文章编号：1003-0077(2009)01-0050-08

汉蒙统计机器翻译中的形态学方法研究

杨 攀[1,2]，张 建[1]，李 淼[1]，乌达巴拉[1]，雪 艳[3]

（1. 中国科学院 合肥智能机械研究所，安徽 合肥 230031；
2. 中国科学技术大学 信息科学技术学院，安徽 合肥 230027；
3. 内蒙古大学 蒙古学学院，内蒙古 呼和浩特 010021）

摘 要：该文将形态学方法引入到汉蒙统计机器翻译的研究中，尝试解决译文词形选择及语序混乱问题。首先介绍语料库的准备：对原始汉蒙平行语料库进行词法分析及标注，得到两组基础语料库，再由基础语料库生成两组用于形态学实验的派生语料库。其次阐述统计模型的训练，包括语言模型、翻译模型及生成模型。同时讨论了解码的扩展问题。最后重点分析两组形态学方法实验：词素模型实验和 factored 方法实验。结果表明，相对于基线(baseline)实验，引入形态学方法后两组实验的 BLEU 评分均有所提高，译文词形选择及语序混乱问题得到了一定程度的解决。

关键词：计算机应用；中文信息处理；形态学；统计机器翻译；语料库；统计模型；解码
中图分类号：TP391 **文献标识码**：A

Morpholog-Processing in Chinese-Mongolian Statistical Machine Translation

YANG Pan[1,2], ZHANG Jian[1], LI Miao[1], Wudabala[1], XUE Yan[3]

(1. Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China;
2. School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China;
3. School of Mongolian, Inner Mongolia University, Huhhot 010021, Inner Mongolia AR, China)

Abstract: This paper presents an approach to morphology processing in Chinese-Mongolian statistical machine translation, attempting to resolve problems of the word form selection and the word re-ordering in translation generation. On the basis of the original Chinese-Mongolian parallel corpus which is morphologically analyzed and POS tagged, two corpora are derived for the morphological experiments. Then the statistical models, including the language model, the translation model and the generation model, are established. The issue of decoding expansion is also discussed. Finally we analyze the two experiments based on different morphological processing methods: morpheme model experiment and factored method experiment. The results show that the BLEU scores of on the two morphological processing methods are better than the baseline system, revealing our method partially solved the problem of word form selection and word ordering.

Key words: computer application; Chinese information processing; morphology; statistical machine translation; corpus; statistical model; decoding

1 引言

随着信息化时代的到来，各民族之间的交流日趋频繁，语言充当其中的载体发挥着重要的作用。而语言的差异为信息交流带来了障碍，所以民族语言之间的翻译对于促进民族间的交流具有重要意义。汉语和少数民族语言间的机器翻译研究中汉蒙

收稿日期：2007-12-29 **定稿日期**：2008-06-16
资助项目：中国科学院知识创新工程重要方向资助项目(KGCX2-SW-511)
作者简介：杨攀(1983—)，男，硕士生，研究方向为统计自然语言处理与机器翻译；张建(1954—)，男，副研究员，研究方向为人工智能及其应用；李淼(1955—)，女，研究员，博导，研究方向为人工智能与农业知识工程。

翻译是比较成熟的，其中比较典型的有内蒙古大学蒙古学学院和中国科学院计算技术研究所合作的采用基于实例方法实现的汉蒙机器翻译系统[1]。另外中国科学院合肥智能研究所智能农业实验室与内蒙古大学蒙古学学院合作开发的基于统计的汉蒙机器翻译系统 PanGu 已投入实际应用。

汉语和蒙古语的差异极大。汉语属于孤立语，构词主要通过语素的合成来实现，在构形方面被认为“基本上没有形态变化”。从句子的基本语序看，属于 SVO(主谓宾)型语言。而蒙古语属于黏着语，在构词和构形上与汉语有很大的不同。蒙古语以在词干之后附加构词词缀为派生新词的主要手段，以在词干之后附加构形词缀为构形的主要手段，从句子的基本语序看，属于 SOV(主宾谓)型语言。当前，由于缺乏大规模汉蒙平行语料库，汉蒙统计机器翻译研究所面临的数据稀疏问题比较严重，而要通过扩大语料规模来提高系统翻译质量在短时间内是无法实现的。单从译文方面来分析，词形变化方面的错误以及句子语序混乱问题比较明显和突出。本文探索将形态学的方法引入到汉蒙统计机器翻译的研究中，旨在解决译文词形选择及语序混乱问题。

近年来针对形态变化丰富的语言，许多学者在统计机器翻译中用到了形态学知识。Niessen 和 Ney[2]用形态分解的方法来提高德语和英语的词对齐质量。Yang 和 Kirchhoff[3]利用源语言的形态分解信息构造分层回退模型，以解决翻译模型中未出现词的翻译问题，并以德语和芬兰语作为源语言，英语作为目标语言做了相应的实验。Lee[4]在阿拉伯到英语的统计机器翻译中对平行语料库进行了形态信息的分析及标注。Zollmann 等[5]也同样在阿拉伯到英语的统计机器翻译中探索使用形态学的分析方法。Popovic 和 Ney[6]通过在源语言为曲折语的统计机器翻译中使用词干、词缀及词性标注信息来提高翻译质量。Goldwater[7]在捷克语到英语的统计机器翻译中对捷克语语料库进行了形态分析与处理。Minkov 等[8]在解码器解码后通过使用源语言的结构信息及形态信息对译文进行后处理以提高译文翻译质量。Oflazer 和 El-Kahlout[9]则把形态学的研究方法应用到英语到土耳其语的统计机器翻译中，取得了较好的翻译效果。Koehn 等[10-11]在开源的基于统计机器翻译的 Moses 系统上引入了“factored”的方法，其中“factor” 可代表词的表面词形(Surface Forms)、词性(Part-of-speech)、形态信息(Morphology)及词类(Word Classes)等语言学特征。

本文亦是将开源的 Moses 系统作为公共实验平台，除了基于短语的统计机器翻译[12]基线(Baseline)实验外，我们还设计了两组形态学方法实验：一组是利用蒙古语的词素(词干或构形词缀)信息进行实验；另外一组针对 factored 方法，将汉语的词、词性，蒙古语的表面词形、词干、词缀部分、词性、词素标注信息分别作为 factor 引入进行相关实验。从语言学角度需要说明的是，形态学中的词素覆盖“词根”、“词干”、“构词词缀”和“构形词缀”等多个概念。在统计机器翻译中引入形态学知识时，既应考虑语言的构形特征也应考虑其构词特征。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，因此本文仅考虑蒙古语词的构形信息，“词素”仅代表“词干”或者“构形词缀”。

本文其他小节安排如下：第 2 节介绍包含形态信息的语料库的准备，第 3 节重点讨论统计模型的训练，第 4 节介绍解码的扩展，第 5 节是实验及分析，最后是结束语。

2　语料库的准备

统计机器翻译是基于语料库的方法之一，所以语料的准备工作是非常重要的。蒙古语的词形变化是通过将构形词缀黏附于词干后来实现的，且一个词干后可以层层附加多个构形词缀以表达词语之间复杂的语法关系。蒙古语词与词之间有空格，所以蒙古语语料的处理工作中不涉及分词。蒙古语动词有式、态、体以及形动词和副动词变化形式，如 YABV 是意为“走”的动词词干，后接使动态构形词缀 GVL 后形成 YABVGVL，表示“让……走”，再接过去时陈述式构形词缀 BA 形成 YABVGVLBA，表示“让……走了”。体词类有格、复数、反身领属等变化形式，如，SVRVGCI 意为“学生”，加上复数构形词缀 D 后 SVRVGCID 表示“学生们”，再接一个属格构形词缀 VN，形成 SVRVGCID-VN 这样一个词形，表示“学生们的”。根据《蒙古语语法信息词典》中的数据统计，蒙古语包含约 30 000 余个词干，297 个构形词缀，由此派生出的蒙古语词形从理论上说是指数级增长的。因此充分利用蒙古语的形态信息，在一定程度上可以缓解由汉蒙平行语料库规模不大所造成的数据稀疏问题。

2.1　基础语料库

我们的原始语料库是句对齐的汉蒙双语平行语

料库，其中蒙古文以拉丁转写形式录入。汉语部分，利用中国科学院计算技术研究所的汉语词法分析系统ICTCLAS进行分词与词性标注处理，词性标注采用北京大学计算语言所的下位词性标注[13]。蒙古语部分，利用内蒙古大学的蒙古语词法分析系统Darhan进行词的切分和标注，得到蒙古语词素及其标注信息，并通过人工校对尽量确保词法分析结果的准确性[14]。在完成以上预处理后，我们根据模型训练的需要，选择不同的词法加工信息，得到了基础语料库1和基础语料库2。

基础语料库1是汉语“词”到蒙古语“表面词形”（注：汉语“词”到蒙古语“表面词形”指的是汉语句子基本单位为“词”，蒙古语为“表面词形”，下同）的语料库，其中汉语部分只做分词处理，不进行标注，蒙古语部分为表面词形，不进行词素切分及标注，示例如下：

汉语句子（词）：

一个 诚实 的 人 对 行贿 理所当然 感到 愤慨 。

蒙古语句子（表面词形）：

UNENCI HOMON HEGELI TULHIHU YABVDAL-DV JIBEGUCEHU NI MEDEGEJI .

基础语料库2是汉语“词|词性”到蒙古语“词素:词素标注信息”的语料库，其中汉语部分进行分词处理，并有词性标注信息，蒙古语部分进行词干、构形词缀的切分标注，构形词缀以“+”“-”号为起始标志。示例如下：

汉语句子（词|词性）：

一个|mq 诚实|a 的|udel 人|n 对|p 行贿|vi 理所当然|al 感到|v 愤慨|an 。|wj

蒙古语句子（词素：词素标注信息）：

|UNENCI:Ac| |HOMON:Ne1| |HEGELI:Ne2| |TULHI + HU:Ve1 + Ft12| |YABVDAL-DV:Ne1-Fc21| |JIBEGUCE + HU:Ve2 + Ft12| |NI:Sf| |MEDEGEJI:H| |.:Wp1|

2.2 派生语料库

为研究形态学信息对汉蒙统计机器翻译的影响，我们在基线实验基础上增加两组形态学方法实验。其中，除了基线实验可以直接利用基础语料库1外，其他两组实验都要以派生的新语料库作为基础。

派生语料库1为汉语“词”到蒙古语“词素”的句对齐双语平行语料库，其汉语部分来自基础语料库1，蒙古语部分是在基础语料库2的基础上通过在词干及构形词缀之间按“+”“-”号判断添加空格而得到的，示例如下：

汉语句子（词）：

一个 诚实 的 人 对 行贿 理所当然 感到 愤慨 。

蒙古语句子（词素）：

UNENCI HOMON HEGELI TULHI + HU YABVDAL-DV JIBEGUCE + HU NI MEDEGEJI .

派生语料库2为一组factor形式的语料库。在factored方法中，句子的基本单位是一组factor向量，可以包含词、词干、词性等。派生语料库2形式如图1所示，其中括号中的数字是factor的索引编号。

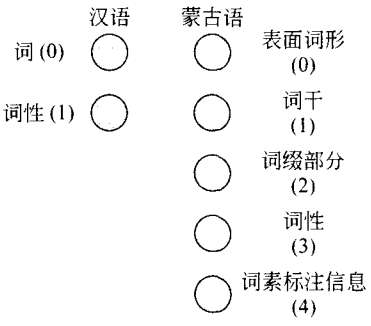

图1 派生语料库2形式

在这组语料库中，汉语句子的基本单位为“词|词性”，蒙古语句子的基本单位为“词|词干|词缀部分|词性|词素标注信息”，其中蒙古语部分是在基础语料库1及基础语料库2的基础上构造的，句子的基本单位以空格为分割标志，factor向量内部以“|”为分割标志，对于词缀部分为空的蒙古语词，我们以“NoSuf”替代，示例如下：

汉语句子（词|词性）：

一个|mq 诚实|a 的|udel 人|n 对|p 行贿|vi 理所当然|al 感到|v 愤慨|an 。|wj

蒙古语句子（词|词干|词缀部分|词性|词素标注信息）：

UNENCI|UNENCI|NoSuf|Ac|Ac

HOMON|HOMON|NoSuf|Ne1|Ne1

HEGELI|HEGELI|NoSuf|Ne2|Ne2

TULHIHU|TULHI|+ HU|Ve1|Ve1 + Ft12

YABVDAL-DV|YABVDAL|-DV|Ne1|Ne1-Fc21

JIBEGUCEHU|JIBEGUCE|+ HU|Ve2|Ve2 + Ft12

NI|NI|NoSuf|Sf|Sf MEDEGEJI|MEDEGEJI|NoSuf|H|H . |. |NoSuf|Wp1|Wp1

3　统计模型的训练

对数线性模型是当前统计机器翻译的基本模型，如公式(1)：

$$p(m \mid c) = \frac{1}{Z}\exp\sum_{i=1}^{n}\lambda_i h_i(m,c) \qquad (1)$$

其中 Z 是一个归一化常量可以忽略，$h_i(m,c)$ 为各种统计模型(如语言模型，翻译模型，扭曲模型等)的特征函数，λ_i 为特征函数系数。在引入形态信息的汉蒙统计机器翻译系统中，除了基本的语言模型(Language Model)，翻译模型(Translation Model)外，主要是引入了生成模型(Generation Model)作为对数线性模型的一个特征函数。

3.1　语言模型

语言模型用于评价译文的忠实度和流利度。本文利用统计机器翻译领域公认的开源语言模型训练工具 SRILM 进行 N-gram 语言模型的训练，其中统一采用改进的 Kneser-Ney 平滑算法(Modified Kneser-Ney discounting)。训练用的蒙古语语料库从基础语料库 1 和 2 中抽取，我们分别抽取了蒙古语“表面词形”语料库(注：句子基本单位为表面词形，下同)，蒙古语“词素”语料库，蒙古语“词干”语料库，蒙古语“词性”语料库，蒙古语“词素标注信息”语料库等。

我们在蒙古语“表面词形”语料库基础上训练了三元语言模型，用于基线测试及对词素模型解码后的重新评分；在“词素”语料库基础上训练了五元的“词素”语言模型，用于词素模型实验的解码；考虑到词性标记数量有限，我们在“词性”语料库基础上训练了七元的“词性”语言模型，用于翻译过程中对蒙古语句子的词性进行评分，以选择更符合蒙古语语序的译文。除此之外，我们还分别训练了五元的“词干”语言模型，七元的“词素标注信息”语言模型等。表 1 是我们对于这几组语言模型的统计(语料库规模 38 000 句，统计至 3-gram)。

表 1　语言模型统计信息

n-gram	表面词形	词　素	词　干	词　性	词素标注信息
1-gram	28 203	10 734	10 357	97	2 052
2-gram	145 659	100 712	106 175	2 607	29 187
3-gram	26 347	43 561	27 473	12 174	31 882

3.2　翻译模型

我们采用开源的 GIZA++ 进行翻译模型的训练。为了利用蒙古语的形态信息，除传统的汉语“词”到蒙古语“表面词形”的短语翻译模型外，我们还训练了几组含形态信息的翻译模型：汉语“词”到蒙古语“词素”的翻译模型；汉语“词”到蒙古语“表面词形|词性”的翻译模型；汉语“词”到蒙古语“词干”的翻译模型；汉语“词性”到蒙古语“词性”的翻译模型等。

我们利用派生语料库 1 进行汉语“词”到蒙古语“词素”的翻译模型训练，其中蒙古语句子的基本单位是词素，即词干或者构形词缀。通过对 GIZA++ 训练后得到的部分对齐结果进行分析，发现除了“错误对应”外，存在如下两个较明显的问题：1)一些汉语词没有找到与之对应的蒙古语单位，或反过来一些蒙古语词素也没能在汉语中找到相应的对应单位；2)部分汉语实词与表达语法意义的蒙古语构形词缀发生对应关系。对于第 1 种问题，目前考虑一是由于语言之间差异造成的正常的“空对应”，二是由于语料库规模限制造成的不正常的“空对应”。第 2 种问题，从汉蒙语言对比研究的角度来看，汉语实词通常与蒙古语实词产生对应关系，而与蒙古语构形词缀有对应关系的往往是汉语虚词，因此出现实词对应词缀现象是不合理的。针对这一点，我们对派生语料库 1 的蒙古语部分做了有选择性的合并：首先我们依据语言对比研究成果建立了“蒙古语构形词缀与汉语虚词的对应表”，表中列出了所有蒙古语构形词缀，其中一部分词缀在汉语中没有对应单位，我们将这部分词缀在语料库中与其词干进行了相应的合并，而其他词缀则不予处理。

其他模型则是在派生语料库 2 的基础上训练。我们分别观察了三个翻译模型中的“读书”翻译选项，如下所示：

汉语“词”到蒙古语“词干”：

读书 ||| NOM VNGSI ||| 0.571 429　0.028 366　0.8　0.150 892　2.718

读书 ||| NOM ||| 0.003 636 36　0.018 121 9　0.2　0.407 407　2.718

汉语“词”到蒙古语“表面词形”：

读书 ||| NOM VNGSIHV ||| 0.5　0.051 033 6　0.333 333　0.025 510 2　2.718

读书 ||| NOM VNGSIJV BAYIHV ||| 0.5　0.037 966 3　0.333 333　0.001 822 16　2.718

读书 ||| N0M VNGSIN_A ||| 1 0.148 256 0.333 333 0.012 755 1 2.718

汉语“词”到蒙古语“表面词形|词性”：

读书||| N0M|Ne1 VNGSIHV|Ve1 ||| 0.5 0.054 757 3 0.4 0.030 178 3 2.718

读书 ||| N0M|Ne1 VNGSIJV|Ve1 BAYIHV|Vz1 ||| 0.5 0.038 419 1 0.2 0.002 235 43 2.718

读书||| N0M|Ne1 VNGSIJV|Ve1 ||| 0.5 0.056 595 6 0.2 0.060 356 6 2.718

读书 ||| N0M|Ne1 VNGSIN_A|Ve1 ||| 1 0.150 346 0.2 0.015 089 1 2.718

可以看出，随着蒙古语句子形态信息的不断增加，翻译模型中的翻译选项也随之丰富起来。

3.3 生成模型

在 factored 方法中，引入了生成模型(Generation Model)，其特征函数如公式(2)：

$$h_g(m,c) = \sum_k \gamma(m_k) \qquad (2)$$

生成模型主要计算目标语言不同 factor 间相互映射的条件概率关系，在目标语言语料库中采用最大似然法训练。如蒙古语“词干”与“词性”生成模型形式如下：

蒙古语“词干”与“词性”：

VNTARG_A Ve1 1.000 000 0 0.000 053 4

ESEHU Cz 1.000 000 0 0.008 888 9

AUIt0MAt Ne2 0.333 333 3 0.000 038 6

AUIt0MAt Ne1 0.666 666 7 0.000 043 7

ANI Ve1 1.000 000 0 0.000 080 1

B0S Ve2 1.000 000 0 0.004 465 6

DEBCIDE Ve2 1.000 000 0 0.000 075 1

其中每一行包含 4 项，第 1 项 factor 为蒙古语词干，第 2 项 factor 为词性，第 3 项数值为 P(词性|词干)，表示在已知词干的情况下是该词性的条件概率，最后一项数值为 P(词干|词性)，表示已知词性的情况下，是该词干的条件概率。除了此生成模型外，我们针对汉蒙统计机器翻译的形态学方法实验又训练了另外几个生成模型，如蒙古语“表面词形”与“词性”的生成模型，蒙古语“表面词形”与“词干|词性”的生成模型，蒙古语“表面词形”与“词干|词缀部分”的生成模型等。

4 解码的扩展

与传统的基于短语的统计机器翻译解码过程相比，增加了 factored 方法的解码过程不再是单一的翻译短语的过程，而是对于输入句子的任一短语，通过几种不同的 factor 的映射步骤(Mapping Steps)最终生成目标短语的翻译选项。映射步骤一般包含翻译及生成，翻译是将源短语的某个 factor 翻译为目标短语的 factor(主要利用 factored 翻译模型)，生成是在目标短语端已有的 factor 的基础上生成新的 factor(主要利用 factored 生成模型)。我们以派生语料库 2 为例，在这组语料库基础上我们设计了一个 factored 方法实验，其短语翻译包含如下 3 个映射步骤：

1. 翻译汉语“词”到蒙古语“词干”(主要利用汉语“词”到蒙古语“词干”的翻译模型)；

2. 翻译汉语“词性”到蒙古语“词性”(主要利用汉语“词性”到蒙古语“词性”的翻译模型)；

3. 在已翻译的蒙古语词干，词性的基础上生成蒙古语表面词形(利用蒙古语“词干|词性”与蒙古语“表面词形”的生成模型)。

短语翻译过程如图 2 所示，其中箭头中的数字表示相应的映射步骤。

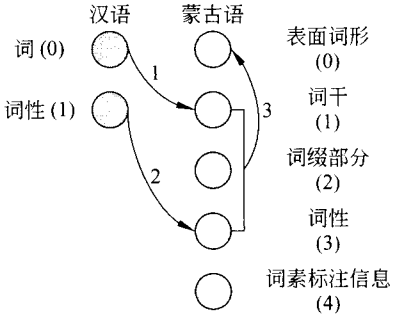

图 2 短语翻译映射步骤

由这个过程可以看出，factored 方法中翻译选项得到了很大程度的扩充，但整个映射过程都是在解码之前完成的，因此，不需要对解码算法做任何修改，只是在解码器加载翻译选项时，适当地提高翻译选项的最大个数即可。

5 实验及其分析

为了进行相关的实验及模型训练，我们按 ACL07 统计机器翻译讨论会(ACL 2007 Second Workshop on Statistical Machine Translation)的介绍搭建了 Moses 测试平台，其核心框架仍是 Kohen 的基于短语的 SMT 框架。平台采用 GIZA++ 进行翻译模型的训练，Srilm 工具进行语言模型的训练，Moses 工具包进行生成模型的训练及解码，Mteval

工具包对实验结果进行 BLEU 评分。

基础语料库来源于内蒙古大学提供的 38 000 句对日常用语汉蒙平行语料库，派生语料库由基础语料库生成。由于目前还没有汉蒙机器翻译公共测试平台，我们选用训练集以外的 200 句日常用语作为测试集，由以蒙古语为母语的专业人员进行翻译，每个汉语句子对应 3 种译文。蒙古语部分采用拉丁转写形式。

基线测试是在基础语料库 1 的基础上使用 Moses 工具箱里默认的训练参数训练汉语“词”到蒙古语“表面词形”的翻译模型，利用 Srilm 训练三元的蒙古语“表面词形”语言模型。对测试集解码时除了将翻译选项个数(ttable-limit)设置为 50 外(注：默认值为 20，并且对于下面两组形态学方法实验，均设置为 50)，其他仍采用 Moses 解码时默认的配置参数。基线测试 BLEU 评分结果如表 2 所示。

表 2　基线测试结果

基线实验	BLEU
表面词形	0.184 7

5.1　词素模型实验

词素模型的实验以派生语料库 1 作为基础，该语料库已做过有选择性的合并处理，如 3.2 节所述，实验过程如下：

我们训练了五元的蒙古语“词素”语言模型及汉语“词”到蒙古语“词素”的翻译模型。测试集解码后输出的蒙古语句子仍是“词素”形式的，通过合并相应的词素，得到“表面词形”级别的输出句子。所谓合并相应的词素是指，将词干与相应构形词缀间的空格去掉，并根据蒙古语词干反向还原规则得到表面词形。在此过程中，可能会出现部分词的“词素结构”混乱问题，我们仅对其中一些极端情况做了修正，如对于句首出现构形词缀的情况，直接将其去掉。

但是这样的输出句子由于没有考虑更多的目标语言知识，并不一定是最佳译文，因此我们又对合并后的译文进行了 re-ranking。所谓 re-ranking 就是指针对某个评估标准对机器翻译程序输出的多个结果进行重新选择，致力于从中选择出使该标准达到最优时的翻译结果[15]。我们采用的 re-ranking 方法如下：对于解码后输出的“词素”形式的 100-best 译文，合并其相应的词素，再利用三元的蒙古语“表面词形”语言模型对其进行重新评分，最后从这些评分后的 100-best 译文句子中选择最佳译文。

表 3 分别给出了词素模型实验 re-ranking 前后的 BLEU 评分。

表 3　词素模型实验结果

词素模型实验	BLEU
re-ranking 前	0.195 8
re-ranking 后	0.203 7

与基线(Baseline)实验相比，词素模型实验 re-ranking 前的 BLEU 评分提高约 0.011，而 re-ranking 后的 BLEU 评分较之 re-ranking 前又提高了约 0.008。通过分析，我们发现词素模型实验的 BLEU 评分在 1-gram 上较之基线实验有很大提高，这表明词素模型实验可以较好地在译文中选择正确的蒙古语表面词形。

下面再以汉语句子“我的爸爸安静地站在那里。”为例，给出基线(Baseline)实验以及词素模型实验 re-ranking 前后的翻译结果，中括号内标识出当前词进行词素合并前的形式：

baseline：MINU ABV NAM GEJU TENDE J0GS0JV BAYIHV .

re-ranking 前：MINU ABV NAM-IYAR [NAM -IYAR] TENDE J0GS0N_A[J0S0 +N_A] .

re-ranking 后：MINU ABV NAMHAN[NAM +HAN] TENDE J0GS0N_A[J0GS0 +N_A] .

从以上翻译结果来看，基线实验中，“站”的翻译为“J0GS0JV BAYIHV”，而动词的这一形式出现在句末，通常被认为是“不合法”的。词素模型实验中，“站”被翻译为“J0GS0N_A”，这种形式能够体现与源文对应的时态，且出现在句末是符合蒙古语语法的。再进一步看 re-ranking 实验，re-ranking 前“安静地”被翻译为“NAM-IYAR”，re-ranking 后被翻译为“NAM+HAN”，前者虽然基本上表达了“安静 地”之意，但后者为形容词“NAM”的比较级形式，相对而言能够使句子更加流畅。

5.2　factored 方法实验

词素模型实验虽然利用了词干，构形词缀等形态信息，却无法考虑包括词性在内的其他词法信息。而 Koehn 的 factored 方法提供了一个更为灵活的框架，在该框架下我们可以利用更多的语言学知识。

因此，针对蒙古语复杂多变的形态特征，我们在派生语料库2的基础上，设计了一组factored方法实验，有关该语料库的介绍，详见2.2节。以下我们选取了其中3个典型实验进行详细说明：

A. 汉语“词”到蒙古语“表面词形|词性”的翻译：自然语言的句子可以被看做是词的序列，如果为每一个词标注词性信息，句子又可被视做是词性序列。而句子的词性是按照一定的规律排列成一个序列的。我们所训练的七元词性语言模型正是要从统计学意义上体现这一规律。本实验中，首先利用汉语“词”到蒙古语“表面词形|词性”的翻译模型进行源短语的翻译，再利用蒙古语三元表面词形语言模型及七元词性语言模型对输出结果进行评分，旨在通过利用词性信息使得译文更加符合蒙古语句子的表达习惯。

B. 汉语“词”到蒙古语“词干”的翻译，汉语“词性”到蒙古语“词性”的翻译，蒙古语“词干|词性”生成蒙古语表面词形：该实验用到了两个翻译模型（汉语“词”到蒙古语“词干”，汉语“词性”到蒙古语“词性”），三个语言模型（三元的蒙古语“表面词形”语言模型，五元的蒙古语“词干”语言模型，七元的蒙古语“词性”语言模型）以及一个生成模型（蒙古语“词干|词性”与“表面词形”的生成模型）。在短语的映射过程中，我们首先将汉语词翻译为蒙古语词干，而不是直接翻译为表面词形，主要目的在于得到一个目标短语的“骨架”；其次将汉语“词性”映射为蒙古语“词性”；最后利用词干及相应词性生成最终的蒙古语表面词形。该实验考虑了蒙古语词干信息，相对于表面词形，在词干基础上得到的相关统计知识更加准确；另外，该实验不单单利用目标语言的词性组合关系，同时还在翻译过程中引入了汉蒙两种语言的词性序列对应关系，从而得到更合理的译文。

C. 多路径解码实验：在Moses中设置两条翻译路径，分别对应实验A与B，同时进行解码，择优选择最佳译文。

图3是3组实验的综合图示意，箭头上的字母分别代表相关实验，实验C覆盖了实验A与B。

表4是这组实验的BLEU评分，相对于基线实验，实验A，B均提高了约0.013，通过分析，我们发现其BLEU值的提高主要由3-gram，4-gram的BLEU值增加推动的，这表明在译文的局部语序调整上，词性模型起到了一定的作用。但是与我们的预期不同的是，实验B较之实验A的BLEU评分几乎没有增加。通过分析两组测试集的翻译结果，我

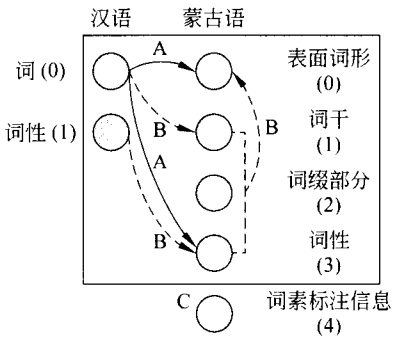

图3 factored实验A、B、C

们发现虽然实验B考虑了汉语句子方面的词性信息，但是由于实验A的翻译选项更加丰富，因此一些输出译文在局部语块的选词方面表现反而更好。实验C取这二者之长，表现相对出色，其BLEU评分较之基线实验提高了约0.02。

表4 factored方法实验结果

factored方法实验	BLEU
实验A	0.197 5
实验B	0.198 1
实验C	0.204 5

我们选择了一个较为典型的简单句“小李读小说。”来依次观察基线实验、实验A、B和C的翻译结果。基线实验与实验A对输入的例句仅做分词处理，实验B与C对输入的例句进行分词及词性标注。翻译结果如下，其中中括号内为Moses的详细输出形式：

baseline: JIJIG LI VNGSIJV UGULELGE .

实验A: JIJIG LI UGULELGE VNGSIJV .

[JIJIG | Ac LI | Nel UGULELGE | Nel VNGSIJV | Vel . | Wpl]

实验B: SIY0V LI UGULELGE VNGSIJV .

[SIY0V | SIY0V | Ntl LI | LI | Ntl UGULELGE | UGULELGE | Nel VNGSIJV | VNGSI | Vel . |. | Wpl]

实验C: SIY0V LI UGULELGE VNGSIJV .

[SIY0V | SIY0V | Ntl LI | LI | Ntl UGULELGE | UGULELGE | Nel VNGSIJV | VNGSI | Vel . |. | Wpl]

基线实验中，系统没有能够将“小李”翻译成人名，而是把“小”译为“JIJIG（体积小的）”，把“李”音译为“LI”；另外蒙古语译文以“主谓宾”形式输出，这显然不符合正常的蒙古语语序。实验A中，“小

李”的翻译仍与基线实验结果一样，但译文语序得到调整，输出结果符合蒙古语基本句式的“主宾谓”表达习惯。我们分析认为，这是七元的词性语言模型在起作用。实验B和C在翻译人名“小李”方面，给了我们一个惊喜，系统将“小”和“李”都当作人名以音义形式输出。与此同时，译文继续保持正确的“主宾谓”语序。仅就这个例子来看，“词性”到“词性”的翻译模型发挥了一定的作用。需要指出的是，以上实验中“读”均被翻译为“VNGSIJV”，虽在核心意义上表达了“读”的意思，但按照蒙古语语法，“动词词干＋构形词缀JV(阳性)/JU(阴性)”这一表面词形不能出现在句末，而且它不表示一般现在时的时态特征。

6　结束语

考虑到蒙古语是形态变化丰富的语言，本文尝试采用词素模型实验解决译文词形正确选择的问题，采用factored方法实验解决译文语序混乱的问题。实验结果表明，在汉蒙统计机器翻译中，引入形态学的研究方法具有合理性和可操作性，但仍存在一些问题有待进一步探索。

在词素模型实验中，我们仅对个别的“词素结构”混乱问题做了简单的处理，对于其他情况，可以考虑采用一些后处理方法，如利用蒙古语“拼写检查器”进行译文词形的修正等。factored方法实验中，虽然在Moses框架下可以利用多种语言学知识，但是随着实验中用到的factor数目及相应映射步骤的增加，计算复杂度也迅速增大，其理论意义远大于实际意义，因此，设计更为适用的解码算法是非常必要的。

参考文献：

[1] 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.

[2] Sonja Niessen, Hermann Ney. Statistical Machine translation with Scarce Resources Using Morpho-syntatic Information[J]. Computational Linguistics, 2004, 30(2): 181-204.

[3] Mei Yang, Katrin Kirchhoff. Phrase-based Backoff Models for Machine Translation of Highly Inflected Languages[C]// Proceedings of EACL. 2006: 41-48.

[4] Young-Suk Lee. Morphological analysis for statistical machine translation[C]//Proceedings of HLT-NAACL 2004-Companion Volume. 2004: 57-60.

[5] Andreas Zollmann, Ashish Venugopal, Stephan Vogel. Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation[C]//Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume. 2006: 201-204.

[6] Maja Popovic, Hermann Ney. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation[C]//Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC). 2004:1585-1588.

[7] Sharon Goldwater, David McClosky. Improving Statistical MT Through Morphological Analysis[C]// Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005: 676-683.

[8] Einat Minkov, Kristina Toutanova, Hisami Suzuki. Generating Complex Morphology for Machine Translation[C]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07). 2007: 128-135.

[9] Kemal Oflazer, Ilknur Durgar El-Kahlout. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation[C]//Proceedings of the Second Workshop on Statistical Machine Translation (ACL'07). 2007: 25-32.

[10] P. Koehn, Hieu Hoang, Alexandra Birch et al. Moses: Open Source Toolkit for Statistical Machine Translation[C]//Proceedings of the ACL 2007 Demo and Poster Sessions (ACL'07). 2007: 177-180.

[11] P. Koehn, Hieu Hoang. Factored Translation Models [C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(ACL'07). 2007: 868-876.

[12] P. Koehn, F. J. Och, D. Marcu. Statistical Phrase-Based Translation[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003. Edmonton, Alberta, Canada.

[13] 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.

[14] 那顺乌日图，雪艳，叶嘉明. 现代蒙古语语料库加工技术的新进展—新一代蒙古语词语自动切分与标注系统(Darhan Tagging System)[C]//第十届全国少数民族语言文字信息处理学术研讨会论文集. 青海：2005.

[15] 付雷，刘群. 单纯形算法在统计机器翻译Re-ranking中的应用[J]. 中文信息学报，2007，21(3)：28-33.