# 内蒙古师范大学 CWMT2011 蒙汉机器翻译系统评测技术报告

王春荣 宝美荣 王斯日古楞
内蒙古师范大学计算机与信息工程学院 呼和浩特 010022
E-mail: {20104019012,jsj05547}@mail.imnu.edu.cn
siriguleng@imnu.edu.cn

摘要：本文介绍了内蒙古师范大学计算机与信息工程学院自然语言信息处理实验室参加 CWMT2011 机器翻译评测中的蒙汉日常用语评测情况。为了提高蒙汉统计机器翻译系统的性能，对蒙古语的格、数和领属附加成分进行切分处理，并提交了一个系统翻译结果。本文对参加评测的基于短语的蒙汉统计机器翻译系统进行了详细说明。

关键字：统计机器翻译、基于短语的蒙汉统计机器翻译、蒙古文附加成分

# Technical Report of Mongolian-Chinese Statistical Machine Translation Based on Phrase

Chunrong Wang,Meirong Bao, Siriguleng Wang
Collage of Computer and Information Engineering, Inner Mongolia Normal University, Hohhot 010022, China

E-mail: {20104019012,jsj05547}@mail.imnu.edu.cn
siriguleng@imnu.edu.cn

**Abstract:** *This paper describes our Mongolian-Chinese machine translation of daily texts system for the machine translation evaluation task of the CWMT2011. In order to improve the performance of the phrase-based Mongolian-Chinese statistical machine translation system, we separate the additional composition of which Mongolian cases, plural forms and genitive cases from the words, and submitted a translation result. In this paper, we gave a detailed description of the phrase-based Mongolian-Chinese statistical machine translation system for the machine translation evaluation task.*

**Keywords:** *Statistical Machine Translation, Phrase-based Mongolian-Chinese Statistical Machine Translation, Mongolian Additional Composition*

## 1 引言

CWMT2011 机器翻译评测开展了汉英新闻领域机器翻译、英汉新闻领域机器翻译、英汉科技领域机器翻译、日汉新闻领域机器翻译、蒙汉日常用语机器翻译、藏汉政府文献机器翻译、维汉新闻领域机器翻译、哈汉新闻领域机器翻译及柯汉新闻领域机器翻译等 9 项评测，我们参加了其中的蒙汉日常用语机器翻译评测，并提交了一个系统翻译结果。该系统对蒙古文的分写附加成分进行切分处理来提高了系统性能。下面对该系统进行详细说明。

## 2 系统

### 2.1 翻译模型

目前，基于统计的方法在机器翻译领域占据着主导地位，并出现了多种不同类型的统计机器翻译方法，例如基于词的、基于短语的、基于句法的以及基于混合策略的翻译方法。其中，基于短语的翻译模型是当前统计机器翻译领域的一个研究热点。

本文采用基于短语的统计翻译模型，并利用该模型建立各种特征函数和相应参数的线性组合。根据柱搜索算法获得最佳译文 $e_{best}$。最优译文的计算公式为：

$$e_{best} = \arg\max_{e_1^*} \left\{ p\left(e_1^i, f_1^j\right) \right\} = \arg\max_{e_1^*} \left\{ \sum_{m=1}^{M} \lambda_m h_m\left(e_1^i, f_1^j\right) \right\} \quad \text{公式 1}$$

该模型采用如下翻译模型作为特征函数。

短语正向翻译概率（Phrase Translation Probability）；

短语反向翻译概率（Inverse Phrase Translation Probability）；

词汇正向翻译概率（Lexical Translation Probability）；

词汇反向翻译概率（Inverse Lexical Translation Probability）；

因子化翻译模型（Factored Translation Model）；

语言模型（Language Model）；

MSD 重新调序模型（MSD Reordering Model）；

短语长度惩罚（Phrase Length Penalty）；

生成模型（Generation Models）；

通过短语翻译模型训练，从双语平行语料库中抽取双语短语对并计算其概率。在对数线性模型中考虑各个特征函数的估计值外，还要考虑对应的参数值。因此，通过最小错误率训练来调整特征函数的参数，以便获得译文和参考答案在某种评价方法下错误率最小的一组参数值。

## 2.2 基本系统

本系统的运行环境：操作系统为 ubuntu 10.10 版的 Linux 平台；双核；两个 CPU 都为 Intel(R) Core(TM)2Duo CPU E8400 @3.00GHz；内存为 2.00G。

该系统由若干模块组成，分别为语言模型、词语对齐模块、解码器及评测工具。下面分别介绍各个模块。

### 2.2.1 语言模型

语言模型是采用统计分析和模拟自然语言的结构规律，用统计概率来衡量某句子符合语法的程度。

语言模型从大量的训练语料中获取语言结构知识。因此采用设计领域广泛、规模足够大的语料库作为训练语料，使得训练结果充分反映目标语言句子的特点，这是建立统计语言模型的重要基础。本文用到的语言模型的训练语料库是规模为 797311 条句子的汉语单语语料库。

本系统使用了语言模型工具 SRILM，并进行了 3-gram 语言模型训练。

语言模型工具 SRILM 是由 SRI 口语技术与研究实验室开发的，建立和使用统计语言模型的开源工具包。该工具包含一组 C++类库、一组执行标准任务的可执行程序和其它各种脚本。

### 2.2.2 词语对齐模块

词语对齐部分，本系统使用了开源工具 GIZA++。

词语对齐工具 GIZA++是著名的训练翻译模型模块 GIZA 的升级版。GIZA++能够方便地对双语语料库进行词语对齐训练，由句子对齐的双语平行语料库得到词语对齐的模型。GIZA 是独立于语言的，能够对任何两种语言进行训练。

词语对齐的具体步骤如下：

（1）用 CIZA++提供的 plain2snt.out 工具生成 GIZA++所需要的四个输入文件。生成的文件分别为源语言词汇文件、目标语言词汇文件、源语言到目标语言的数字对齐文件及目标语言到源语言的数字对齐文件等。

(2)利用 GIZA++进行汉语到蒙语、蒙语到汉语两个方向的训练，获得双向词语对齐结果。

(3)采用 grow-diag-final-and 对双向对齐结果进行优化。GIZA++实现了基本的 IBM 统计翻译模型，但得到的对齐结果只考虑了一对一的情况，忽略了多对多及一对多的情况。对此，按照 och 等提出的 refined alignment 的思路进行优化。

### 2.2.3 解码器

解码的目标是根据建立好的短语翻译概率表和汉语语言模型，翻译测试语料并输出最恰当的译文。

解码器部分，本系统使用了 Moses 开源解码工具。

解码工具 Moses 是由英国爱丁堡大学等 8 家单位的研究人员联合开发的目前比较成熟的基于短语的统计机器翻译系统。它是由美国南加州大学信息科学实验室的菲利普·科恩(Philipp Koehn)开发的统计机器翻译系统 Pharaoh 的升级版本。

### 2.2.4 评测工具

本系统使用了由 CWMT2011 提供的自动评测工具 mteval_sbp，进行评测。该评测工具提供了 NIST、BLEU、BLEU_SBP、GTM、mWER、mPER 及 ICT 方法等的得分结果。

## 3 数据

本系统采用了 CWMT2011 提供的 67288 条句对的蒙汉对齐语料库作为训练集，利用 CWMT2011 提供的 1000 条句子的蒙古文语料库和对应的 4 个汉语参考答案的开发集作为训练最小错误率的语料库，从搜狗语料库中取 730023 条句子，并与训练语料库的 67288 条句子的汉语语料库合并起来作为训练语言模型的语料库，最终对 CWMT2011 提供的 500 条蒙古文句子进行了翻译。

为了完成评测，对语料库进行相应的预处理。预处理分为蒙古文语料库预处理和汉语语料库预处理。

蒙古文语料库的预处理包括利用内蒙古大学蒙古文拉丁转写工具把蒙古文转写成拉丁形式并将蒙古文拉丁转写形式中的单词与标点符号进行分割。

利用我们原有的 66932 条句对的蒙汉双语语料库作为训练集，用独立于训练集的 400 条句子的开发集和测试集上分别做了老蒙古文和蒙古文拉丁转写形式的两组实验，实验结果表明，用蒙古文拉丁转写形式的评测结果比用老蒙古文形式的评测结果高，而且用蒙古文拉丁转写形式有运行速度快，处理方便等优点，所以，我们选择了蒙古文拉丁转写形式。

汉语语料库预处理包括 A3 区全角字符转换和汉语分词。用“丝路 1.0”的汉语 A3 区全角字符转换工具对汉语进行全半角转换。汉语的 A3 区全半角转换是，把汉语的 A3 区全角符号Ａ－Ｚ，ａ－ｚ，０－９，共 62 个字符，转换为相应的半角符号(英语 ASCII 码)A-Z,a-z,0-9。用中国科学院计算所开发的汉语词法分析系统 ICTCLAS2011，对汉语进行分词。

语料库预处理的例子：

汉语句子

（1） 预处理之前的汉语句子

７１３房间在这里，请进。

（2） 全角到半角转换的汉语句子

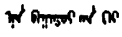
713 号房在这儿，请进。

（3） 分词处理以后的汉语句子

713 号 房 在 这儿 ， 请 进 。

蒙古文句子

（1） 预处理之前的蒙古文句子



（2） 拉丁转写后的蒙古文句子

ENE BAGATVR-VN GER MON UU?

（3） 分割标点符号以后的蒙古文句子

ENE BAGATVR-VN GER MON UU ?

## 4 实验结果及分析

对实验涉及到的所有语料库进行语料库预处理，得到规范的语料库。用规范的语料库进行训练，然后对 CWMT2011 提供的蒙古文测试语料库进行翻译，对得到的翻译译文，用 CWMT2011 提供的在线评测工具进行评测，其结果如表 1 所示。

**表 1. 基准实验评测结果**

| 评测项 | NIST6 | NIST7 | BLEU5 | BLEU6 | BLEU5_SBP | GTM | mWER | mPER | ICT |
|---|---|---|---|---|---|---|---|---|---|
| 实验结果 | 5.5970 | 5.6032 | 0.1784 | 0.1369 | 0.1624 | 0.5951 | 0.5881 | 0.4951 | 0.4624 |

蒙古文是个形态变化丰富的语言。其中一个导致形态变化的因素就是附加成分。不同的变形附加成分将同一个词变形为不同的词形，这对统计建模带来了不利，如：数据稀疏问题进一步恶化，词表过于庞大等。

在蒙古文拉丁转写方案中，对词干与分写的附加成分之间用“-”符号连写，而且该符号在其它地方不会出现。

所以，利用“-”进行分写的附加成分的切分，其准确率达到100%。

对规范的蒙古文语料库进行分写的附加成分切分处理后，进行训练，然后对CWMT2011提供的蒙古文测试语料库也进行同样的切分处理之后，进行翻译，对得到的翻译译文，用CWMT2011提供的在线评测工具进行评测，其结果为如表2所示。

**表2. 对蒙古文的分写附加成分进行切分处理后的评测结果**

| 评测项 | NIST6 | NIST7 | BLEU5 | BLEU6 | BLEU5_SBP | GTM | mWER | mPER | ICT |
|---|---|---|---|---|---|---|---|---|---|
| 实验结果 | 5.7263 | 5.7323 | 0.1819 | 0.1357 | 0.1643 | 0.6072 | 0.5582 | 0.4786 | 0.4879 |

从评测结果看，对蒙古文的分写附加成分进行切分处理后，其bleu值大约提高了0.002。该增值小于我们之前做实验得到的增值(约0.03)。为了找到其中原因，我们做了如下测试：

（1） 把翻译结果中的未登录词转写成老蒙古文，其评测结果为如表3所示。

**表3. 把译文中的未登录词转写成老蒙古文后的评测结果**

| 评测项 | NIST6 | NIST7 | BLEU5 | BLEU6 | BLEU5_SBP | GTM | mWER | mPER | ICT |
|---|---|---|---|---|---|---|---|---|---|
| 实验结果 | 5.7299 | 5.7361 | 0.1826 | 0.1365 | 0.1658 | 0.6093 | 0.5532 | 0.4756 | 0.4923 |

（2） 把翻译结果中的未登录词删除之后，其评测结果为如表4所示。

**表4. 把译文中的未登录词删除后的评测结果**

| 评测项 | NIST6 | NIST7 | BLEU5 | BLEU6 | BLEU5_SBP | GTM | mWER | mPER | ICT |
|---|---|---|---|---|---|---|---|---|---|
| 实验结果 | 5.7351 | 5.7417 | 0.1895 | 0.1432 | 0.1721 | 0.6213 | 0.5454 | 0.4731 | 0.5210 |

从这两个评测结果来看，对CWMT2011蒙汉日常对话测试集进行翻译时，如果蒙古文语料库采用老蒙古文形式的话，评测结果可能更好，故进行了下面的第3组测试。

（3） 我们用老蒙古文形式再做了一次同样的一组实验。其基准实验和对蒙古文的分写附加成分进行切分处理后的评测结果分别如表5和表6所示。

**表5. 用老蒙古文形式的基准实验评测结果**

| 评测项 | NIST6 | NIST7 | BLEU5 | BLEU6 | BLEU5_SBP | GTM | mWER | mPER | ICT |
|---|---|---|---|---|---|---|---|---|---|
| 实验结果 | 5.3035 | 5.3114 | 0.1694 | 0.1272 | 0.1535 | 0.5757 | 0.5876 | 0.5042 | 0.46 |

**表6. 对蒙古文的分写附加成分进行切分处理后的评测结果**

| 评测项 | NIST6 | NIST7 | BLEU5 | BLEU6 | BLEU5_SBP | GTM | mWER | mPER | ICT |
|---|---|---|---|---|---|---|---|---|---|
| 实验结果 | 5.6596 | 5.6668 | **0.1991** | **0.154** | **0.1792** | 0.5993 | 0.5666 | 0.4873 | 0.4796 |

该评测结果表明，对CWMT2011蒙汉日常对话测试集翻译时，在蒙古文的分写附加成分进行切分处理实验中，用老蒙古文训练得到的译文评测结果，比用蒙古文拉丁转写形式训练得到的译文评测结果更好。用蒙古文拉丁转写形式的情况下，分写附加成分切分处理实验结果不理想的原因是老蒙古文到拉丁蒙古文的转换工具对控制符的处理不恰当，导致一些控制符丢失。

从老蒙古文到拉丁转写程序存在控制符丢失问题：蒙古文变体选择符1的拉丁转写形式为[‘]；蒙古文变体选择符2  的拉丁转写形式为[“]；蒙古文变体选择符3  的拉丁转写形式为[`]；但转换的时候把这些拉丁字符丢失了。具体例子见表7所示。

**表7. 拉丁转写程序转换时丢失控制符举例**

| 蒙古文 | 输入形式 | 正确拉丁转写形式 | 错误拉丁转写形式 |
|---|---|---|---|
|  | duN | d’ uN | DUn |
|  | biqig | biqig” | BICIG |
|  | oghu | og`hu | OGHU |

## 5总结

本文介绍了内蒙古师范大学参加 CWMT2011 蒙汉日常用语翻译评测的基于短语的蒙汉统计机器翻译系统情况。为了提高蒙汉机器翻译性能，在蒙古文语料库上进行了分写的附加成分切分处理。

在实验过程中，我们发现老蒙古文和拉丁蒙古文之间相互转换工具不够成熟，影响了蒙古文统计机器翻译系统的性能，所以，完善和改进该工具是非常必要的。另外，在蒙汉语料的训练集，开发集和测试集的蒙古语拼写上都存在一些不一致现象，所以，我们需要进一步提高蒙汉机器翻译的语料库质量。

目前，我们的蒙古文切分标注工具还不够成熟，导致蒙古语语料库的加工处理受到了阻碍。所以，我们下一步工作重点放在蒙古文切分标注系统的完善上，以便在统计机器翻译模型中融入更多的蒙古文形态学信息提高译文质量。

## 参考文献


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.

那顺乌日图, 刘群, 巴达玛敖德斯尔. 关于“汉蒙机器辅助翻译系统”[J]. 韩国阿尔泰学报. 2001(11):135-141.

王斯日古楞. 基于混合策略的汉蒙机器及相关技术研究[D]. 呼和浩特:内蒙古大学蒙古学学院, 2009.

银花. 基于短语的蒙汉统计机器翻译研究[D]. 呼和浩特:内蒙古师范大学, 2011.

宗成庆. 机器翻译研究进展:第四届全国机器翻译研讨会论文集. 北京, 2008.