒介构建出来的翻译知识集成到现有机器翻译系统中，集成方式简单有效。实验结果表明，本文方法是有效的。该方法具有通用性，不仅适用于汉蒙统计机器翻译，同时也适用于其它形态非对称且平行语料规模较小的语言对。然而，本文采用的双调序模型会产生特征冗余。因此，下一步工作将考虑如何对特征冗余进行消解，同时将本文方法用在其它形态非对称语言对的统计机器翻译中。

参考文献:

[1] 杨攀, 张建, 李淼等. 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.

[2] Tran K, Bisazza A, Monz C. Word Translation Prediction for Morphologically Rich Languages with Bilingual Neural Networks [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha :ACL, 2014:1676–1688.

[3] Al-Haj H, Lavie A. The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation[J]. Machine translation, 2012, 26(1-2): 3-24.

[4] El Kholy A, Habash N. Translate, predict or generate: Modeling rich morphology in statistical machine translation [C]// Proceedings of EAMT. Trento: European Association for Machine Translation, 2012:27-34.

[5] Luong M T., Nakov P, Kan M Y. A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages [C]// Proceedings of EMNLP. Massachusetts:ACL, 2010: 148–157.

[6] Goldwater S, McClosky D. Improving statistical MT through morphological analysis [C]// Proceedings of HLT-EMNLP. Vancouver :ACL, 2005:676-683.

[7] Singh N, Habash N. Hebrew morphological preprocessing for statistical machine translation [C]// Proceedings of EAMT. Trento: European Association for Machine Translation, 2012: 43-50.

[8] Salameh M, Cherry C, Kondrak G. Lattice desegmentation for statistical machine translation. [C]// Proceedings of ACL, Maryland:ACL, 2014:100-110.

[9] 骆凯, 李淼, 乌达巴拉等. 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.

[10] Li Wen, Chen Lei, Li Miao et al. Chained machine translation using morphemes as pivot language [C]// Proceedings of the 8th Workshop on Asian Language Resources. Beijing:ACL, 2010: 169-177.

[11] Koehn P, Och F J, Marcu D. Statistical phrase-based translation [C]// Proceedings of NAACL-HLT. Stroudsburg :ACL, 2003: 48-54.

[12] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C]// Proceedings of ACL on interactive poster and demonstration sessions. Prague :ACL, 2007: 177-180.

[13] Och F J, Ney H. Improved statistical alignment models [C]// Proceedings of ACL. Stroudsburg:ACL, 2000: 440-447.

[14] Stolcke, A. SRILM - an extensible language modeling toolkit. [C]// Proceedings of International Conference on Spoken Language Processing, 2002:901-904.

[15] Chen, S F, Goodman, J. An empirical study of smoothing techniques for language modeling. [C]// Proceedings of ACL. Stroudsburg :ACL, 1996:310-318.

[16] 刘群, 张华平, 俞鸿魁等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8):1421-1429.

[17] Och F J. Minimum error rate training in statistical machine translation [C]// Proceedings of ACL. Stroudsburg:ACL, 2003: 160-167.

[18] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proceedings of ACL. Philadelphia:ACL, 2002: 311-318.

[19] He Mian-tao, Li Miao, Chen Lei. Mongolian Morphological Segmentation with Hidden Markov Model [C]// Proceedings of IALP. Hanoi: IEEE, 2012: 117-120.