# 中科院自动化所 CWMT2009 评测技术报告

张家俊，李茂西，周玉，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: {jjzhang, mxli, yzhou, chenfy, cqzong}@nlpr.ia.ac.cn

**摘 要**：本文主要介绍了中科院自动化所参加 CWMT2009 研讨会的技术报告，我们一共参加了四个项目的评测任务，包括汉英新闻单一系统评测任务、英汉新闻、英汉科技的机器翻译评测任务以及汉英新闻系统融合任务。文章主要介绍了我们参加各个评测任务的系统的主要框架、模型、实现细节及其评测结果。

**关键词**：基于短语翻译模型；MEBTG 翻译模型；句法增强的 MEBTG 翻译模型；层次短语；系统融合

# CASIA Technical Report for CWMT2009 Evaluation

ZHANG Jiajun, LI Maoxi, ZHOU Yu, CHEN Yufeng and ZONG Chengqing
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
E-mail: {jjzhang, mxli, yzhou, chenyf, cqzong}@nlpr.ia.ac.cn

***Abstract***: *This paper describes an overview of CASIA technical report for CWMT2009. We participated in four tasks: Single system translation on Chinese-to-English News, Machine translation on English-to-Chinese News, Machine translation on English-to-Chinese Science and Technology and System Combination on Chinese-to-English News. This paper mainly introduces the overview of our system, the primary modules, the key techniques, and the evaluation results.*

***Keywords***: *phrase-based translation model; MEBTG-based translation model; syntax augmented MEBTG-based translation model; hierarchical phrase; system combination*

## 1 引言

2009 年第五届全国机器翻译评测项目（CWMT2009）一共包括五个子任务：汉英新闻领域单一系统翻译任务、英汉新闻领域机器翻译任务、英汉科技领域机器翻译任务、汉英新闻领域系统融合任务以及汉蒙日常用语机器翻译任务。自动化所(CASIA)作为参评单位之一参加了除汉蒙翻译任务以外的其它四个评测任务，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 6 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）句法增强的基于最大熵括弧转录文法（SynMEBTG）的统计机器翻译系统、（3）开源基于短语的翻译系统（Moses[1]）、（4）开源基于层次短语的翻译系统（Joshua[2]）、（5）词语级系统融合系统（WordComb）以及（6）句子级系统融合系统（SenComb）。

[1] http://www.statmt.org/moses/
[2] http://sourceforge.net/projects/joshua/

对于每个翻译任务，我们分别使用如下的系统：

a) 汉英新闻：MEBTG、SynMEBTG 和 Moses。

b) 英汉新闻：Moses。

c) 英汉科技：Moses、Joshua 和 SenComb.

d) 系统融合：SenComb 和 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 $A \to (x, y)$ 以及两种二元合并规则：顺序合并规则 $A \to [A^l, A^r]$ 和 逆序合并规则 $A \to < A^l, A^r >$。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：

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

其中 $\Omega$ 是调序分值，$\lambda_8$ 为相应特征的权重。与[xiong et al., 2006]相似，调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 SynMEBTG

SynMEBTG 系统是 MEBTG 系统的句法增强版本。由于 MEBTG 的核心思想就是将顺序合并和逆序合并看成一个最大熵的二元分类问题。因此，分类所采用的特征将成为决定系统性能的关键因素。MEBTG 系统只采用了词汇化的特征，分类的正确率不是很高。SynMEBTG 系统就是设法在不降低实际解码速度的情形下，将源语言的句法信息高效地融入调序模型。

SynMEBTG 的基本思想就是：如果被合并的两个短语都是句法短语，我们就采用句法调序信息，否则我们采用 MEBTG 的词汇化调序信息。不同于在解码过程中计算句法调序信息，我们将句法调序信息的计算作为翻译前的预处理模块。类似于[Li et al., 2007]，我们从一棵句法树上获得句法调序信息。[Li et al, 2007]处理含有两个或三个孩子节点的子树，然后决定孩子节点间是否需要调序，最终得到调序后的源语言句子。我们的方法如下：如果一个节点有两个孩子节点，我们即可以构造一个规则决定他们是否需要交换顺序；如果一个节点含有三个以上的孩子节点，我们首先判断孩子节点中是否有中心节点(VP 或者 NP)，有的话，我们便设计一个规则决定位于中心节点前的修饰节点是否需要调至中心节点后。综合而得我们设计的规则如下：

$$P: N^l <> N^r \Rightarrow \begin{Bmatrix} N^l <> N^r & straight \\ <> N^r N^l & inverted \end{Bmatrix}$$

如果节点 P 只有两个孩子节点，则<>为空；否则<>便为修饰节点 $N^l$ 与中心节点 $N^r$ 之间的其他节点。由于 $N^l <> N^r$ 中 $N^l$ 调至 $N^r$ 之后与 $N^l$、$<> N^r$ 交换顺序是等价的，所以我们采用后者的表示方法，从而以上规则就表示两个连续短语之间的顺序关系。同时，我们假设 $<> N^r$ 也为句法短语，从而我们的规则也就变成两个连续句法短语之间的顺序关系。最后我们以训练语料的中文句法树，双语对齐为输入，抽取调序实例，并利用最大熵训练得到调序模型，特征包括边界词与词性标记、上下文词与词性标记以及各节点及父节点的句法标记。

同时，与以前所有方法将句法短语、非句法短语融入一个调序模型不同，我们认为句法与非句法调序依赖于不同层次的特征，在翻译中扮演着不同重要程度的角色。从而，我们将调序模型细分为句法调序模型与非句法调序模型，因此 BTG 翻译模型中合并规则的打分可以由下式计算：

$$Prm(\Omega\Omega(\ )_N^{\lambda_8 \cdot I_N(A)} \cdot \ _S^{\lambda_9 \cdot I_S(A)} \cdot P_{LM}^{\lambda_7} \ y$$

其中，$\Omega_S$ 和 $\Omega_N$ 分别是句法调序分值和非句法调序分值，$I_S(A)$ 和 $I_N(A)$ 为指示函数，表明当合并句法短语时，使用 $\Omega_S$，否则使用 $\Omega_N$。

为了增强句法调序的作用，我们也增加了一个二元特征 $R_S$ 对句法调序进行奖励。

对于一个待翻译的句子，我们首先从其句法树中得到句法调序规则集合，在其后的 CKY 解码过程中，如果待合并的两个连续短语有句法调序规则与之匹配，则采用句法调序，否则采用词汇化调序，最后选取概率最大的翻译假设。

### 2.3 Moses

Moses 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。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。我们在评测时利用了当时的最新版本 Version 2009-04-13。

### 2.4 Joshua

Joshua 是一个开源的基于层次短语的统计机器翻译系统。该系统实现了上下文无关文法所需的所有算法，并采用了基于后缀数组的文法规则抽取算法。它使得层次短语模型能够工作在大规模训练语料上。在评测中，我们使用了当时的最新版本 version 1.1。

### 2.5 WordComb

我们的词级别系统融合方法是参考文献[Rosti et al., 2007a, Rosti et al., 2007b]实现的，但是我们用 WER[Bangalore et al., 2001]词对齐代替 TER[Snover et al., 2006]词对齐构建混淆网络。

### 2.6 SenComb

利用词级别的系统融合产生的结果，从原始的系统 1-Best 中挑出一个与词级别融合结果最相似的一个，该方法主要参考文献[Sim et al., 2007]，主要的不同在于我们利用很多特征来进行词级别的系统融合，而文献[Sim et al., 2007]只是使用了词的后验概率来进行词级别的系统融合。在计算两个翻译假设的距离时，我们分别使用了平滑的句级别 BLEU 打分方法和 Meteor 打分方法。

### 2.7 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：

表 1：机器硬件配置与操作系统

| CPU | 内存 | 操作系统 |
|---|---|---|
| Intel Xeon E5420 2.5G | 32G | Ubuntu-server 8.04 |

# 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训练数据，规模合计为3,783,921 句对，其中包括科技语料 920,985 句对。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：在汉英单系统评测中，我们使用了 2003 到 2005 以及 2008 的测试集的并集共3,276 句中文及 4 个相应的参考译文；英汉新闻机器翻译我们使用了 2003 到 2005 以及 2008 的测试集的并集共 3,481 句英文及 4 个相应的参考译文；英汉科技机器翻译我们使用了 2008 年的测试语料 1,008 句英文及其 4 个相应的参考译文；系统融合的开发集我们使用的是 SSMT2007 的测试语料，共 1,002 句。

## 3.2 数据处理方法及工具

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开源工具 ICTCLAS3.0[3]。

词对齐工具采用 GIZA++(Och 2003)。当利用 GIZA++进行双语语料双向词对齐后，我们评估多种启发式合并规则得到不同的词对齐结果，最后选择效果最好的词对齐方式。

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。

SynMEBTG 中用到的句法分析采用了 Stanford Parser(Klein et al., 2003)，MEBTG 与 SynMEBTG 中使用的调序模型训练工具采用的是(Zhang 2004)的最大熵训练工具。

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）。其中目标语言选择与开发集参考答案相同的形式。

对于命名实体，针对中文，我们采用[Wu，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[4]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 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。

[3] http://www.nlp.org.cn
[4] http://mallet.cs.umass.edu/index.php/Main_Page

数据的后处理：对于汉语的后处理主要是合并空格，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

由于本次评测所用系统除 Joshua 外都是基于短语的模型，因此需要对翻译中允许的源短语最大长度进行设置。而且每个系统都要用到词对齐结果、目标语言模型，因此我们利用开源工具 Moses 作为解码器，科技领域 CWMT08 的测试集作为开发集，BLEU-4 作为评价指标来对每个选择进行评估，最终选择效果最好的组合。在实验过程中，每组对比实验只是所考察特征的值不同，其余设置一样。

- 短语长度选择：我们仅考察允许的短语最大长度为 7 或 10，并从中选择较好者，实验比较如表 2 所示，我们发现最大短语长度为 10 时效果更好。

表 2：最大短语长度选择实验

| 短语最大长度 | Bleu-4 |
|---|---|
| maxphlen=7 | 0.255987 |
| **maxphlen=10** | **0.258018** |

- 词对齐中启发式规则的选择：我们考察两种启发式规则 grow-diag-final 和 grow-diag-final-and，以及这两种启发式规则的合并（将语料、词对齐都合并，从合并的词对齐中抽取短语，计算短语概率），表 3 为我们的实验结果，该结果表明，两者结合产生的效果最好。

表 3：词对齐启发式规则选择实验

| 不同词对齐结合方式 | Bleu-4 |
|---|---|
| grow-diag-final-and | 0.258018 |
| grow-diag-final | 0.247290 |
| **gdfa+gdf** | **0.258851** |

- 语言模型的选择：我们设置并训练了三种目标语言模型：1）以训练语料的目标语言部分作为训练集得语言模型 Origin.lm4，2）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语言模型 Mid.lm5。其中“4、5”表示 4 元或 5 元语言模型。实验结果如表 4 所示，该结果显示 Mid.lm5 与 Big.lm5 结合最好。

表 4：不同语言模型实验

| 不同语言模型 | Bleu-4 |
|---|---|
| Origin.lm4 | 0.258018 |
| Big.lm5 | 0.247963 |
| Big.lm5+Origin.lm4 | 0.267523 |
| **Big.lm5+Mid.lm5** | **0.269800** |

- 训练语料的选择：针对训练语料中句子对齐质量的参差不齐，我们设置了三种语料（针

对科技)：1）科技领域训练语料，2）科技领域与新闻领域的合并语料，3）利用 GIZA++ 训练得到的词典对新闻领域训练语料进行过滤得到的过滤语料。实验结果如表 5 所示，该结果说明科技领域本身的语料加上新闻领域过滤后语料得到的效果最好，但由于效果并不明显，我们最终选择的是新闻与科技语料的组合。

表 5：语料的选择实验

| 不同语料 | Bleu-4 |
|---|---|
| 科技语料 | 0.258018 |
| 新闻+科技语料 | 0.273732 |
| **过滤后新闻+科技语料** | **0.275532** |

从以上四组对比实验中，我们设定最后的实验配置如下：短语的最大长度选择 10，词对齐选择 grow-diag-final 与 grow-diag-final-and 的结合，语言模型选择过滤后五元语言模型 Mid.lm5 与大规模单语五元语言模型 Big.lm5，训练语料选择科技与新闻语料的组合。Joshua 只使用了 grow-diag-final-and 词对齐与过滤后的语言模型 Mid.lm5。

### 3.4 调序模型

系统 MEBTG 与 SynMEBTG 所用的词汇化调序模型由最终的所有训练语料经最大熵工具训练得到；系统 SynMEBTG 所用的句法调序模型由句子对齐质量较好的“中英百万对齐（点通）_book”和“CLDC-LAC-2003-006”经最大熵工具训练而得。

### 3.5 实验结果与分析

#### 3.5.1 单系统汉英翻译评测结果与分析

表 6 列出了我们参加单系统评测的所有翻译系统在开发集和测试集上的性能表现。由于 Moses 工具中没有按 BLEU-SBP 训练的选项，所以产生 BLEU 值的变化与 BLEU-SBP 值变化不一致的情形，从而导致我们在“Moses(gdfa+gdf)”与” Moses(gdfa)”中选择错误。另外一个重要问题是：为什么我们没有选择 SynMEBTG 作为主系统？主要原因是：在开发集打分时，参考译文都是小写，Moses 产生的结果也是小写，而 MEBTG 与 SynMEBTG 产生的结果都是首字母大写，性能比较时采用了大小写敏感，导致 MEBTG 与 SynMEBTG 比表中的结果少了近 1.5 的 BLEU 值，从而直接导致了选择主系统的失败。

从表中数据，我们可以看到采用最大熵的词汇化调序模型(MEBTG)要优于采用 MSD 词汇化调序模型(Moses)；更进一步，利用源语言句法的调序知识并将句法与非句法分开调序(SynMEBTG)相比于词汇化调序模型(Moses, MEBTG)能显著改善译文质量。

表 6：单系统评测结果

| 参评系统 | 类别 | 开发集（BLEU-SBP,忽略大小写） | CWMT2009 测试集（BLEU-SBP,大小写敏感） |
|---|---|---|---|
| Moses(gdfa+gdf) | ia-primary | 0.2673 | 0.2223 |
| Moses(gdfa) | ia-contrastc | 0.2652 | 0.2251 |
| MEBTG | ia-contraste | 0.2714 | 0.2317 |
| SynMEBTG | ia-contrastd | **0.2791** | **0.2409** |

### 3.5.2 机器翻译评测结果与分析

- 英汉新闻机器翻译：在该项评测中，我们只使用了 Moses 系统，且用 BLEU-4 对开发集打分，与汉英单一系统相似，Moses 的训练只能最大化 BLEU 值，导致 BLEU 值的变化与 BLEU-SBP 值变化的不一致，从表中实验结果，我们看到两种词对齐启发式规则 gdfa+gdf 的合并并没有提高测试集的翻译质量，其原因我们认为是开发集与测试集的语言现象分布差距较大。

表 7：英汉新闻机器翻译评测结果

| 参评系统 | 类别 | 开发集（BLEU-4,忽略大小写） | CWMT2009 测试集（BLEU-SBP,大小写敏感） |
|---|---|---|---|
| Moses(gdfa+gdf) | ia-primary | **0.2641** | 0.3352 |
| Moses(gdfa) | ia-contrastb | 0.2626 | **0.3406** |

- 英汉科技机器翻译：在该项评测中，我们使用了三个系统：Moses、Joshua 和 SenComb。为了便于比较，这三个系统在开发集训练时采用的都是基于词级别的 BLEU-4 打分。与前面的实验一致，两种启发式规则的合并 gdfa+gdf 并没有提高最终的翻译质量。句子级融合系统非常微弱地改善了翻译结果，主要原因可能是可供选择的翻译结果比较少（只有 3 个）。

表 8：英汉科技机器翻译评测结果

| 参评系统 | 类别 | 开发集（BLEU-4,忽略大小写） | CWMT2009 测试集（BLEU-SBP,大小写敏感） |
|---|---|---|---|
| SenComb | ia-primary | **0.2859** | **0.4767** |
| Moses(gdfa+gdf) | ia-contrastb | 0.2786 | 0.4688 |
| Moses(gdfa) | ia-contrastc | 0.2744 | 0.4733 |
| Joshua | ia-contrastd | 0.2801 | 0.4656 |

### 3.5.3 系统融合评测结果与分析

我们采用 mteval-v13 脚本程序来计算 BLEU 得分，和进行词级别的系统融合的参数调整。对参与融合的系统，我们选用了开发集上排名前七的系统进行融合，即 U12, U4, U9, U6, U10, U5, U11。表 9 是三种系统融合方法在开发集与测试集上的得分。

由于三种方法在开发集上性能差异较小，在为测试集挑选最后的 primary 系统和 contrastive 系统时，我们并没有把在开发集上表现最好的词级别融合结果作为 primary 系统，而是参考 CWMT' 08 的系统融合方法，把基于 BLEU 打分的句级系统融合方法作为 primary 系统。

表9：系统融合评测结果

| 参评系统 | 类别 | SSMT2007,（BLEU-4,忽略大小写） | CWMT2009 测试集（BLEU-SBP，大小写敏感） |
|---|---|---|---|
| WordComb | ia-contrastb | **30.05** | **0.2349** |
| SenComb-Bleu | ia-primary | 29.98 | 0.2327 |
| SenComb-Meteor | ia-contrastc | 29.83 | 0.2342 |

## 4 总结

本文主要介绍了中科院自动化所参加 CWMT2009 评测的情况。在四个子项的评测中，我们都取得了较好的成绩，特别是汉英翻译任务中，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 SynMEBTG 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。然而，我们还有很大的提升空间，尤其是系统融合还不够稳定。我们需要向国内外同行学习，改善我们现有的系统。

## 4 致谢

在这次评测中，我们实验室很多同学付出了许多艰辛的劳动，如语料的预处理、时间数字等命名实体的识别与翻译，在此我们对夏睿、汪昆、刘鹏、庄涛、鉴萍、周可艳、曹文洁、王志国、翟飞飞和吴晓锋一并表示感谢！

## 参 考 文 献

[Li 2007] Chi-Ho Li, Dongdong Zhang, Mu Li, Ming Zhou Minghui Li and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. *In Proceedings of 45th Meeting of the Association for Computational Linguistics.*

[Li 2008] Maoxi Li, and Chengqing Zong. Word Reordering Alignment for Combination of Statistical Machine Translation Systems. *In the International Symposium on Chinese Spoken Language Processing (ISCSLP)*, December 16-19, 2008. Kunming, China.

[Kumar 2004] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," *In Proc. of HLT*, 2004

[Mangu 1999] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," *In Proc. Eur. Conf. Speech Commun.* Technol, 1999.

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

[Sim 2007]K.C. Sim, W.J. Byrne, M.J.F. Gales, H. Sahbi & P.C. Woodland, "Consensus network decoding for statistical machine translation system combination," *In proc. ICASSP*, volume 4, pages 105-108, 2007.

[Kemal 1996] O. Kemal, "Error-tolerant finite-state recognition with application to morphological analysis and spelling correction," *Computational Linguistics*, 22(1):73-89, 1996.

[Lin 2004] C.Y Lin, F.J Och. Orange: A method for Evaluating Automatic Evaluation Metrics for Machine Translation, *In Proceedings of COLING*, 2004.

[Wu 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.

[Xiong 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.

[Chen 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.

[Rosti 2007a] A.-V. I. Rosti, N. F. Ayan, B. Xiang et al., "Combining outputs from multiple machine translation systems," *In Proceedings of NAACL HLT 2007*, Rochester, NY, 2007, pp. 228–235.

[Rosti 2007b] A.-V. I. Rosti, S. Matsoukas, and R. Schwartz, "Improved Word-Level System Combination for Machine Translation," *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,* Prague, Czech Republic, 2007, pp. 312–319.

[Bangalore 2001] S. Bangalore, F. Bordel, and G. Riccardi, "Computing consensus translation from multiple machine translation systems," *In IEEE Workshop on Automatic Speech Recognition and Understanding, 2001.* ASRU'01, 2001, pp. 351-354.

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

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