# 2013 全国机器翻译研讨会计算所评测技术报告

# The ICT Technique Report of All Language Tracks of CWMT 2013

李响, 宋林峰, 孟凡东, 刘凯, 张海波, 王星, 蔡洽吾, 陈宏申, 刘洋, 杨似彤, 吕雅娟, 刘群
中国科学院计算技术研究所 智能信息处理重点实验室 北京 100190
Xiang Li, Linfeng Song, Fandong Meng, Kai Liu, Haibo Zhang, Xing Wang, Qiawu Cai, Hongshen Chen, Yang Liu, Sitong Yang, Yajuan Lü, Qun Liu
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology Chinese Academy of Sciences
P.O.Box 2704,Beijing,China,100190

E-mail:{lixiang, songlinfeng, mengfandong, liukai, lvyajuan, liuqun}@ict.ac.cn

## 摘要

本文简要地介绍了中国科学院计算技术研究所自然语言处理研究组在 2013 年全国机器翻译研讨会机器翻译评测中所使用的机器翻译技术及相应语言处理技术。今年本组参加了本次评测的所有评测任务。对于不同的语言对，我们采取了不同的语言处理技术及对应的翻译策略。本文简要地介绍了各个语言对相应的针对性的技术。对于所有语言，我们使用了多种主流的统计翻译模型，并利用词级和句级的融合方法对上述不同的模型及系统进行融合得到最终结果。本文简要地介绍了各个系统的理论模型和系统框架，并对评测相关实验进行了说明。

## Abstract

This paper describes the ICT systems involved in the tracks fie of CWMT 2013 evaluation campaign. We participated in all the tracks. We adopt different techniques and strategies for different tracks in this campaign. This paper briefly describes all the specific techniques used for each track. For all the tracks, we use various popular SMT models, including formally syntax-based system and phrase-based system. Then we adopt various system combination techniques to combine the results of all the systems. We will describe the framework and model of these systems and report the results on both development and test sets.

## 1 引言

2013 年中国机器翻译研讨会（CWMT2013）机器翻译评测共含六个项目，分别为：英汉新闻领域，汉英新闻领域，英汉科技领域，维汉新闻领域、藏汉政府领域、蒙汉口语领域。中科院计算所参加了所有六个项目，并取得了优秀的成绩。其中维汉、蒙汉、英汉新闻为所有参评单位的第一，藏汉、英汉科技为第二。

## 2 参评翻译系统描述

这次机器翻译评测中我们使用了四种统计机器翻译系统：Silenus、Chiero、Moses-Chart和Moses。其中，Silenus是内部实现的基于压缩森林到串模型的翻译系统；Chiero和Moses-chart分别是内部实现的和开源的基于层次短语模型的系统；Moses是开源的基于短语模型的翻译模型。

### 2.1 Silenus

Silenus[Mi et al.,2008;Mi and Huang,2008]为基于句法森林的树到串模型，衍生于单树树到串系统Lynx[Liu et.al, 2006,2007]，最关键的不同在于Silenus在规则抽取和解码时使用共享压缩森林而不是1-best句法分析树。

**规则抽取:**

在给定上下文无关语法下，一个句子的句法森林是其所有可能的推导（即，句法分析树）的紧凑表示形式[Billot and Lang,1989]。压缩森林可以形式化定义为一个四元组 Hp=<V, E, t, R>, 其中：V 为有限的节点集合，E 为有限的超边集合， 是句法分析根节点，R 为权重集合。给定一个句子 $w_{1:n}=w_1 \ldots w_n$，v∈V 的每个节点形式化表示为 $X_{i,j}$（表示已经识别的跨度为 的非终结符）。每个超边 e∈E 为一个三元组 e= <T(e), h(e), f(e)>，其中 h(e)∈V 为超边的头节点，T(e)∈V*为超边的尾节点向量，f(e)为 R|T(e)|到 R

的权重函数。图 1 给出了一个中文森林和英文串的例子。

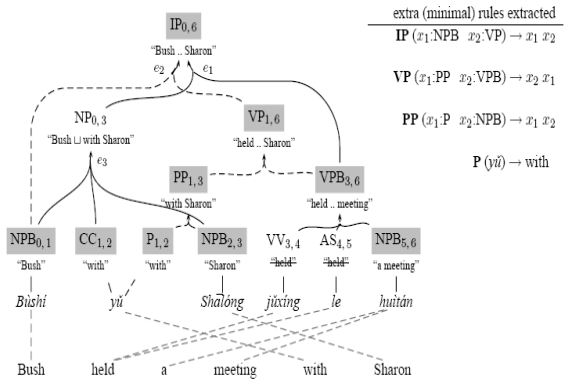


图 1: 源语言端压缩森林和目标语言端串

**解码：**

在给定句法分析森林 $H_p$ 下，Silenus 首先使用匹配转换算法将句法分析森林转换成翻译森林 $H_t$，然后利用树替换语法将源语言的翻译森林转换为目标语言串搜索最优推导（翻译规则序列）。

给定一个句法森林 $H_p$ 和一个翻译规则集合，转换算法最基本的想法是自顶向下遍历整个句法森林的每一个节点 v，然后对每一条规则尝试当前以 v 为根节点的子森林，如果匹配成功则生成相应的翻译超边。相应的翻译森林如图 2 所示。

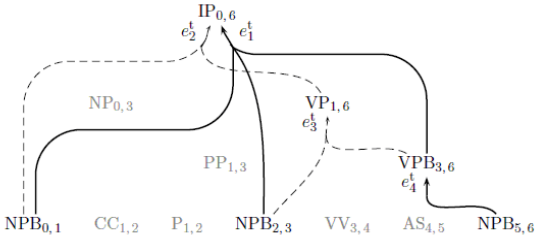


图 2：相应的翻译森林

## 2.2 Chiero

Chiero 系统是基于层次短语模型[Chiang, 2007]实现的。层次短语模型可以认为是基于短语的模型的扩展，其可以抽取源句子中非连续的部分，并将其翻译为目标句子的非连续部分。该模型可以形式化为同步上下文无关文法（SCFG），其规则形式为：$X \rightarrow \langle \gamma, \alpha, \sim \rangle$，其中，$X$为非终结符，$\gamma$ 和 $\alpha$ 为终结符和非终结符组成的原端串和目标端串，~为 $\gamma$ 和 $\alpha$ 中非终结符的一一对应关系。

## 2.3 Moses

Moses [Koehn et al.,2007]是由英国爱丁堡大学、德国亚琛工业大学等8家单位联合开发的一个统计机器翻译系统，我们使用了其中包括了基于短语的系统（后面简称Moses）和基于层次短语的系统（后面简称Moses-chart）。

## 2.4 系统融合

### 2.4.1 IHMM-wordcomb：基于 IHMM 的词级系统融合

IHMM-wordcomb 是基于 He 等人[He et.al,2008]的工作实现的词级系统融合系统。

首先会针对每个系统构建一个混淆网络，所有单系统的混淆网络构成一个多混淆网络，再采用一些特定的特征搜索多混淆网络，分值最高的路径对应的译文便作为基于 IHMM 的系统的输出。

对于某个单系统对应的单混淆网络，构建过程如下：

(1) 用 Minimum Bayes Risk 从该单系统的 nbest 中选择一个代价最小的候选翻译作为 skeleton，选取方法如下：

$$E_s = \arg\min_{E \in E_i} \sum_{j=1}^{N_s} \mathrm{TER}(E_j, E_i)$$

(2) 用基于 Indirect-HMM 的对齐方法将所有单系统的 nbest 与 skeleton 进行对齐;

(3) 以 skeleton 的单词顺序为基准，将所有的候选翻译根据与 skeleton 的对齐进行 normalize，如果对齐到空，则要在适当的位置添加空词;

(4) 逐句的将 normalize 后的句对加入好混淆网络中;

(5) 分别添加一个头结点和尾结点，将所有单混淆网络合成一个多混淆网络。

搜索多混淆网络时，采用如下特征计算分数：每个词的概率，语言模型，非空词个数，非空词个数。采用 Powell 算法调参。

### 2.4.2 TER-wordcomb：基于 TER 的词级系统融合

TER-wordcomb 是基于 Rosti 等人 [Rosti et.al, 2007]的工作实现的词级系统融合系统。任意选取一个系统，从不重复的 N-best（具体取 10）中选取第一个翻译作为混淆网络的骨架，然后将所有系统输出的 N-best 均与该骨架对齐，构造基于当前系统的混淆网络。此过程依次作用于每个系统。原文的方法为了在 N-best 中选择一个合适的骨架，需要多次计算所有翻译与候选骨架之间的 TER 对

齐，这在实际使用中比较消耗时间，我们的实现中直接选择第一个翻译结果，实际使用中速度较快，可忽略损失。

我们采用线性模型，使用的特征包括:

(1) 构成当前候选翻译的词语的置信度分数之和，具体请参考 [Rosti et al., 2007]；

(2) 语言模型的权重；

(3) 空词数惩罚；

(4) 非空词个数惩罚。

其中参数调整的过程，本文方法采用MERT算法，目标函数是BLEU-SBP。同时将四个系统权重设置为1，为了在实际中获得适中且稳定的提高。而采用BLEU-SBP最大化作为训练目标的原因是为了消除开发集中句子的长度不一而对参数的训练的影响，请参考[Chiang et.al, 2008]，我们实验证明采用BLEU-SBP比采用BLEU效果更好。

#### 2.4.3 SentComb:句级系统融合

SentComb是基于[Macherey and Och,2007]的工作实现的句级系统融合系统，使用global linear模型对合并后的n-best结果进行重排序：

$$\hat{y} = \arg\max_{y \in GEN(x)} f(x,y) \cdot W$$

其中，$x$ 为源语言句子，$y$ 为翻译结果，$f(x,y)$为特征向量，$W$ 为权重向量，$GEN(x)$为可能的候选译文结果的集合。

所采用的特征类型包括：

(1) 与其他候选译文的相对BLEU值；

(2) 语言模型分数；

(3) 译文结果的长度。

各个特征权重由最小错误率训练得到。

## 3 少数民族语言处理技术

本节主要介绍藏语、维语和蒙语等少数民族语言的处理技术。

### 3.1 拼音文字拉丁转换（维语、蒙语）

由于维语及蒙语为拼音文字并且形式较为独特，难以被非该语种使用者理解并进一步处理。因此，我们将传统的维语及蒙语转换成较易接受的拉丁形式。其中维语的拉丁转换与新疆大学合作完成，转换标准依据维语相应的国家拉丁标准。目前维语转换工具支持utf8老文字维文到utf8拉丁维文的转换。而蒙语转换参照内蒙古大学制定的“内大拉丁”标准形式，相应转换工具与内蒙古大学合作完成。目前蒙语转换拉丁工具可以支持utf8编码及蒙科立编码的老蒙文到拉丁蒙文的转换。

### 3.2 维语词法分析

维语的词法分析使用的是计算所研制开发的基于判别式有向图的维语词法分析器。通过判别式方法来对当前的词素特征进行建模，既考虑词内词素间的关系，也考虑词间词素的关系，可以包含更多的全局特征。基于判别式有向图的词法分析模型，可以提高维语词法分析质量，尤其是 OOV 的切分结果。在整词切分上，维语的切分正确率达到了 96. 10%。整个分析流程如图 3 所示。

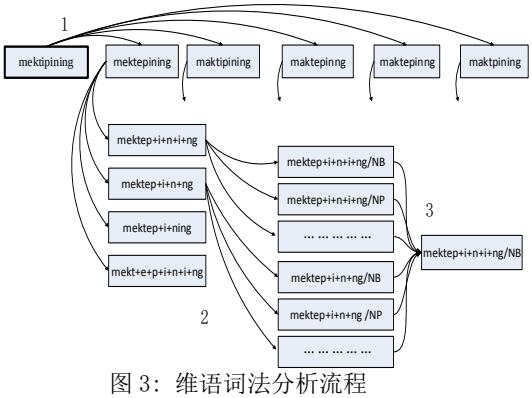


图 3：维语词法分析流程

### 3.3 蒙语词法分析

蒙语的词法分析使用的是计算所研制开发的蒙语的词法分析器，依靠基本的规则对蒙语（内大拉丁形式）进行准形态分析，具有速度快，准确率高的优点。

### 3.4 藏语分词

本次评测采用计算所研制开发的基于统计的藏语分词系统，分词的F值可以达到90%以上。首先按照音节以及一定的启发式信息将藏文切分为最小分词粒度，因此藏文的分词可以作为序列标注任务。采用感知机作为序列标注的模型，进行藏文的粗切分，并生成词图以存储藏文句子的多种切分结果。最后利用词典信息赋予词图中边不同的权重，采用最短路径的动态规划算法求解出最优的切分结果。

### 3.5 维吾尔语人名翻译

使用计算所研制开发的维吾尔于的汉语人名翻译系统对汉语人名进行音译翻译。基本原理就是使用语言模型选择符合汉族人名习惯的人名。由于本次评测测试集中未出现

较为明显的汉语人名，在最终提交结果中我们名没有采用维语人名翻译模块，但是该模块在实际线上系统中具有较好的效果。

### 3.6 拼写校对系统

在本次评测中，我们对蒙语及维语采用了拼写校对系统。与一般拼写校对系统不同的是，本拼写校对系统采取“联合形态切分拼写校对”技术。即例如维语“1931-yilidiki”在词法分析后我们将会将其分析为“1931-yil i diki”，由于人书写的词汇容易出现拼写错误并且形态分析器存在一定的错误率，实际情况中极容易出现类似情况：“1931-yildiki”(拼写错误)或者形态切分结果错误地切分为“1931-yili diki”(形态切分错误)。显然“1931-yildiki”、“1931-yili diki”在包括空格的前提下与正确结果“1931-yil i diki”的编辑距离为1。如果使用原始的不考虑空格（多个词的情况）拼写校对系统，“1931-yildiki”与形态分析后的正确拼写编辑距离将会很大，并且“1931-yili diki”拼写校对的结果将会是“1931-yil diki”缺少了“i”。因此，我们开发了针对多词的“联合形态切分拼写校对”技术，在形态切分的同时考虑相应的拼写校对问题。在本次评测中的拼写校对系统仅考虑未登入词及两个词的情况。

### 3.7 蒙、维、藏的数字时间词识别及翻译

在本次评测中，我们使用与相关合作单位合作研发的规则方法为基础的数词时间词识别翻译工具对这些少数民族语言进行数字和时间词识别及翻译。

## 4 数据处理及使用

### 4.1 英汉新闻方向

英汉新闻评测使用官方提供的训练语料的other data 部分（CLDC-LAC、Datum、HIT-IR、HIT-MT、ICT-web、NEU、XMU），采用评测官方提供的开发集做开发，对包括训练集、开发集、测试集在内的所有语料的中文端进行了编码统一、全半角转换、转义字符替换和 ictclas 分词，对应的英文端进行了编码统一、全半角转换、转义字符替换、token 和全小写转换。除了采用 ictclas 切分，我们另外对原始语料的汉语部分采用 Stanford 切分，跟采用 ictclas 切分的语料一样，我们用这些语料训练每一个单系统，在最终系统融合中，我们额外添加用 Stanford 切分的语料训练的、在开发集和内部测试集上性能突出的单系统作为子系统进行系统融合，以增加子系统的多样性。

我们只对主系统的测试集翻译结果的英文单词进行 TrueCase 还原，而一般的对比系统的测试集翻译结果我们不做任何后处理。

### 4.2 英汉科技方向

英汉科技评测使用官方提供的训练语料（2008&2009 ISTIC、2011 ISTIC），采用评测官方提供的开发集做开发，对包括训练集、开发集、测试集在内的所有语料的预处理方式与“4.1 英汉新闻方向”采用的相同。特别的是，我们并没有对测试语料中圆括号“（）”内的单词进行全小写转换，而是保留了原有的大小写风格。

对测试集翻译结果的 OOV 我们采用如下策略处理：首先，如果该词原先是带有大写字符的，那么我们对它进行 TrueCase 还原；其次，我们对剩下的 OOV 进行了“形态枚举”，我们枚举三种形式的变化：二分切分、词性变换、编辑距离。其中二分切分考虑将一个词分成两个词的所有可能情况，间隔符只考虑空格或“-”符，比如“preprocess”可以变成“pre process”或“pre-process”。词形变换考虑词的时态、单复数形态和词性变换，比如“netcards”可以变成“netcard”，“transliteration”可以变成动词形式“transliterate”。对于编辑距离，由于编辑距离导致的歧义性较大，这里只考虑原词的所有编辑距离为 1 的词，编辑距离操作包括：增加、删除、修改、相邻字母调换这四项操作。最后，对于前两个方法都处理不了的 OOV，我们加以保留。

我们只对主系统的测试集翻译结果的英文单词进行 TrueCase 还原，而一般的对比系统的测试集翻译结果我们不做任何后处理。

### 4.3 汉英新闻方向

英汉新闻评测使用官方提供训练语料的other data 部分（CLDC-LAC、Datum、HIT-IR、HIT-MT、ICT-web、NEU、XMU），采用评测官方提供的开发集做开发，对包括训练集、开发集、测试集在内的所有语料的中文端进行了编码统一、全半角转换、转义字符替换和 ictclas 分词，对应的英文端进行了编码统一、全半角转换、转义字符替换、token 和全小写转换。我们也额外对语料的汉语部分采用 Stanford 切分，在最终系统融合

（句级和 TER 词级）中，我们额外添加在开发集和内部测试集上性能突出的使用 Stanford 切分的单系统作为子系统，以增加子系统的多样性。对于官方提供的训练语料，我们还根据标记将其分为两类，新闻部分（News）和其他部分（Others），在 Silenus 单系统上，我们尝试分为用新闻部分和其他部分语料分别抽取规则，然后做翻译规则表平滑（Smooth），即将 Others 规则表中源端不出现在 News 规则表中的规则添加到 News 规则表中，构成一个新规则表。

我们只对主系统的测试集翻译结果的英文单词进行 TrueCase 还原。

### 4.4 蒙汉方向

蒙语语料使用评测方提供的训练语料（内蒙古大学汉蒙平行语料库，中国科学院合肥智能机械研究所蒙汉双语语料）。

前处理包括把蒙古文转写成内大拉丁蒙文形式，拉丁蒙文的 Tokenization。

针对拉丁蒙语存在的同一单词具有不同书写形式的问题，我们对语料进行了拼写校对和形式归一化，例如后缀 SAN 和 SEN 意思相同，转为其中一个即可，可以在一定程度上解决训练语料中存在的数据稀疏问题。

由于蒙古文是一种形态变化丰富的语言，不同的形态成分将同一个词变形为不同的词，导致了数据稀疏问题，因此我们除了拉丁蒙文词级系统外，还采用了拉丁蒙文的词干级系统来缓解这种问题针。

另外，对转换后的拉丁蒙语的同一单词存在个别原因导致的拉丁形式的不统一。因此我们使用将拉丁蒙文全部归一化转为小写形式进行了尝试，即蒙语小写词级系统。

汉语端采用按字切分的方式进行处理。

我们采用了 GIZA++和 Berkeley Aligner 的融合对齐结果用来抽取短语。

### 4.5 维汉方向

维吾尔语语料使用评测方提供的训练语料（新疆大学维汉双语句子对齐语料库(2013 版)，中国科学院新疆理化技术研究所维汉双语语料库）。

维吾尔语处理包括老维文到拉丁维文的转码、Tokenization、处理不合理符号，以及在此基础之上进行的形态分析、翻译粒度选择等。

针对拉定维语存在的同一单词具有不同书写形式的问题，我们对语料进行了联合形态分析及拼写校对和还原。

我们采用 2 种粒度的系统：词干词缀级和词干级。

汉语端采用按字切分的方式进行处理。

我们采用 GIZA++的对齐结果抽取短语。

### 4.6 藏汉方向

藏语语料使用评测方提供的训练语料（青海师范大学藏汉平行语料库，央金藏汉平行语料库，西北民族大学、西藏大学与厦门大学藏汉语料）。

藏文在逻辑语法体系中属于拼音文字，但是词与词之间没有明显的间隙，因此藏文分词是藏文前处理的必要步骤。

汉语端采用计算所开发的 ICTCLAS 汉语分词工具进行分词。

我们采用了 GIZA++和 Berkeley Aligner 的对齐结果作为最终对齐用来抽取短语。

### 4.7 语料通用处理技术

除了各少数民族语言特有的语言处理技术，我们对于所有语言采用了其他通用处理技术，主要包括：全角转半角，繁体转简体，非法和特殊符号的处理，数字、序号和时间符号的处理，数字格式化，句首无效标点的处理，长度比过大的平行语料处理和校对等。

### 4.8 语言模型

除了所有语言对自身训练语料的语言模型外，目标端为中文的使用了搜狗新闻语料库，目标端为英文的使用了路透社语料。

语言模型统一采用 SRILM 训练的 5 元模型，采用 modified kndiscount 平滑技术和 interpolate 差值技术，在训练过程中保留所有 n-gram。

### 4.9 词语对齐

对齐工具主要使用 GIZA++ 和 Berkeley Aligner。其中 GIZA++的训练参数采用 Moses 训练脚本的默认参数配置；Berkeley Aligner 除采用 10 轮外迭代训练，其他配置均为默认配置。

## 5 实验结果

所有实验都是基于评测方提供的相应的开发集进行的。其中少数民族语言的三个任务的开发集的评测指标采用 BLEU5-SBP，其他任务的开发集的评测指标采用 BLEU-4，所有任务的测试集的评测指标均采用 BLEU5-SBP，以“△”号为标记的结果是最后提交的主系统结果。

少数民族的三个方向的翻译结果均为去除未登录词的结果。

对于英汉汉英的三个方向：在调参阶段和测试阶段，我们采用按照开发、测试集过滤后的训练集作为实际训练集，我们采用训练集的目标端训练第一语言模型，我们采用搜狗语料或路透社语料训练第二语言模型，所有语言模型均为 5 元。

对于少数民族语言任务：在调参阶段，我们采用过滤开发集的训练集作为实际训练集，语言模型采用过滤开发集的训练集目标端训练。在测试阶段，我们采用全部训练集合和开发集作为实际训练集，语言模型采用训练集和开发集的目标端训练。

### 5.1 英汉新闻领域机器翻译评测结果与分析

在评测组织方提供的开发集下，各个系统在开发测试集上的结果见下表：

表 1: 英汉新闻领域机器翻译结果

| | 开发集 | progress cwmt09 | current cwmt11 |
|---|---|---|---|
| 参评系统 | BLEU-4 | BLEU5-SBP | BLEU5-SBP |
| Chiero | 0.2735 | 0.3536 | 0.3316 |
| Chiero +Stanford 切分 | 0.2765 | 0.3421 | 0.3205 |
| Silenus | 0.2724 | 0.3638 | 0.3425 |
| MosesChart | 0.2718 | 0.3448 | 0.3318 |
| MosesChart +Stanford 切分 | 0.2759 | 0.3528 | 0.3364 |
| 句级系统融合 | 0.2942 | 0.3687△ | 0.3528△ |

系统融合是以上所有单系统的融合结果。我们只对主系统结果中的英文单词做 TrueCase还原，最终提交主系统的结果为大小写还原及句子级系统融合叠加后的结果。

从结果中可以看出基于森林到串的系统在测试集上表现出明显的优势，我们分析这是因为句法系统在调序上较非句法系统更有优势。与此同时，采用 Stanford 切分的 MosesChart系统也表现出不错的性能，并且在系统融合中贡献了多样性。

### 5.2 英汉科技领域机器翻译评测结果与分析

在评测组织方提供的开发集下，各个系统在开发测试集上的结果见下表。

表 2：英汉科技领域机器翻译结果

| | 开发集 | progress 测试集 | current 测试集 |
|---|---|---|---|
| 参评系统 | BLEU-4 | BLEU5-SBP | BLEU5-SBP |
| Chiero | 0.4075 | 0.3873 | 0.3576 |
| Silenus | 0.4249 | 0.4015 | 0.3856 |
| Moses | 0.4131 | 0.3975 | 0.3590 |
| MosesChart | 0.4124 | 0.3920 | 0.3577 |
| 句级系统融合 | 0.4425 | 0.3929△ | 0.3976△ |

系统融合是以上所有单系统的融合结果。我们只对主系统结果中的英文单词做 TrueCase和“形态枚举”还原，最终提交主系统的结果为大小写还原及句子级系统融合叠加后的结果。

从结果中可以看出基于森林到串的系统在测试集上表现出明显的优势，我们分析这是因为句法系统在调序上较非句法系统更有优势。

### 5.3 汉英新闻领域机器翻译评测结果与分析

在评测组织方提供的开发集下，各个系统在开发测试集上的结果见表 3，如下:

表 3：汉英新闻领域机器翻译结果

| | 开发集 | progress cwmt09 |
|---|---|---|
| 参评系统 | BLEU-4 | BLEU4-SBP |
| Chiero | 0.2789 | 0.1784 |
| Chiero+Stanford 切分 | 0.2753 | 0.1916 |
| Moses | 0.2678 | 0.1729 |
| Moses+Stanford 切分 | 0.2660 | 0.1747 |
| Silenus | 0.2888 | 0.1915 |
| Silenus Smooth | 0.2862 | 0.1955 |
| 句级系统融合 | 0.2998 | 0.1997 |
| TER 词级系统融合 | 0.3240 | 0.2334△ |

系统融合是以上所有单系统的融合结果。我们只对主系统结果做完整的英文 TrueCase 还原，所以最终提交主系统的结果为大小写还原及 TER 词级系统融合叠加后的结果。

从结果中可以看出，在汉英新闻领域，TER 词级系统融合的结果要远好于句级系统融合。与此同时，在森林到串系统上，采用 Smooth 技术要比一般使用全部语料效果好，我们分析原因是因为测试集是新闻领域的，较一般使用全部语料，采用 Smooth 技术更能偏向新闻领域训练语料。最后，在 Chiero 和 Moses 系统上，Stanford 切分的效果要比 ictclas 切分效果好。

### 5.4 藏汉政府领域机器翻译评测结果与分析

在评测组织方提供的开发集下，各个系统在开发测试集上的结果见下表。

表4: 藏汉政府领域机器翻译结果

| | 开发集 | 测试集 |
|---|---|---|
| Moses | 0.6251 | 0.2472 |
| Moses-Chart | 0.6555 | 0.2446 |
| Chiero | 0.6299 | 0.2485 |
| 系统融合 | 0.6651 | 0.2504△ |

系统融合是以上所有单系统的融合结果。

### 5.5 维汉新闻领域机器翻译评测结果与分析

在评测组织方提供的开发集下，各个系统在开发测试集上的结果见表 5。

最后的系统融合结果是在所有不同粒度融合结果上再次进行融合的结果。

### 5.6 蒙汉日常口语领域机器翻译评测结果与分析

在评测组织方提供的开发集下，各个系统在开发测试集上的结果见表 6。

最后的系统融合结果是所有不同粒度融合结果上再次进行融合的结果。

由于此次评测中，蒙汉翻译结果较好。为了进一步解释蒙汉翻译性能并验证相关技术的可靠性，我们后期就各项相应评测技术进行了一系列的对比实验。其中翻译系统采用的 Moses 层次短语系统，对于目标端汉语统一采用按字切分方式，评价指标仍然采用 BLEU5-SBP。具体实验结果如表 7 所示。

以下分别解释表 7 中所使用的不同的系统配置：

- 基线
  只对汉语端做了切字，其他语料未做任何处理。
- 拉丁
  将蒙文通过 3.1 节中的方式转为拉丁形式，以下系统配置都将训练集根据开发集进行了深度过滤，减少了 230 句左右的训练语料句数，否则会导致系统严重过拟合。
- token
  对拉丁蒙文进行准形态分析，如 3.3 节中所述。
- 小写
  对 token 后的拉丁蒙文全部转为小写，减小语料得数据稀疏问题同时会带来部分词汇歧义问题。
- berkeley
  简单将 giza++和 berkeley 对齐结果融合并用来抽取短语的翻译结果。
- mysri
  使用了非过滤方式对训练集训练得到的语言模型，训练参数如灰箱评测数据所示。
- 搜狗
  加入搜狗语言模型作为第二语言模型，其中 2009 及 2012 年搜狗全网语料库均有使用。
- 语料处理
  对语料进行部分预处理工作，例如简单的根据长度的语料筛选等。
- 其他
  用全部训练语料和开发集抽取短语表和训练语言模型，用于最后的测试解码，同时针对测试集翻译结果出现的未登录词进行了拼写校对和还原。

通过以上实验，我们可以得出结论，蒙文的拉丁转换对于性能有非常重要的影响，通过分析转换表可以发现，在转换中多个不同的蒙文字会转换到相同的拉丁形式，减少了数据的稀疏性，极大提高了翻译性能；同时拉丁蒙文 tokenization 和小写化，融合不同对齐结果，语言模型规模的扩大及训练方式的优化，以及语料处理技术都有效提升了最终的翻译性能。需要注意的是，以上验证实验的目的是解释我们的蒙汉系统的性能，和参赛期间训练的系统有所差别。

表 5: 维汉新闻领域机器翻译结果

| | Moses | | Moses-Chart | | Chiero | | 同粒度系统融合 | |
|---|---|---|---|---|---|---|---|---|
| | 开发集 | 测试集 | 开发集 | 测试集 | 开发集 | 测试集 | 开发集 | 测试集 |
| 词干级 | 0.5111 | 0.5000 | 0.5122 | 0.5112 | 0.5218 | 0.5016 | 0.5227 | 0.5087 |
| 词干词缀 | 0.5054 | 0.5002 | 0.5169 | 0.5082 | 0.4998 | 0.4944 | 0.5281 | 0.5089 |
| 系统融合 | | | | | | | 0.5276 | 0.5168△ |

表 6: 蒙汉日常口语领域机器翻译结果

| | Moses | | Moses-Chart | | Chiero | | 同粒度系统融合 | |
|---|---|---|---|---|---|---|---|---|
| | 开发集 | 测试集 | 开发集 | 测试集 | 开发集 | 测试集 | 开发集 | 测试集 |
| 词级 | 0.3607 | 0.1777 | 0.3733 | 0.1797 | 0.3610 | 0.1842 | 0.3857 | 0.1914 |
| 小写词级 | 0.3591 | 0.1781 | 0.3760 | 0.1851 | 0.3652 | 0.1820 | 0.3843 | 0.1891 |
| 词干级 | 0.3576 | 0.1655 | 0.3673 | 0.1714 | 0.3622 | 0.1756 | 0.3805 | 0.1853 |
| 系统融合 | | | | | | | 0.3907 | 0.1964△ |

表 7: 不同配置的蒙汉系统实验

| 系统配置 | 测试集 |
|---|---|
| 基线 | 0.0745 |
| 基线+拉丁 | 0.1356 |
| 基线+拉丁+token | 0.1447 |
| 基线+拉丁+token+小写 | 0.1462 |
| 基线+拉丁+token+小写+berkeley | 0.1678 |
| 基线+拉丁+token+小写+berkeley+mysri | 0.1626 |
| 基线+拉丁+token+小写+berkeley+mysri+搜狗 | 0.1655 |
| 基线+拉丁+token+小写+berkeley+mysri+搜狗+语料处理 | 0.1714 |
| 基线+拉丁+token+小写+berkeley+mysri+搜狗+语料处理+其他 | 0.1727 |

## 6 总结

通过评测结果我们可以发现，系统融合技术对于改善翻译结果有一定的帮助，词级系统融合和句子系统融合在不同的语言对上表现出不同的效果,加入不同切分产生的单系统结果对系统融合有正面的影响。对于少数民族语言而言，语言处理技术仍然是提升翻译质量的关键手段：编码转换、词法分析、拼写校对等技术均能够显著提升翻译效果。在词法分析后，通过对黏着语进行相同粒度内部的系统融合能够提升翻译结果，在此之后再进行多粒度融合能够进一步得到更好的翻译结果。

## 致谢

此次评测中少数民族语言翻译项目所用到的部分技术是在各少数民族合作单位的大力支持下完成的。特别感谢内蒙古大学的那顺乌日图老师，乌日力嘎和苏传捷等同学；新疆大学的吐尔根和麦热哈巴老师，米日姑和米莉万等同学；青海师范大学的才让加、华却才让等老师。少数民族语言翻译的相关研究及进展均离不开这些老师及同学的大力帮助和支持，在此表示感谢！

## 参考文献

Gulsen Eryigit, Kemal Oflazer. Statistical Dependency Parsing of Turkish[C]. 11th Conference of the European Chapter of the Association for Computational Linguistics(EACL). Trento, Italy, 2006: p.89-96.

Ryan McDonald, Koby Crammer, Fernando Pereira. Online Large-Margin Training of Dependency Parsers [C]. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). University of Michigan, USA, 2005: p.91-98.

Wenbin Jiang, Qun Liu. Dependency Parsing and Projection Based on Word-Pair Classification[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). Uppsala, Sweden 2010: p.12-20.

Klein, D. and Manning, C.D. Corpus-based induction of syntactic structure: Models of dependency and constituency[C]. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL). Barcelona, Spain 2004: p.478-485.

王志洋. 面向形态丰富语言的机器翻译关键技术研究[D]. 中国科学院大学（计算技术研究所）,2013.

姜文斌等，蒙古语词法分析的有向图模型[C]。中文信息学报。2010 3 卷.

玉素甫·艾白都拉，维语句法分析器中的词义排岐问题的研究. 计算机应用与软件. 2002 年 19 卷 4 期: p59-62.

力提甫·托乎提，从短语结构到最简方案——阿尔泰语言的句法结构[M]. 北京：中央民族大学出版社，2004.

亚热·艾拜都拉，维吾尔语句法分析方面存在的一些问题[J]. 新疆大学学报：哲学社会科学维文版. 2010 年 31 卷 2 期: p.46-53.

阿布都克力木·阿不力孜，哈里旦木·阿布都克里木，吐尔根·依布拉音，等。基于自顶向下算法的维吾尔语句法分析初探[J]。电脑知识与技术. 2010 年 02Z 期: p.1182-1183,1185.

E. Charniak. A maximum-entropy-inspired parser[C]. In Proc. NAACL. Seattle, Washington, USA, 2000: p.1396－1400.

A. Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models[J]. Machine Learning, 34: p.151－175.

A. Culotta and J. Sorensen. Dependency tree kernels for relation extraction[C]. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL). Barcelona, Spain 2004: p.423-es.

S. Billot and B. Lang. 1989. “The structure of shared forests in ambiguous parsing,” in Proceedings of ACL 1989, pages 143－151.

E. Charniak and M. Johnson. 2005. “Coarse-to-fine n-best parsing and maxent discriminative reranking,” in Proceedings of ACL 2005, Ann Arbor, Michigan, pages 173－180.

D. Chiang. 2007. “Hierarchical phrase-based translation,” Computational Linguistics, vol. 33, no. 2, pages 201－228.

M. Galley, M. Hopkins, K. Knight, and D. Marcu. 2004. “What’s in a translation rule?” in Proceedings of HLT/NAACL 2004, Boston, Massachusetts, USA, pages 273－280.

Z. He, Q. Liu, and S. Lin. 2008. “Partial matching strategy for phrase-based statistical machine translation,” in Proceedings of ACL/HLT 2008 (Short Paper), Columbus, Ohio, pages 161－164.

L. Huang. 2008. “Forest reranking: Discriminative parsing with non-local features,” in Proceedings of ACL 2008, Columbus, Ohio, pages 586－594.

L. Huang and D. Chiang. 2005. “Better k-best parsing,” in Proceedings of IWPT 2005, Vancouver, Canada, pages 53－64.

L. Huang and D. Chiang. 2007. “Forest rescoring: Faster decoding with integrated language models”. In Proceedings of ACL, pages 144－151, Prague, Czech Republic, June.

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. 2007. Moses: Open source toolkit for statistical machine translation. In Proc. of ACL 2007 (demonstration session).

Y. Liu, Q. Liu, and S. Lin. 2006. “Tree-to-string alignment template for statistical machine translation,” in Proceedings of COLING/ACL 2006, Sydney, Australia, pages 609－616.

Y. Liu, Y. Huang, Q. Liu, and S. Lin. 2007 “Forest-to-string statistical rules,” in Proceedings of ACL 2007, Prague, Cezch Republic, pages 704－711.

H. Mi, L. Huang, and Q. Liu. 2008. “Forest-based translation,” in Proceedings of ACL/HLT 2008, Columbus, Ohio, USA, pages 192－199.

W. Macherey and F. J. Och. 2007. “An empirical study on computing consensus translations from multiple machine translation systems,” in proceedings of EMNLP-CoNLL2007, Prague, pages 986－995.

F. J. Och. 2003. “Minimum error rate training in statistical machine translation,” in Proceedings of ACL 2003, pages 160－167.

F. J. Och and H. Ney. 2002. “Discriminative training and maximum entropy models for statistical machine translation,” in Proceedings of ACL 2002, pages 295－302.

A. Rosti, S. Matsoukas, and R. Schwartz. 2007. ”Improved Word-Level System Combination for Machine Translation”, in Proceedings of ACL 2007, pages 312－319.

D. Wu. 1997. “stochastic inversion transduction grammars and bilingual parsing of parallel corpora,” Computational Linguistics, vol. 23, pages 377－404, 1997.

D. Xiong, Q. Liu, and S. Lin. 2006. “Maximum entropy based phrase reordering model for statistical machine translation,” in Proceedings of COLING/ACL 2006, Sydney, Australia, pages 521－528.

D. Xiong, S. Li, Q. Liu, and S. Lin. 2005. “Parsing the penn chinese treebank with semantic knowledge,” in Proceedings of IJCNLP 2005, pages 70－81.

X. He, M. Yang, J. Gao, P. Nguyen and R. Moore. 2008. “Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems”, in Proceedings of EMNLP 2008, pages 98－107.