# 多语言文本机器翻译系统

--中科院自动化所CWMT2011评测技术报告

周玉，翟飞飞，张家俊，涂眉，陈钰枫，宗成庆
中国科学院自动化研究所 模式识别国家重点实验室 北京 100190
E-mail: {yzhou, ffzhai, jjzhang, mtu, chenfy, cqzong}@nlpr.ia.ac.cn

**摘 要：** 本文是中科院自动化所参加CWMT2011机器翻译系统评测的技术报告。在本次评测中我们一共参加了九个项目的评测任务，包括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 、哈萨克斯坦语－汉语（哈汉）新闻领域机器翻译、柯尔克孜语－汉语（柯汉）新闻领域机器翻译。文章主要介绍了各评测任务的系统框架、模型、实现方法及评测结果。

**关键词：** 翻译评测；翻译模型；评测结果

# Multi-lingual Machine Translation System

--CASIA Technical Report for CWMT2011 Evaluation

ZHOU Yu, ZHAI Feifei, ZHANG Jiajun,TU Mei, CHEN Yufeng and ZONG Chengqing
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
E-mail: {yzhou, ffzhai, jjzhang, mtu, chenfy, cqzong}@nlpr.ia.ac.cn

***Abstract:*** *This paper describes an overview of CASIA technical report for CWMT2011. We participated in nine tasks: Chinese-to-English Translation for News, English-to-Chinese Translation for News, English-to-Chinese Translation for Scientific and Technological Text, Japanese-to-Chinese Translation for News, Mongolian-to-Chinese Translation for Daily Expressions, Tibetan-to-Chinese Translation for Government Documents, Uyghur-to-Chinese Translation for News, Kazakh-to-Chinese Translation for News and Kirgiz-to-Chinese Translation for News. This paper mainly introduces the overview of our system, the primary modules, the key techniques, and the evaluation results.*

***Keywords:*** *machine translation evaluation, translation model, evaluation results*

## 1 引言

2011年第六届全国机器翻译评测项目（CWMT2011）共包括九个子任务：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 、哈萨克斯坦语—汉语（简称“哈汉”）新闻领域机器翻译、柯尔克孜语—汉语（简称“柯汉”）新闻领域机器翻译评测任务。中科院自动化所(CASIA)参加了所有的评测任务，本文主要介绍自动化所的各个参评系统和相关技术以及在各个翻译任务上的性能表现。

## 2 参评系统描述

在这次机器翻译评测中我们使用了 5 个翻译系统，即：（1）基于最大熵括弧转录文法（MEBTG）的统计机器翻译系统、（2）基于功能（RoleTrans）的统计机器翻译系统（提交报告上使用的是 SynMEBTG 这个名字）、（3）开源基于短语的翻译系统（Moses-BP[1]）、（4）开源基于层次短语的翻译系统（Moses-HP[2]））、（5）词语级系统融合系统（WordComb）。下面我们将对各个系统进行简要地介绍。

### 2.1 MEBTG

MEBTG 是基于最大熵括弧转录文法的统计机器翻译系统，它是我们对文献[Xiong et al., 2006]的一个重实现。该系统的翻译过程类似于一个单语分析过程，该过程只允许一种词汇化规则 $A \rightarrow (x, y)$ 以及两种二元合并规则：顺序合并规则 $A \rightarrow [A^l, A^r]$ 和 逆序合并规则 $A \rightarrow < A^l, A^r >$。在解码时，首先利用词汇化规则将源语言的每个短语 $x$ 翻译成目标语言短语 $y$，并形成一个块 $A$。然后利用合并规则将两个相邻的块合并为一个更大的块，直至源语言句子被一个块完全覆盖，最后选择一个打分最高的目标翻译。

词汇化规则的分值由下面的公式计算：

$$\Pr^l(A) = p(y|x)^{\lambda_1} \cdot p(x|y)^{\lambda_2} \cdot p_{lex}(y|x)^{\lambda_3} \cdot p_{lex}(x|y)^{\lambda_4} \cdot exp(l)^{\lambda_5} \cdot exp(|y|)^{\lambda_6} \cdot P_{LM}^{\lambda_7}(y)$$

其中右边项的前两个是正向与逆向的短语翻译概率，$p_{lex}(y|x)$ 和 $p_{lex}(x|y)$ 是正向与逆向词汇翻译概率，$exp(l)$ 和 $exp(|y|)$ 分别是短语个数惩罚与译文长度惩罚，$P_{LM}(y)$ 是语言模型概率。

合并规则的分值由如下的公式计算：

$$Pr^m(A) = \Omega^{\lambda_8} \cdot P_{LM}^{\lambda_7}(y)$$

其中 $\Omega$ 是短语重排序分值，$\lambda_8$ 为相应特征的权重。与[xiong et al., 2006]相似，短语重排序的分值由基于词汇化（边界词）特征的最大熵模型训练得到。

### 2.2 RoleTrans

基于功能的统计机器翻译系统是集短语翻译、基于最大熵的 BTG 短语重排序模型以及非连续短语翻译为一体的翻译系统。该系统的基本框架可由图-1 来表示：

[1] http://www.statmt.org/moses/
[2] http://www.statmt.org/moses/

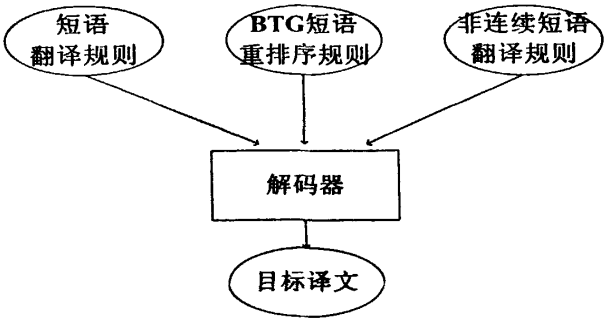


**图-1 RoleTrans 基本框架**

其中，短语翻译规则形如 $A \to x / y$，表示源短语 $x$ 翻译为 $y$；BTG 短语重排序规则形如 $A \to \left[A^l, A^r\right]$ 和 $A \to \left\langle A^l, A^r \right\rangle$，表示两个相邻短语翻译对 $A^l$ 和 $A^r$ 的目标端可进行顺序合并或逆序合并，根据 $A^l$、$A^r$ 源端是否属于句法短语，我们将短语重排序规则分为句法短语重排序规则与非句法短语重排序规则[Zhang et al., 2009]，非句法规则的概率由最大熵分类器依据短语边界词为特征计算得出，而句法规则的概率由最大熵分类器依据短语边界词以及句法标记为特征计算获得；非连续短语翻译规则包括源端非连续 $A \to f_L\ X\ f_R / e_L\ X \mid f_L\ X\ f_R / X\ e_R$、目标端非连续 $A \to f_L\ X / e_L\ X\ e_R \mid X\ f_R / e_L\ X\ e_R$ 和两端非连续 $A \to f_L\ X\ f_R / e_L\ X\ e_R$，我们按照层次短语翻译模型中的规则抽取方法抽取非连续短语翻译规则，并限制非终结符的个数为 1。从而，基于功能的翻译模型可由下式表示：

$$P(A) = P(PhrTrans)^{\lambda_1} \cdot P_{syn}(BTG)^{\lambda_2} \cdot P_{non-syn}(BTG)^{\lambda_3} \cdot P(SrcDis)^{\lambda_4} \cdot P(TgtDis)^{\lambda_5} \cdot P(BothDis)^{\lambda_6}$$

$P(PhrTrans)$ 包括双向短语翻译概率、双向词汇化翻译概率，当然，我们在翻译的过程中也采用了一些其他常用的特征：语言模型、翻译规则的个数以及译文长度，另外在基于功能的翻译系统中，除了传统的前向语言模型特征，我们还采用了后向语言模型特征[Xiong et al., 2011]，即对于一个目标语言串 $e = e_1 e_2 \cdots e_{n-1} e_n$，传统的前向 m-gram 语言模型计算概率为：

$$P_{forward}(e) = P_{forward}(e_1 e_2 \cdots e_{n-1} e_n) = P(e_1) P(e_2 \mid e_1) \cdots P(e_n \mid e_{n-m+1} \cdots e_{n-1})$$

而后向 m-gram 语言模型概率由下式计算：

$$P_{backward}(e) = P_{backward}(e_1 e_2 \cdots e_{n-1} e_n)$$
$$= P_{forward}(e_n e_{n-1} \cdots e_2 e_1)$$
$$= P(e_n) P(e_n | e_{n-1}) \cdots P(e_1 | e_m \ldots e_2)$$

所有特征的权重利用最小错误率训练（MERT）获得。

### 2.3 Moses-BP

Moses-BP 是当前最流行也是最稳定的基于短语的统计机器翻译系统[Koehn et al., 2007]。该系统利用 log-linear 模型将多个翻译特征融合，它采用了 MSD(Monotone, Swap, Discontinuous)词汇化的调序模型。我们在评测时利用的版本是 Version 2011-01-12。

### 2.4 Moses-HP

Moses-HP 是一个开源的基于层次短语的统计机器翻译系统。所谓层次短语模型就是基于上下文无关文法，不使用任何语言学知识的句法模型，该层次短语模型的翻译效果要好于基于普通短语的翻译模型[Chiang, 2005] [Chiang, 2007]。我们在评测时利用的版本是 Version 2011-01-12。

### 2.5 WordComb

我们的词级别系统融合方法主要用了三种，一种是单纯的 WER[Snover et al., 2006]，一种是单纯 IHMM[He et al., 2008]，第三种是对这两种融合后产生的新假设和各个单系统生成的假设利用 IHMM 进行再次融合。

### 2.6 系统性能

在这次评测中，机器翻译评测采用的计算机配置如表 1 所示，我们在这次评测中主要用了两台服务器：

**表 1：机器硬件配置与操作系统**

| 服务器 | CPU | 内存 | 操作系统 |
|---|---|---|---|
| 服务器 1 | Xeon(R) L5508 2.5G 2.0GHz ×2CPU | 64G | Ubuntu-server 8.04 |
| 服务器 2 | Intel Xeon E5620 2.40GHz ×2CPU | 128G | Red Hat Enterprise Linux Server release 5.5 |

## 3 实验

### 3.1 数据使用

- **训练语料**：我们此次评测使用的语料完全是主办方提供的所有训练数据。另外，英文的语言模型训练也用到了主办方提供的路透社语料；中文的语言模型训练用到了搜狗语料。
- **开发集**：我们在这次评测中使用的开发集也是主办方发布的所有的开发集作为我们调参的开发集。

## 3.2 关键知识和模型的处理与获取方法

- 数据的预处理：对中文数据进行的处理有：中文的分词和全角变半角；对英文数据进行的处理为：大写转小写和标点符号的分离处理。其中中文的分词是利用我们自主开发的分词工具 Urheen2.2[3]。对各个小语种也进行了标点符号的分离处理。
- 词对齐和翻译模型的获取：词对齐工具采用两种开源词对齐工具：GIZA++[Och et al., 2000][Koehn et al., 2003]和 Berkeley[4]对齐工具，其中 GIZA++的词对齐我们采用 grow-diag-final-and 的扩展方式来得到多对多的词对齐结果。我们在评测中分别利用这两个词对齐工具生成两个词对齐文件，然后直接合并这两个词对齐文件，最后在合并的词对齐文件上获取翻译模型。
- 语言模型的获取：语言模型训练工具采用 SRILM 工具[Stolcke, 2002]，并且我们评估不同的语言模型规模对翻译性能的影响，最终选择最好的语言模型组合。
- 短语重排序模型训练工具：RoleTrans 中用到的句法分析采用了 Stanford Parser[Klein et al., 2003]，MEBTG 与 RoleTrans 中使用的短语重排序模型训练工具采用的是[Zhang, 2004]的最大熵训练工具。
- 多语种时间数字的识别和翻译：时间数字识别和翻译主要是利用规则方法。考虑到时间和数字信息的多样性，我们将各个语种（这里我们一共做了 6 个语种，包括：汉语、英语、日语、藏语、蒙语和维语）的时间数字细化为六类来进行处理，分别如下所示：1、数量（Number）；2、序数词（Ordinal）；3、号码（Figure）；4、月份（Month）；5、日期（Date）；6、星期（Week）。其中目标语言的翻译方式主要选择与开发集参考答案相同的形式。该部分工作主要是在[翟飞飞等，2009]的工作基础上进行开发、完善和调整，并将程序与规则严格分离开来，从而使之能够迅速进行扩展和移植，从而成功实现了多语种的时间数字识别及翻译。
- 命名实体识别和翻译：对于命名实体，针对中文，我们采用[Wu et al., 2005]开发的多知识源融合的汉语实体识别系统进行汉语命名实体的识别；针对英文，我们采用公开的 Mallet[5]软件包中的基于条件随机场模型（Conditional Random Fields，CRF）的英语实体标注工具进行英语命名实体的识别标注。在汉英实体翻译中，我们对人名和地名采用字典音译方式进行翻译，而机构名的翻译则利用基于语块的层次翻译模型[Chen et al., 2008]。针对英汉实体翻译，我们对各类实体都采用音译方式进行翻译。
- 数据的后处理：对于汉语没有采用任何后处理，直接提交输出文件，对于英文主要是处理字母大小写和标点符号的合并。

## 3.3 实验设置

在本次评测中，我们对翻译模型和语言模型的获取中的一些重要设置和选择如下所示。

- 短语长度的设置

本次评测系统中对于基于短语的翻译模型，我们短语长度是这么设定的：1）对于

[3] http://www.openpr.org.cn/index.php/NLP-Toolkit-For-Natural-Language-Processing/68-Urheen-A-Chinese/English-Lexical-Analysis-Toolkit/View-details.html
[4] http://code.google.com/p/berkeleyaligner/downloads/list
[5] http://mallet.cs.umass.edu/index.php/Main_Page

汉英新闻和英汉新闻，设置的短语最大长度为 7；2）对于其他的翻译评测任务，设置的短语最大长度为 10。对于基于层次短语的翻译模型，我们的短语最大长度设置为 5。

- 词对齐的选择

  这次翻译模型的训练我们采用了不同词对齐（包括 giza++--gdfa 和 berkeley）的融合来训练得到翻译模型。

- 语言模型的选择

  我们设置并训练了三种目标语言模型：1）以发布的双语训练语料的所有目标语言作为训练集得到 5 元的语言模型，我们称之为 else.lm5，2）以大规模单语语料（中文搜狗语料或英文路透社语料）作为训练集得 5 元的语言模型 big.lm5，3）以训练语料的目标语言对大规模单语语料进行过滤，用过滤后的部分作为训练集得到的语言模型 mid.lm5。我们在试验中尝试了不同语言模型的结合对开发集的影响。最终对各个参评系统采用了如下的结合方式：

表 2：最终不同语言模型的结合方式

| 参评项目 | 最终采用的语言模型 |
|---|---|
| 汉英新闻 | else.lm5+big.lm5 |
| 英汉新闻 | else.lm5+big.lm5 |
| 英汉科技 | else.lm5+big.lm5 |
| 日汉新闻* | else.lm5+mid.lm5 & else.lm5+big.lm5 |
| 蒙汉日常用语* | else.lm5+mid.lm5 & else.lm5+big.lm5 |
| 藏汉政府文献* | else.lm5+mid.lm5 & else.lm5+big.lm5 |
| 维汉新闻* | else.lm5+mid.lm5 & else.lm5+big.lm5 |
| 哈汉新闻* | else.lm5+mid.lm5 & else.lm5+big.lm5 |
| 柯汉新闻* | else.lm5+mid.lm5 & else.lm5+big.lm5 |

其中加“*”的用了两种语言模型的结合方式，主要是考虑到我们在不同的翻译引擎中会利用不同的语言模型的结合方式。

- 短语重排序模型训练语料的选择

  系统MEBTG与RoleTrans所用的词汇化短语重排序模型由最终的所有训练语料经最大熵工具训练得到；用于汉英和英汉新闻的 RoleTrans 系统所用的句法短语重排序模型在所有训练语料的基础上去除 ict-web 和 neu corpus 经最大熵工具训练而得。用于科技的 RoleTrans 系统所用的句法短语重排序模型采用整个科技训练语料经最大熵工具训练而得。

## 3.4 实验结果与分析

### 3.4.1 汉英新闻领域机器翻译评测结果与分析

我们汉英新闻的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 5,746,951 个双语句对，开发集包含 2,270 个源语言句子和 4*2,270 个参考答案。表 3 列出了我们参加汉英新闻领域机器翻译系统在开发集和测试集上的性能表现。

表 3：汉英新闻领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集（BLEU-SBP，忽略大小写） | 测试集（BLEU-SBP，大小写敏感） |
|---|---|---|---|
| **RoleTrans（primary-systema）** | **else.lm5+big.lm5** | **0.282726** | **0.2283** |
| **Moses-HP (contrast-systemc)** | **else.lm5+big.lm5** | **0.276139** | **0.2016&** |
| MEBTG | else.lm5+big.lm5 | 0.264644 | 0.2103 |
| Moses-BP1 | else.lm5+big.lm5 | 0.268082 | 0.2162 |
| Moses-BP2 | else.lm5 | 0.268239 | 0.2131 |
| IHMM | big.lm5 | 0.2521* | 0.2223 |
| WER | big.lm5 | 0.2495* | 0.2182 |
| **IHMM-WER* (contrast-systemb)** | **big.lm5** | **0.2533*** | **0.2252** |

所有表 3 中加黑的地方都表示是提交的结果。其中带&表示结果奇怪，后来查看才知道参数搞错了，所以导致得分竟然还没有 BP 的高。上面的*表示所用的开发集不一致，所以打分仅供参考，这个融合参数是通过下面表 4 训练得到的见表 4。这个融合系统是从汉英新闻原始的开发集中挑选了其中的 966 句作为系统融合的开发集：后面三个系统为融合的结果，其中 IHMM 和 WER 是分别把前面四个系统进行融合得到的结果，而 IHMM-WER 是把前面 6 个系统的结果再次采用 IHMM 的方式进行融合的结果。

表 4：融合系统与单系统性能得分在开发集上的对比

| Moses-HP | RoleTrans | Moses-BP1 | Moses-BP2 | IHMM | WER | IHMM-WER |
|---|---|---|---|---|---|---|
| 0.2363 | **0.2414** | 0.2271 | 0.2263 | 0.2521 | 0.2495 | **0.2533** |

### 3.4.2 英汉新闻领域机器翻译评测结果与分析

我们英汉新闻的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 5,746,951 个双语句对，开发集包含 2,481 个源语言句子和 4*2,481 个参考答案。表 5 列出了我们参加英汉新闻领域机器翻译系统在开发集和测试集上的性能表现。

表 5：英汉新闻领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集（BLEU-SB | 测试集（BLEU-SBP,字 | 测试集（BLEU-SBP, |
|---|---|---|---|---|

| | | P，词 4-gram） | 5-gram）<br>（Progress） | 字 5-gram）<br>（Current） |
|---|---|---|---|---|
| **RoleTrans<br>（primary-systema）** | **else.lm5+big.lm5** | **0.307266** | **0.3522** | **0.3284** |
| Moses-HP | else.lm5+big.lm5 | 0.296592 | – | – |
| MEBTG | else.lm5+big.lm5 | 0.292321 | – | – |
| Moses-BP1 | else.lm5+big.lm5 | 0.300011 | – | – |
| Moses-BP2 | else.lm5 | 0.288688 | – | – |

### 3.4.3 英汉科技领域机器翻译评测结果与分析

我们英汉科技的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 900,765 个双语句对，开发集包含 1,116 个源语言句子和 4*1,116 个参考答案。表 6 列出了我们参加英汉科技领域机器翻译系统在开发集和测试集上的性能表现。

表 6：英汉科技领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集（BLEU-SBP，以词的 4-gram 打分） | 测试集（BLEU-SBP，以字的 5-gram 打分） |
|---|---|---|---|
| **RoleTrans<br>（primary-systema）** | **else.lm5+big.lm5** | **0.344728** | **0.3722** |
| Moses-HP | else.lm5+big.lm5 | 0.3417890 | – |
| MEBTG | else.lm5+big.lm5 | 0.337509 | – |
| Moses-BP1 | else.lm5+big.lm5 | 0.335377 | – |
| Moses-BP2 | else.lm5 | 0.331819 | – |
| IHMM | big.lm5 | 0.3465 | – |

我们发现提交的英汉科技结果表现一般，分析其主要原因可能是因为科技的句子很长，句法分析质量很差，所以获取的调序模型不好。

这里需要说明的是：因为时间紧张，而且语料规模庞大，加载内存较大，英汉新闻和英汉科技没做融合系统，最终只提交了单系统最好的 RoleTrans 作为我们的主系统。

### 3.4.4 日汉新闻机器翻译评测结果与分析

我们日汉新闻的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 282,483 个双语句对，开发集包含 500 个源语言句子和 4*500 个参考答案。表 7 列出了我们参加日汉科技领域机器翻译系统在开发集和测试集上的性能表现。其中日语分词我们采用的是开源软件 Mecab[6]分词工具。

表 7：日汉新闻领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集（BLEU-SBP, 以词的 4-gram 打分） | 测试集（BLEU-SBP, 以字的 5-gram 打分） |
|---|---|---|---|
| **Moses-HP (contrast-systemc)** | **else.lm5+big.lm5** | **0.355022** | **0.4947** |
| Moses-HP | else.lm5+mid.lm5 | 0.335697 | 0.4535 |
| MEBTG | else.lm5+mid.lm5 | 0.30785 | 0.4348 |
| Moses-BP1 | else.lm5+big.lm5 | 0.335505 | 0.4718 |
| Moses-BP2 | else.lm5+mid.lm5 | 0.326798 | 0.4303 |
| IHMM | big.lm5 | 0.3627 | 0.4925 |
| **WER (contrast-systemb)** | **big.lm5** | **0.3633** | **0.4958** |
| **WER- IHMM (primary-systema)** | **big.lm5** | **0.3668** | **0.4957** |

### 3.4.5 蒙汉日常用语机器翻译评测结果与分析

我们蒙汉日常用语的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 67,288 个双语句对，开发集包含 1,000 个源语言句子和 4*1,000 个参考答案。表 8 列出了我们参加蒙汉日常用语领域机器翻译系统在开发集和测试集上的性能表现。

表 8：蒙汉日常用语领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集（BLEU-SBP, 以词的 4-gram 打分） | 测试集（BLEU-SBP, 以字的 5-gram 打分） |
|---|---|---|---|
| **Moses-HP (contrast-systemc)** | **else.lm5+big.lm5** | **0.205476** | **0.2002** |

[6] http://mecab.sourceforge.net/

| | | | |
|---|---|---|---|
| Moses-HP | else.lm5+mid.lm5 | 0.2029 | 0.1774 |
| MEBTG | else.lm5+mid.lm5 | 0.191505 | 0.1605 |
| **Moses-BP1 (contrast-systemd)** | **else.lm5+big.lm5** | **0.199823** | **0.1906** |
| Moses-BP2 | else.lm5+mid.lm5 | 0.198202 | 0.1844 |
| **IHMM (contrast-systemb)** | **big.lm5** | **0.2251** | **0.1932** |
| **WER_IHMM (primary-systema)** | **big.lm5** | **0.2275** | **0.1851** |

### 3.4.6 藏汉政府文献机器翻译评测结果与分析

我们藏汉政府文献的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 101,629 个双语句对，开发集包含 650 个源语言句子和 4*650 个参考答案。表 9 列出了我们参加藏汉政府文献领域机器翻译系统在开发集和测试集上的性能表现。

**表 9：藏汉政府文献领域机器翻译评测结果**

| 参评系统 | 语言模型 | 开发集（BLEU-SBP, 以词的 4-gram 打分） | 测试集（BLEU-SBP, 以字的 5-gram 打分） |
|---|---|---|---|
| **Moses-HP (primary-systema)** | **else.lm5+big.lm5** | **0.521852** | **0.58** |
| Moses-HP | else.lm5+mid.lm5 | 0.4954 | 0.5598 |
| **MEBTG (contrast-systemc)** | **else.lm5+mid.lm5** | **0.485657** | **0.5471** |
| **Moses-BP1 (contrast-systemb)** | **else.lm5+big.lm5** | **0.500306** | **0.5337** |
| Moses-BP2 | else.lm5+mid.lm5 | 0.482979 | 0.5291 |
| **IHMM (contrast-systemd)** | **big.lm5** | **0.5429** | **0.5843** |

### 3.4.7 维汉新闻领域机器翻译评测结果与分析

我们维汉新闻领域的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 50,000 个双语句对，开发集包含 700 个源语言句子和 4*700 个参考答案。表 10 列出了我们参加维汉新闻领域机器翻译系统在开发集和测试集上的性能表现。

表 10：维汉新闻领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集（BLEU-SBP, 以词的 4-gram 打分） | 测试集（BLEU-SBP, 以字的 5-gram 打分） |
|---|---|---|---|
| **Moses-HP (primary-systema)** | **else.lm5+big.lm5** | **0.430743** | **0.4761** |
| **Moses-HP (contrast-systemd)** | **else.lm5+mid.lm5** | **0.418251** | **0.4714** |
| **MEBTG** | **else.lm5+mid.lm5** | 0.407152 | 0.4599 |
| **Moses-BP1 (contrast-systemb)** | **else.lm5+big.lm5** | **0.413311** | **0.4636** |
| Moses-BP2 | else.lm5+mid.lm5 | 0.401184 | 0.4545 |
| **IHMM (contrast-systemc)** | **big.lm5** | **0.4405** | **0.4754** |

### 3.4.8 哈汉新闻领域机器翻译评测结果与分析

我们哈汉新闻领域的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 50,000 个双语句对，开发集包含 700 个源语言句子和 4*700 个参考答案。表 11 列出了我们参加哈汉新闻领域机器翻译系统在开发集和测试集上的性能表现。

表 11：哈汉新闻领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集（BLEU-SBP, 以词的 4-gram 打分） | 测试集（BLEU-SBP, 以字的 5-gram 打分） |
|---|---|---|---|
| **Moses-HP (primary-systema)** | **else.lm5+big.lm5** | **0.318463** | **0.398** |
| Moses-HP | else.lm5+mid.lm5 | 0.317747 | – |
| **Moses-BP1 (contrast-systemb)** | **else.lm5+big.lm5** | **0.308735** | **0.3906** |

此处没有提交 Moses-HP(else.lm5+mid.lm5)的结果，是考虑到它跟 Moses-HP (else.lm5+big.lm5)在开发集上的性能差不多，就没有翻译。

### 3.4.9 柯汉新闻领域机器翻译评测结果与分析

我们柯汉新闻领域的训练集和开发集所有语料都来自于主办方发布的语料，经预处理后我们的训练集包含 50,000 个双语句对，开发集包含 700 个源语言句子和 4*700 个参考答案。表 12 列出了我们参加柯汉新闻领域机器翻译系统在开发集和测试集上的性能表现。

表 12：柯汉新闻领域机器翻译评测结果

| 参评系统 | 语言模型 | 开发集 BLEU-SBP, 以词的 4-gram 打分） | 测试集（BLEU-SBP, 以字的 5-gram 打分） |
|---|---|---|---|
| **Moses-HP (primary-systema)** | **else.lm5+big.lm5** | **0.388467** | **0.4346** |
| Moses-HP | else.lm5+mid.lm5 | 0.381098 | 0.4293 |
| **MEBTG (contrast-systemc)** | **else.lm5+mid.lm5** | **0.362873** | **0.4177** |
| **Moses-BP1 (contrast-systemd)** | **else.lm5+big.lm5** | **0.377783** | **0.4229** |
| Moses-BP2 | else.lm5+mid.lm5 | 0.368006 | 0.4134 |
| **IHMM (contrast-systemb)** | **big.lm5** | **0.3982** | **0.4391** |

这里需要解释一下的是：其中藏汉、维汉和柯汉的之所以没有将融合系统 IHMM 作为主系统，是因为我们融合的时候是在开发集上没有-drop-unknown 的情况下调参的，而测试集我们都 drop-unknown 了，所以怕参数不稳定，就没有选择其作为主系统，而作为对比系统提交。

## 4 总结

本文主要介绍了中科院自动化所参加 CWMT2011 评测的情况。在九个子项的评测中，我们都取得了较好的成绩，特别是汉英新闻和英汉新闻翻译任务中，我们发现句法知识能够帮助基于短语的系统显著地改善翻译性能，同时我们提出的系统 RoleTrans 有效地集成了各个翻译功能，并且可以适用于大规模的训练语料，因为它不需要对所有训练语料的中文部分进行句法分析，而且不增加解码的复杂度；其次，对于小语种的翻译，采用的词对齐合并技术、两种语言模型结合的

技术，尤其是开发的多语种的时间数字识别和翻译模块，评测结果证明都非常有效。然而，我们还有很大的提升空间，比如系统融合对于大规模语料还不够稳定，翻译引擎速度还有待提高，加载内存较大等。我们需要向国内外同行学习，改善我们现有的系统。

## 5 致谢

在这次评测中，模式识别国家重点实验室的很多同学付出了许多艰辛的劳动，给予了很多工作上和精神上的支持，在此对汪昆、庄涛、王志国、李小青四位同学表示衷心的感谢！并特别感谢西北民族大学教育部重点实验室-中国民族语言文字信息技术实验室的于洪志教授、江涛老师、加羊吉、格根塔娜和努尔比亚吐拉甫几位同学给予的大力帮助和支持。

## 参 考 文 献

[Chen et al., 2008] Yufeng Chen and Chengqing Zong. 2008. A Structural-Based Model for Chinese Organization Name Translation. *ACM Transactions on Asian Language Information Processing (ACM TALIP)*, 7(1): 1-30.

[Chiang, 2005] Chiang, David. 2005. A hierarchical phrasebased model for statistical machine translation. *In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263 - 270, Ann Arbor, MI.

[Chiang, 2007] Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201 - 228.

[He et al., 2008] Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, Robert Moore. 2008. Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems. In *Proceedings of the 2008 Conference on EMNLP*, Honolulu, pp. 98-107.

[Och et al., 2000] Franz Josef Och, Hermann Ney. 2000. "Improved Statistical Alignment Models". *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447, Hongkong, China, October 2000.

[Koehn et al., 2003] Philipp Koehn, Franz Josef Och, Daniel Marcu. 2003. "Statistical Phrase-Based Translation", In *Proc HLT-NAACL*, 2003

[Koehn et al., 2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

[Snover et al., 2006] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, & J. Makhoul. 2006. "A study of translation edit rate with targeted human annotation," In *Proc. Assoc. for Machine Trans. in the American*, 2006.

[Stolcke, 2002] A. Stolcke. 2002. SRILM -- An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, Denver.

[Klein et al., 2003]Klein D. and C.D. Manning. 2003. Accurate Unlexicalized Parsing. *In Proceedings of ACL.*

[Xiong et al., 2006] Xiong, D.Y., Q. Liu and S.X. Lin. 2006. Maximum Entropy based Phrase Reordering Model for Statistical Machine Translation. *In Proceedings of ACL-COLING* 2006.

[Xiong et al., 2011] Deyi Xiong, Min Zhang, Haizhou Li. 2011. Enhancing Language Models in Statistical Machine Translation with Backward N-grams and Mutual Information Triggers. In *Proceedings of ACL 2011.*

[Wu et al., 2005] Youzheng Wu, Jun Zhao and Bo Xu. 2005. Chinese Named Entity Recognition Model Based on Multiple Features. In *Proceedings of HLT/EMNLP 2005*, pages 427-434. October 6-8, Vancouver, B.C., Canada.

[Zhang, 2004] L. Zhang, 2004. Maximum Entropy Modeling Toolkit for Python and C++. *http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.*

[Zhang et al., 2009] Jiajun Zhang and Chengqing Zong. 2009. A Framework for Effectively Integrating Hard and Soft Syntactic Rules into Phrase-Based Translation. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23).* 3-5 December 2009, Hong Kong. Pages 579-588

[翟飞飞等，2009]翟飞飞，夏睿，周玉，宗成庆. 2009. 汉英双向时间和数字命名实体的识别与翻译系统，第五届全国机器翻译研讨会（CWMT2009）论文集，2009 年 10 月 16—17 日，南京。第 172—179 页。