第七届全国机器翻译研讨会（CWMT2011）

富士通研究开发中心技术报告

李贤华 郑仲光 孟遥 于浩

富士通研究开发中心 北京 100025

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

摘 要：本文介绍了富士通研究开发中心参加2011年第七届全国机器翻译研讨会的评测情况。今年FRDC参加了两个项目：英汉科技领域的机器翻译以及日汉新闻领域的机器翻译。本文介绍了FRDC做统计机器翻译的主要情况、本次参加评测所采用的系统情况、语料情况、在开发集和测试集的实验结果，并对实验结果进行了分析与总结。

关键字：自然语言处理、机器翻译、短语模型、层次短语模型

FRDC Technical Report for CWMT2011

Xianhua Li Zhongguang Zheng Yao Meng Hao Yu

Fujitsu R&D Center CO., LTD, Beijing, 100025, China

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

Abstract: *This paper is an overview of FRDC technical report for the 7th China workshop on machine translation. We participated in two tasks: English-Chinese science machine translation and Japanese-Chinese news machine translation. This paper introduces the main history of FRDC in statistical machine translation, describes the systems and corpora adopted in the evaluation. We also present the results and have a discussion on the results.*

Keywords: *natural language processing, machine translation, phrase-based model, hierarchical phrase-based model*

1 引言

富士通研究开发中心（FRDC）的统计机器翻译研究始于2008年10月，至今经过了将近三年时间的发展。今年，FRDC参加了第七届全国机器翻译研讨会（CWMT2011）机器翻译评测中的两个项目：英汉科技领域机器翻译（EN-CH-SCIE）以及日汉新闻领域机器翻译（JP-ZH-NEWS）。这也是FRDC第二次参加全国机器翻译研讨会。

FRDC今年参加CWMT的评测系统，主要有开源的机器翻译系统Moses（phrase-based和hierarchical phrase-based），以及根据[Chiang，2005，2007]中介绍的方法实现的一个基于层次短语的机器翻译系统“鉴真”（Jianzhen）。

本文第二章介绍了主要参评系统；第三章介绍实验所用数据，第四章是实验设置和结果，以及错误分析；第五章进行了总结。

2 系统

2.1 Moses

摩西是一套开源的统计机器翻译系统。最初的 Moses 只有短语模型，在众多研究人员的集体开发下，Moses 已经扩展成为一个集短语模型、层次短语模型、句法模型等为一体的统计机器翻译系统。

2.1.1 Phrase-based Moses

基于短语模型的 Moses 是当前应用最广泛、最成熟的统计机器翻译系统。它采用对数线性模型[Och and Ney, 2003]将候选译文的得分描述为若干特征的线性组合，最后通过 Beam Search 算法来获得最佳译文。对数线性模型的使用，使得在系统中添加新的特征十分方便。对数线性模型的参数通过在开发集上进行最小错误率训练[Och, 2003]得到。

$$P\gamma(e|f) \propto \sum_{i} \lambda_i h_i(\alpha, \gamma) \quad (1)$$

在公式（1）中，$h_i(\alpha,\gamma)$是特征函数，λ_i其对应的h_i的权重。对于不同的候选翻译，通过对数线性模型计算其得分，得分最高的即为最佳译文。

在本次评测中，FRDC 使用了 Moses 的短语系统作为参评的对比系统，可以衡量 Moses 的短语模型和层次短语模型之间的差距。

2.1.2 Hierarchical phrase-based Moses

最初的 Moses 是基于短语模型的统计机器翻译系统。在 2010 年 3 月，Moses 的主要负责人之一 Hieu Hoang 在 Moses 的邮件列表中发了一份新版 Moses 的介绍信，介绍了 Moses 的新增功能。新版的 Moses 增加了 Chiang 的层次短语模型以及句法模型，改进了 MERT，并有许多其他功能上的改进。

层次短语模型的相关介绍见 2.2 节。

在本次评测中，FRDC 使用了 Moses 的层次短语系统作为参评的主系统。经过测试，Moses 的 HPB 系统相对来说，翻译效果比较稳定，翻译质量较好。

2.2 鉴真

鉴真是由富士通研究开发有限公司研发的基于层次短语模型的统计机器翻译系统。它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。层次短语模型的规则具有下面的形式：

$$X \rightarrow \langle \alpha, \gamma, \sim \rangle \quad (2)$$

其中，X 是非终结符，α 是源语言端，γ 是目标语言端，$\sim$ 表示 α 和 γ 中非终结符的对应关系。

层次短语模型中的规则主要分为普通规则和粘贴规则。其中，普通规则又可以分为短语规则和层次化短语规则。

普通规则如下面两个规则所示：

$$X \rightarrow \langle \text{关键 步骤}, \text{ the key steps} \rangle \quad (3)$$

$$X \rightarrow \langle X_1 \text{之一}, \text{ one of } X_1 \rangle \quad (4)$$

其中，规则（3）是一个短语规则，这类规则和短语模型[Koehn et al., 2003]中的规则类似，主要实现了词串的翻译功能；规则（4）是一个层次化规则，层次化规则中引入了变量，使得这类规则具有较好的泛化能力，能够进行远距离调序。

层次短语模型的可用的粘贴规则，主要有如下两条：

$$S \rightarrow \langle X, X \rangle \quad (5)$$

$$S \rightarrow \langle SX, SX \rangle \quad (6)$$

粘贴规则主要实现了顺序翻译的功能。如果系统中只使用短语规则以及粘贴规则，那么层次短语模型就相当于一个没有调序功能的短语模型。

$$\begin{aligned} S &\Rightarrow \langle X_1, X_1 \rangle \\ &\Rightarrow \langle X_1 \text{之一}, \text{ one of } X_1 \rangle \\ &\Rightarrow \langle \text{关键 步骤 之一}, \text{ one of the key steps} \rangle \end{aligned}$$

图 1. 一个推导的例子

在层次化短语模型中，翻译的过程被看作是推导（derivation）的过程，即不断使用规则的过程。图 1 是一个推导的例子，在这个例子中，将汉语短语“关键 步骤 之一”翻译为“one of the key steps”。

层次短语模型与短语模型一样，采用对数线性模型组合多个特征及其权重，并计算推导的得分。特征权重由最小错误率训练得到。FRDC 的系统 Jianzhen 共使用了 7 个特征函数：正反向翻译概率 $P(\gamma|\alpha)$ 和 $P(\alpha|\gamma)$，正反向词汇化翻译概率 $P_w(\gamma|\alpha)$ 和 $P_w(\alpha|\gamma)$，n-gram 语言模型，规则个数以及目标单词数。

系统最终选择得分最高的推导，生成最终翻译结果。

3 数据

富士通在本次评测中使用的数据，全部来自于主办方提供的资源清单。其中，主办方提供的训练数据经过长句过滤后得到的句对，作为本次评测的训练数据。

表 1 列出了 FRDC 在英汉科技翻译中的训练数据。

表 2 列出了 FRDC 在日汉新闻翻译中的训练数据。

表 3 列出了 FRDC 在参评项目中使用的语料情况。

表 1：英汉科技翻译中所用训练数据

C1	名称	中信所英汉科技文献句子级对齐语料库（2011 年版）
	提供单位	中国科学技术信息研究所
	语种	英语→汉语
	领域	科技
	规模	91.1527 万句子对
	说明	英文科技文献及其汉语翻译。 含 CWMT2009 评测中提供的 60 万句子对。

表 2：日汉新闻翻译中所用训练数据

C2	名称	南京大学日汉双语句子对齐语料库
	提供单位	南京大学
	语种	日语-汉语
	领域	综合领域
	规模	15 万句对
	说明	该平行语料库是从 Web 上自动挖掘获得的。
C3	名称	大连理工大学日汉双语对齐语料库
	提供单位	大连理工大学
	语种	日语-汉语
	领域	综合领域
	规模	约 14 万句对
	说明	Web 自动挖掘与人工挖掘相结合，仅供研究目的。
C4	名称	北京大学汉英/汉日双语语料库（汉日部分）
	提供单位	北京大学计算语言学研究所
	语种	汉英，汉日
	领域	综合
	规模	汉英句子级对齐语料 20 万句对；汉日句子级对齐语料 2 万句对；汉英词汇级对齐语料 1 万句对。
	说明	在 863 课题《中文平台总体技术研究与基础数据库建设》子课题《汉英/汉日多语语料库》（编号：2001AA114019）资助下开发而成。 参加日汉翻译项目评测的单位仅提供汉日部分的语料。

表 3：各项目所用语料列表

评测项目代号	训练集	语言模型	开发集
EN-ZH-SCIE	C1	C1 中文部分	评测单位提供的开发集
JP-CH-NEWS	C2+C3+C4	C2+C3+C4 中文部分， 部分搜狗语料	评测单位提供的开发集

4 实验

4.1 语料预处理

在英汉科技翻译中，对于英文语料，我们进行了 token 处理，大小写转换；对于汉语语料，我们进行了全半角转换，中文分词[Meng et al., 2005]。

在日汉新闻翻译中，对于日语语料，我们进行了日语分词（chasen），半角字符合并等处理；对于汉语语料，我们进行了全半角转换，中文分词。

4.2 训练

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 EN-CH-SCIE 和 JP-ZH-NEWS 上分别训练了一个 5-gram 的汉语语言模型，并使用部分搜狗全网新闻语料库训练了一个 5-gram 的汉语语言模型。翻译模型的调参使用最小错误率方法[Och, 2003]。

4.3 后处理

在英汉科技翻译中，删除了未登录词；在日汉新闻翻译中，进行了繁简体转换、删掉结果中的假名等后处理。

4.4 实验结果

表 4 和表 5 列出了正式评测之前，FRDC 的参评系统在主办方提供的开发集上得到的实验结果。英汉科技翻译和日汉新闻翻译均使用 BLEU-5-SBP，以字为单位进行打分。

表 4：在英汉开发集上的结果

系统	英汉科技（BLEU5-SBP）
Moses-PB	0.4240
Moses-HPB	0.4304

表 5：在日汉开发集上的结果

系统	日汉新闻（BLEU5-SBP）
Moses-PB-S-LM	0.3603
Moses-HPB-ori	0.3361
Moses-HPB-pre	0.3386
Moses-HPB-S-LM	0.3549
Jianzhen	0.3191

评测结束之后，我们用搜狗全网新闻语料库训练得到一个 4 元的汉语新闻语言模型，并在日汉新闻翻译中进行了实验。结果表明，使用搜狗全网新闻语料库训练的语言模型，对于系统性能有很大的提升。在此基础上，我们使用了 Moses 提供的 mgiza 取代 GIZA++进行对齐，在解码过程中中断了一次，继续解码后，在开发集上的成绩如表 6 所示。

表 6：使用搜狗大语言模型和 mgiza

系统	日汉新闻（BLEU5-SBP）
Moses-HPB-L-LM	0.3831
Moses-HPB-L-LM+mgiza	0.3781

表 7 列出了 FRDC 正式评测的结果。英汉科技翻译使用 Moses PB 和 Moses HPB 两个系统；日汉新闻翻译使用 Moses PB、Moses HPB 和 Jianzhen 系统。

表 7：CWMT2011 评测结果

	项目	
系统	EC-CH-SCIE	JP-CH-NEWS
Moses-PB	0.3668	0.4008
Moses-HPB	0.3729	0.4108
Jianzhen		0.4033

4.5 错误分析

我们对日汉新闻的翻译结果进行了分析。下面是我们总结的主要错误、案例及分析。

问题 1：日语和汉语的语序问题

日语：主宾谓 VS 汉语：主谓宾，以及一些类似的问题（从...到...等），导致译文的语序出现问题。

翻译结果样例：

1.将来 在 一起 ，互相 晚年 互相 帮助 ” 打算 。
2.会议 市场 价格 的 监督 管理 他 作为 以下 四 方案 提出 了 。
3.作为 精神 的 暴力 ，丈夫 妻子 不但 无视 的 行为 ，也 被 作为 女性 的 身心 大 伤 问题 认知 了 。

可能的解决方法：

1. 语序调序预处理：在训练之前，进行语料的预处理，即将汉语语料和日语语料的语序调整为一致，再进行训练，调参和测试的语料也进行这样的预处理
2. 抽取长距离调序规则：想办法抽取长距离调序规则，例如将中日两端进行语块划分，这样类似于进行了粗粒度的分词，句子结构更明确，也能抽取出长距离调序规则
3. 翻译后处理：对翻译的结果进行后处理，将动词提取到其对应的宾语前面

问题 2：人名识别翻译问题（命名实体翻译）

由于分词过于细致，许多人名和命名实体等被全部切分，再单独翻译，导致翻译出现了很多的问题

翻译结果样例：

1.市 妓女 联 经营 的 DV 商量 ，95.2% 的 受害者 实际 伤害 ，障碍 留下 的 严重 的 情况 ，和 可以 刑事 案件 的 重大 的 例子 也 了 。
2.影响 近平 会见 副 主席 、 长崎 县 知事
3.北京大学 华 大学 ，清朝 东京都 首 次 进 鬼录 世界 前 200 名 ，复 大学 、南京 大学 、上海 交通 大学 、中国 科学 厨艺 大家 讨论 大学 、浙江 大学 五 所 大学 排 在 第 201-300 名 。

4. 目前，北京市 统计局 管辖 区域 内 有 三 路线 —— 集 宁 和 张 家 嘴 （ 河北 ） 蝴蝶结 路线

可能的解决方法：

1. 分词后处理：分词后进行命名实体识别，将命名实体合并，再进行规则抽取和翻译；如果无法翻译，则可以通过简单的繁简转换得到翻译结果（基于中日的命名实体很多是繁简对应的汉字）
2. 双语命名实体获取：通过双语命名实体获取，预先得到双语命名实体词典，在翻译中使用

问题 3：数词的翻译

数词是翻译中经常出现的，但是简单的规则抽取并不能很好的对数词进行翻译

翻译结果样例：

1.三十 天 从 早上 雨 中，广东省 春节 特别 运输 体制 和 市场 供应 的 情况 视察 了 。

2.外交部 副 部长，党委 书记 王 光 亚硝酸（ 六 ）全国 政协 副 主席 的 寥 晖（ 六 十八 ）作为 后任 、 国务院 香港 澳门 事务 办公室 主任 调 、 香港 、 澳门 的 媒体 的 关注 。

可能的解决方法：

1. 规则表清理：在规则表中，有许多数字互相翻译的规则，可以将不正确的规则删除，防止其被使用
2. 规则增加：全半角的数字，对应翻译为全半角的数字（译文与原文相同，不发生变化）。将这样的规则添加进规则表，并赋予较高概率，提高匹配可能
3. 预处理：预处理数字，将其进行翻译后，再进行整句的翻译

问题 4：成对的标点符号的处理

成对的标点符号，比如《》“”‘’ <>等，里面的内容一般可以作为一个整体进行翻译，而成对的标点符号一般不应调序，然而实际的翻译中出现由于调序，导致成对的标点符号翻译混乱的现象

可能的解决方法：

1. 可以将成对的标点符号的内容单独提取出来进行翻译，再与原来的结果进行合并，在训练的时候，也可以做类似的处理，使得抽取的规则不会与成对的标点符号之间的内容交叉

问题 5：大量的片假名没有翻译

应该是训练语料里没有学习到对应的规则

翻译结果样例：

1.水产 品 的 药品 污染 检测 合格率，ニトロ フラン 1.8 个 百分点，マラカイトグリーン，クロロマイセチン 0.3 个 白分点，分别 上升 0.9 个 白分点 。

2.スーダン 访问 为 当天 ハルツーム 的 刘 大使 “ 中国 梦寐以求 的 是 ダルフール 地方 早期 的 和平 实现，尽快 スーダン 稳定 和 发展 实现 同时，这个 地方 实现 和平 。

可能的解决方法：

1. 在翻译结果中，作为未登录词直接删除
2. 可以研究采用音译的方法，毕竟日文中的片假名大多数用来表示外来词汇，而外来词汇一般为英文单词读音或者中文单词读音

当然，日汉翻译中肯定还存在许多问题，我们提出的解决方法也不一定是最适合的。这些都是我们下一步要解决的问题。我们也希望能和大家加强沟通和交流，共同进步。

5 总结

富士通研究开发中心参加了 CWMT2011 机器翻译评测中的英汉科技翻译、日汉新闻翻译两个项目。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二次参加 CWMT 评测。今后，我们将继续完善现有系统，并引入更加先进的模型。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

David Chiang. 2005. *A Hierarchical Phrase-based Model for Statistical Machine Translation*. In Proc. of the 43th Annual Meeting on Association for Computational Linguistics, pages 263-270.

David Chiang. 2007. *Hierarchical Phrase Based Translation*. Computational Linguistics, Vol 33, pages 201-228.

Zhongjun He, 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.

Phillip Koehn, Franz Josef Och, Daniel Marcu. 2003. *Statistical phrase-based translation*. In Proc. of NAACL03.

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

Franz Josef Och. 2003. *Minimum Error Rate Training for Statistical Machine Translation*. In Proc. of the 41st Annual Meeting of the ACL. Vol 1, pages 160-167.

Franz Josef Och, Hermann Ney. 2003. *A systematic comparison of various statistical alignment models.* Computational Linguistics, 29(1).

Franz Josef Och, Hermann Ney. 2004. *The alignment template approach to statistical machine translation.* Computational Linguistics, pages 417-449

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation.* In Proc. of the 40th Annual Meeting on Association for Computational Linguistics, pages 311-318.

Andreas Stolcke. 2002. *Srilm-an Extensible Language Modeling Toolkit.* In Proc. of the International Conference on Spoken Language Processing. Vol 2, pages 901-904.