第五届全国机器翻译研讨会（CWMT'2009）富士通研究开发中心技术报告

何中军 郑仲光 孟遥 于浩
富士通研究开发中心 北京 100025

E-mail：{hezhongjun, zhengzhg, mengyao, yu}@cn.fujitsu.com

摘 要： 本文介绍了富士通研究开发中心参加2009年第五届全国机器翻译评测的情况。我们参加了4个项目：汉英新闻领域单一系统、英汉新闻领域机器翻译、英汉科技领域机器翻译和汉蒙日常用语机器翻译。本文介绍了所采用的技术以及参加评测的结果。

关键字： 自然语言处理、机器翻译、最大熵规则选择

FRDC Technical Report for CWMT'2009

Zhongjun He Zhongguang Zheng Yao Meng Hao Yu
Fujitsu R&D Center CO., LTD, Beijing, 100025, China

E-mail：{ hezhongjun, zhengzhg, mengyao, yu}@cn.fujitsu.com

Abstract: *This paper is an overview of FRDC technical report for the 5th China workshop on machine translation. We participated in four tasks: Chinese-English news single system, English-Chinese news translation, English-Chinese science and technology translation, and Chinese-Mongolian daily expressions translation. This paper describes the techniques we used, as well as the evaluation results.*

Keywords: *natural language processing, machine translation, maximum entropy rule selection*

1 引言

富士通研究开发中心（FRDC）参加了第五届全国机器翻译研讨会（CWMT'2009）机器翻译评测中的4个项目：汉英新闻领域单一系统（ZH-EN-NEWS-SINGL）、英汉新闻领域机器翻译（EN-ZH-NEWS-TRANS）、英汉科技领域机器翻译（EN-ZH-SCIE-TRANS）以及汉蒙日常用语机器翻译（ZH-MN-DAIL-TRANS）。

本文第二章介绍参评系统；第三章介绍实验；第四章进行总结。

2 参评系统

FRDC机器翻译研究开始于2008年10月份，此次参加CMWT评测的系统主要是根据[Chiang, 2005, 2007] 中的方法实现的一个基于层次化短语模型的翻译系统“鉴真（Jian Zhen）”。此外，在翻译模型中又加入了最大熵规则选择（MERS）模型[He et al., 2008]。

$$
\begin{aligned}
S &\Rightarrow < X_1 ,\ X_1 > \\
&\Rightarrow < X_1 \text{之一},\ \text{one of } X_1 > \\
&\Rightarrow < \text{最大的城市之一},\ \text{one of the largest city} >
\end{aligned}
$$

图 1:一个推导的例子

2.1 “鉴真（Jian Zhen）”

该系统是一个基于形式化语法的翻译系统，采用上下文无关语法建立翻译模型，其规则具有如下形式：

$$X \rightarrow < \alpha, \gamma, \sim > \tag{1}$$

其中，X 是非终结符，α 和 γ 分别是源语言端和目标语言端由终结符（单词）和非终结符（变量）组成的字符串，$\sim$ 表示 α 和 γ 中非终结符的一对一的对应关系。

下面是从汉英词语对齐的语料库中自动抽取的两个规则：

$$X \rightarrow < \text{最大的城市},\ \text{the largest city} > \tag{2}$$

$$X \rightarrow < X_1 \text{之一},\ \text{one of } X_1 > \tag{3}$$

其中，规则（2）是一个短语规则，与基于短语的方法[Koehn et al., 2003]类似，能够将连续的汉语词串翻译为英语词串；而规则（3）是一个层次化规则，除了能够将其中的汉语词翻译为英语以外，还能够进行短语调序。这主要是由于层次化规则引入了变量，从而具有比连续短语更强的表达能力，能够进行远距离的短语调序。

在层次化短语模型中，翻译被看作推导（derivation），即一个不断使用规则的过程。图 1 是一个推导的例子，将汉语句子“最大 的 城市 之一”翻译为“one of the largest city”。

翻译模型采用对数线性（log-linear）模型[Och and Ney, 2002]，使用多个特征计算推导的分数：

$$w(D) = \prod_i \varphi_i^{\lambda_i} \tag{4}$$

其中，φ_i 是特征函数，λ_i 是特征函数的权重。[Chiang, 2005]共使用了 7 个特征函数：翻译概率 $P(\gamma \mid \alpha)$ 和 $P(\alpha \mid \gamma)$，词汇化权重 $P_w(\gamma \mid \alpha)$ 和 $P_w(\alpha \mid \gamma)$，n-gram 语言模型，规则个数以及目标单词数。

翻译系统最终选择分数最高的那个推导生成翻译结果。

2.1 最大熵规则选择模型

最大熵规则选择模型[He et al., 2008]改进了层次化短语模型中规则选择缺乏上下文知识的不足。

在层次化短语模型中，每个规则所包含的源语言串和目标语言串的翻译概率是在规则抽取时就计算好的，是一种“静态”的分数。在解码过程中，不能根据具体的上下文“动态”改变，容易造成规则选择的不准确，从而影响翻译质量。

例如，从语料库中抽取到下面三个规则：

$$X \rightarrow \langle 在X_1 的X_2, X_2 \text{ in } X_1 \rangle \tag{5}$$

$$X \rightarrow \langle 在X_1 的X_2, \text{at } X_1\text{'s } X_2 \rangle \tag{6}$$

$$X \rightarrow \langle 在X_1 的X_2, \text{with } X_2 \text{ of } X_1 \rangle \tag{7}$$

这三个规则具有相同的源语言端“在 X_1 的 X_2”，但是却对应完全不同的三个目标语言端。在解码过程中，其源语言端都能匹配下面三个短语：

短语 1：　在 [经济 领域]$_1$ 的 [合作]$_2$

短语 2：　在 [今天]$_1$ 的 [会议 上]$_2$

短语 3：　在 [人民]$_1$ 的 [支持 下]$_2$

如果将规则（6）用于短语 1，则会得到不正确的翻译：

at [economic field]$_1$'s [cooperation]$_2$

而实际上，应该使用规则（5）翻译短语 1，得到：

[cooperation]$_2$ in [economic field]$_1$

造成这一问题的一个主要原因是，在规则选择的过程中，缺乏相应上下文的指导。

最大熵规则选择模型以最大熵模型为基础，融合了丰富的上下文知识，能够在解码过程中根据具体的上下文环境进行规则选择：

$$P_{rs}(\gamma \mid \alpha, f(X_k), e(X_k)) = \frac{\exp[\sum_i \lambda_i h_i(C(\gamma), C(\alpha), f(X_k), e(X_k))]}{\sum_{\gamma'} \exp[\sum_i \lambda_i h_i(C(\gamma'), C(\alpha), f(X_k), e(X_k))]} \tag{8}$$

其中，$X_k(k=1,\cdots K)$ 是非终结符，$f(X_k)$ 和 $e(X_k)$ 是非终结符所对应的具体的短语，$C(\alpha)$ 和 $C(\gamma)$ 是源语言和目标语言端上下文，$h_i(C(\gamma), C(\alpha), f(X_k), e(X_k))$ 是特征函数，λ_i 是特征函数的权重。[He et al., 2008]为最大熵模型定义了三种类型的特征：

- 单词特征：α 所匹配的源语言短语左右紧邻的单词以及规则内部非终结符 X 所对应的子短语的边界词。
- 词性特征：单词特征中源语言单词所对应的词性。
- 长度特征：非终结符 X 所对应的子短语的长度

为了将 MERS 模型融合到翻译模型中，又在翻译模型中增加了两个特征：

- 规则选择特征 $P_{rs}(\gamma \mid \alpha, f(X_k), e(X_k))$；

- 歧义规则个数特征 $P_{rsn}=\exp(1)$。如果规则中的源语言端对应多个目标译文，那么这个规则就是歧义规则。在解码过程中，如果一个规则是歧义规则，那么 $P_{rsn}=\exp(1)$，否则，$P_{rsn}=\exp(0)$。

3 实验

3.1 语料库

表 1 列出了我们使用的全部语料库。

表 2 列出了在不同的评测项目中，所使用的语料库组合。

我们没有使用路透社语料库训练英语语言模型，原因是在实验中我们发现它没有提高系统的 BLEU 值。

表 1：所用语料资源列表

类别	资源编号	描述
C1（2.7M 句对）	CLDC-LAC-2003-004	中科院计算所和自动化所中英句子级对齐双语语料库（扩充版）
	CLDC-LAC-2003-006	北京大学汉英/汉日双语语料库（汉英部分）
		厦门大学英汉电影字幕平行语料库
		哈工大信息检索组英汉句子级对齐语料库
		哈工大机器翻译组英汉句子级对齐语料库
C2（0.9M 句对）		中信所英汉科技文献句子级对齐语料库（2009 年版）
		万方汉英中文科技期刊论文摘要语料库
D05	2005-863-001	2005 年 863 机器翻译评测数据（英汉汉英机器翻译部分）
D08		2008 年 CWMT 评测数据
T	2007-863-001	SSMT2007 机器翻译评测数据（英汉汉英机器翻译部分）
M1（67K 句对）		内蒙古大学汉蒙平行语料库（2009 版）
M2（999K 单词）		内蒙古大学蒙古语单语语料库（2009 版）
S（589M 词语）		搜狗全网新闻语料库(SogouCA)

表 2：各项目所用语料库列表

评测项目代号	双语语料库	单语语料库	开发集
ZH-EN-NEWS-SINGL	C1	-	D05(zh-en write)
EN-ZH-NEWS-TRANS	C1+C2	S	D05（en-zh write）
EN-ZH-SCIE-TRANS	C1+C2	S	D08（en-zh-scie）
ZH-MN-DAIL-TRANS	M1	M2	评测单位提供的开发集

表 3：在 SSMT07、CWMT08 以及汉蒙开发集上的结果

系统	汉英		英汉		汉蒙
	07CE-NEWS	08CE-NEWS	08EC-NEWS	08EC-TECH	09-DEV
Jian Zhen	0.2350	0.2332	0.3220	0.3944	0.2851
+MERS	0.2420	0.2406	-	-	0.3023

表 4：CWMT2009 评测结果

系统	项目			
	ZH-EN-NEWS-SINGL	EN-ZH-NEWS-TRANS	EN-ZH-SCIE-TRANS	ZH-MN-DAIL-TRANS
Jian Zhen	-	0.3345	0.4567	-
+MERS	0.2106	-	-	0.2166

3.2 训练

中文分词和词性标记采用 FRDC 开发的分词程序[Meng et al., 2005]。

词语对齐首先采用 GIZA++进行双向训练，然后使用“grow-diag-final”[Koehn et al., 2003]方法进行优化。

语言模型使用 SRILM 工具[Stolcke, 2002]，在 ZH-EN-NEWS-SINGL 项目上训练了一个 6-gram 的英语语言模型；在 EN-ZH-NEWS-TRANS 和 EN-ZH-SCIE-TRANS 上训练了两个 4-gram 的汉语语言模型；在 ZH-MN-DAIL-TRANS 上训练了两个 6-gram 蒙语语言模型。

最大熵规则选择模型使用 Le Zhang 开发的最大熵模型训练工具[Zhang, 2004]。

翻译模型的调参使用最小错误率方法[Och, 2003]。我们对目标函数进行了修改，以适应评测中使用的 BLEU-SBP[Chiang et al., 2008]指标。

3.3 预处理和后处理

在汉英翻译项目中，使用规则的方法翻译数词和时间词，翻译完毕，删除了未登录词；

对于英汉翻译和汉蒙翻译，没有做特殊的处理。

3.4 实验结果

表 3 列出了正式评测之前，我们的系统在 SSMT2007、CWMT2008 以及汉蒙开发集上的实验结果。其中，汉英和汉蒙翻译使用 BLEU-4-SBP，英汉翻译使用 BLEU-5-SBP。英汉翻译以字为单位进行打分。

汉英和汉蒙翻译给出了两个系统的运行结果，分别是“Jian Zhen”和“Jian Zhen + MERS”。实验结果表明，加入最大熵规则选择模型后，BLEU 值在汉英翻译上提高了约 0.7 个百分点（0.2350->0.2420, 0.2332->0.2406），在汉蒙翻译上提高了约 1.7 个百分点（0.2851->0.3023）。英汉翻译由于时间原因，没有加入最大熵规则选择模型。

表 4 列出了我们参加正式评测的结果。对于汉英和汉蒙翻译，使用“Jianzhen + MERS”系统，对于英汉翻译，使用“Jianzhen”系统。需要注意的是，在汉英单系统项目中，我们仅使用了评测单位提供的部分语料（C1）训练翻译模型，使用 C1 的英语部分训练 6-gram 语言模型。

4 总结

富士通研究开发中心参加了 CWMT2009 机器翻译评测中的汉英新闻领域单一系统、英汉新闻领域机器翻译、英汉科技领域机器翻译以及汉蒙日常用语机器翻译 4 个项目。本文对参评情况进行了介绍。

今年是富士通研究开发中心第一次参加 CWMT 评测。由于系统开发时间较短，目前还不够稳定。今后，我们将继续完善现有系统。同时，也希望能够和各参评单位进行深入的交流与合作。

参考文献

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Pages 263-270.

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

Chiang, David, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of EMNLP 2008*. Pages 610-619.

He, Zhongjun, Qun Liu and Shouxun Lin. 2008. Improving Statistical Machine Translation using Lexicalized Rule Selection. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Pages 321-328.

Koehn, Philipp, Franz J. Och and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*. Pages 127-133.

Meng, Yao, Hao Yu, Fumihito Nishino. 2005. A Lexicon-Constrained Character Model for Chinese Morphological Analysis. In *Proceedings of IJCNLP 2005*. Pages 542-552.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Pages 295-302.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Pages 160-167.

Stolcke, Andreas. 2002. SRILM -- An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken language Processing*. Pages 901-904.

Zhang, Le. 2004. Maximum Entropy Modeling Toolkit for Python and C++. Available at http://homepages.inf.ed.ac.uk/s0450736/maxent_too-lkit.html