第九届机器翻译研讨会厦门大学技术报告

胡金铭[1], 谭 波[1], 黄研州[1], 罗 凌[1], 邬昌兴[1], 徐 伟[1],
何钟豪[1], 陈毅东[1], 史晓东[1], 苏劲松[2]

（1. 厦门大学 智能科学与技术系, 福建 厦门 361005;
2. 厦门大学 软件学院, 福建 厦门 361005）

摘要：本文主要介绍了厦门大学自然语言处理研究室（XMU-NLPRL）参加第九届全国机器翻译研讨会（CWMT2013）翻译评测任务的情况。在本次评测中，XMU-NLPRL 参加了 6 个评测项目中的 3 个子项—汉英、英汉新闻领域机器翻译和藏汉政府文献机器翻译。报告内容主要阐述本次参评系统的实现框架、模型以及它们在评测数据上的性能表现，同时对各翻译结果进行了比较和分析。

关键字：层次短语模型；短语模型；机器翻译

XMU Technical Report for the CWMT 2013

Hu Jinming[1], Tan Bo[1], Huang Yanzhou[1], Luo Ling[1], Wu Changxing[1], Xu Wei[1],
He Zhonghao[1], Chen Yidong[1], Shi Xiaodong[1], Su Jinsong[2]

(1. Cognitive Science Department, Xiamen University, Xiamen, Fujian 361005, China;
2. Software School, Xiamen University, Xiamen, Fujian 361005, China)

Abstract: In this paper, we present an overall introduction of the translation evaluation task of the 9th China Workshop on Machine Translation (CWMT2013), submitted by Xiamen University Natural Language Processing Research Lab (XMU-NLPRL). In this evaluation, XMU-NLPRL takes part in 3 translation sub-tasks of the 6 evaluation tasks, which involve the domain of Chinese-to-English and English-to-Chinese news, Tibetan-to-Chinese government document. The report mainly describes the implementation framework, model and the performance in the evaluation data set of the evaluated translation system. In addition, all translation results are compared and analyzed.

Key words: hierarchical phrase-based model; phrase-based model; machine translation

1 引言

作为参评单位之一，厦门大学参加了三个评测任务：汉英新闻领域机器翻译任务；英汉新闻领域机器翻译任务；藏汉政府文献机器翻译任务。在本次评测中使用了实验室开发的基于层次短语的系统 infdecoder 以及开源工具 Moses (Philipp Koehn, et al, 2007)。对于这三个子项，我们分别提交了汉英新闻领域（3 个系统），英汉新闻领域（3 个系统）和藏汉政府文献（4 个系统）。

在后文中第二部分会简要的介绍各个参评系统的描述及原理，第三部分会介绍实验中前期数据的处理和整个实验的流程。最后会给出评测系统的结果，并进行若干分析，提出本次实验的诸多问题以及影响评测结果的几个因素。

2 系统简介

本次机器翻译评测中我们使用了 2 个翻译系统，即：基于层次短语的机器翻译系统 (infdecoder) 以及基于短语的翻译系统 (Moses)。下面我们将对系统进行简单的介绍。

2.1 infdecoder

infdecoder 是我们实验室开发的解码器，它实现了基于层次短语的模型(Chiang, 2007)。基于层次短语的模型使用具有泛化变量的同步上下文无关文法，其形式为：

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle$$

其中 X 表示非终端符号，γ 和 α 分别是

源端与目标端终端符号以及非终端符号组成的串。~为源端以及目标端非终端符号的对应关系。

infdecoder 支持多个语言模型与翻译模型的使用。它实现了 CYK+ 算法(J.-C. Chappelier and M. Rajman, 1998)，不限制同步上下文文法中非终端符号的个数。但在实验中我们限制非终端符号的数目为 2，与 Chiang 文献中相同。与 Chiang 的文献中不同的是，infdecoder 在 cube pruning 时以翻译候选弹出的个数作为终止条件，而非 Chiang 文献中描述的与最优候选相差的阈值 ε。

2.2 Moses

Moses 是目前非常流行的开源统计机器翻译系统，它包含多个模型，在这里我们使用 Moses 里面的短语模型(Koehn et. al. 2003)。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。

我们在评测中使用的版本是 RELEASE-1.0。

2.3 未登录词替换

我们在藏汉政府文献评测中发现最后的结果中未登录词的数量非常多，受不同分词工具的影响非常大。使用不同的分词工具，有一些未登录词是可以在 GIZA++对齐中找到的。因此我们使用 GIZA++中找到的未登录词，综合考虑 GIZA++分值以及替换后句子语言模型分值，选择合适的词进行未登录词替换。

3 语料处理

3.1 语料的预处理

我们对所有语料（双语训练语料，开发集，测试集，语言模型训练语料）进行了一定的预处理，具体预处理步骤如下：

英汉与汉英：

(1) 行尾转换
(2) 全角转半角
(3) 英文部分的标点处理
(4) 转义字符的处理
(5) 乱码过滤
(6) 部分已分词语料的还原
(7) 针对训练语料中部分语料的引号不匹配问题进行处理
(8) 去除中文语言模型训练语料中带有网址的广告语句，去除英文中的类表格语句
(9) 对中文语言模型训练语料（搜狗）进行分句
(10) 分词
(11) 领域分类
(12) Truecase

藏汉：

(1) 行尾转换
(2) 全角转半角
(3) 转义字符的处理
(4) 分词

3.2 分词与分句

分句我们使用自行开发的 MYLINE 工具。

分词方面，在汉语上使用自行开发的汉语分词工具 segtag；在英语上，使用自行开发的英语 token 工具 tokenize；在藏语上，使用自行开发的藏语分词工具 ts。

3.3 领域分类

我们使用开源工具 tmsvm 对语料进行新闻类和非新闻类的区分工作。

在对汉语语料进行文本分类时，我们在模型训练阶段选用开源的搜狗新闻分类语料 SogouC.tar.gz 进行新闻类语料的训练，选用部分医学、文学以及口语类语料进行非新闻类的训练。

在对英语语料进行文本分类时，我们借用对汉语语料进行文本分类的结果，在中英文对齐的语料中提取阈值大于 0.95 且句子长度大于 150 字符的英文语料句子，并随机选出 4000 个句子作为英文新闻类训练语料；在阈值小于 0.5 的句子中随机抽取 3000 个句子作为非新闻类，并加入部分口语类英文文本作为英文非新闻类的训练语料。

在模型预测时，我们做了大量人工判别工作并细心选择阈值，在最终的文本分类结果文件中随机抽样并进行新闻类的人工判断，预测结果的准确率在 90%以上。

3.4 truecase 与词对齐

truecase 我们使用 Moses 中的 truecase 脚本。

词对齐方面，我们采用 GIZA++。双语语料完成双向词对齐后，采用启发规则合并两个方向的词对齐结果，作为最终的词对齐结果。

3.5 语言模型

语言模型采用 SRILM(Stolcke et al. 2002)。

我们针对参加的项目使用了不同的语言模型组合，在实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结合方式。

对于汉英新闻，以英文路透社语料和该项目双语训练语料的目标语言作为一个集合使用了分类工具进行新闻/非新闻分类，然后将新闻累语料作为一个训练集，训练得到的 5 元模型，我们称之为 CE-LM1，将非新闻语类料作为另一个训练集，训练得到的 3 元模型，我们称之为 CE-LM2。

对于英汉新闻，以中文搜狗语料和该项目双语训练语料的目标语言作为一个集合使用了分类工具进行新闻/非新闻分类，然后将新闻类语料作为一个训练集，训练得到的 5 元模型，我们称之为 EC-LM1，将非新闻类语料作为另一个训练集，训练得到的 3 元模型，我们称之为 EC-LM2。

对于藏汉政府文献，以中文搜狗语料作为一个训练集，训练得到的 5 元模型，我们称之为 TC-LM1。将该项目双语训练语料的目标语言作为一个训练集，训练得到的 5 元模型，我们称之为 TC-LM2。

3.6 数据过滤

对比系统中对双语训练语料进行了过滤处理。根据词对齐单向结果和启发规则后的合并结果对词对齐的好坏做出评价，过滤掉认为较差的词对齐结果。从评测的结果上来看，在英汉方向上有一定的提高。

3.7 语料的使用

在训练集上，我们语料的使用情况如下：

表 1 语料的使用情况

评测项目	训练集	过滤后
英汉新闻	4631610	2901607
汉英新闻	4631610	2901607
藏汉政府文献	109381	109381

语言模型使用情况如下：

表 2 语言模型的使用情况

评测项目	语言模型
英汉新闻	CE-LM1 + CE-LM2
汉英新闻	EC-LM1 + EC-LM2
藏汉政府文献	TC-LM1 + TC-LM2

语言模型的大小为：

表 3 语言模型的大小

评测项目	语言模型大小
英汉新闻	3.8G + 138M
汉英新闻	15G + 379M
藏汉政府文献	15G + 19M

4 实验

4.1 运行环境

系统运行环境如下：

CPU： Intel（R）-Xeon（R）11 个 2.93GHZ

内存：96GB

操作系统：Ubuntu Server-64bit

4.2 系统说明

本节我们简要说明我们提交系统的构成。

汉英新闻评测我们提供了 3 个系统，分别为：

表 4 汉英新闻系统对照

评测系统	解码器
primary-a	infdecoder
contrast-b	Moses
contrast-c	Moses

其中 primary-a 为层次短语主系统。

contrast-b 为短语对比系统，使用了经过过滤的训练语料。contrast-c 为短语系统 baseline。

英汉新闻评测我们提供了 3 个系统，分别为：

表 5 英汉新闻系统对照

评测系统	解码器
primary-a	infdecoder
contrast-b	Moses
contrast-c	Moses

其中 primary-a 为层次短语主系统。contrast-b 为短语对比系统，使用了经过过滤的训练语料。contrast-c 为短语系统 baseline。

藏汉政府文献评测我们提供了 4 个系统，使用了不同的分词工具，分别为：中科院提供的 ICTCLAS 以及我们实验室的 segtag。我们在藏汉政府文献中对未登录词进行了一些替换，在下面用 OOV 进行标记。4 个系统分别为：

表 6 藏汉政府文献系统对照

评测系统	解码器
primary-a	infdecoder (ICT+OOV)
contrast-b	infdecoder (ICT)
contrast-c	infdecoder (segtag)
contrast-d	infdecoder (segtag + OOV)

其中 primary-a 为层次短语主系统，使用计算所分词，对未登录词进行了一些替换。contrast-b 为层次短语系统 baseline，使用计算所分词。 contrast-c 为层次短语系统 baseline，使用实验室内部开发分词。contrast-d 为层次短语系统，使用实验室内部开发分词，对未登录词进行了一些替换。

4.3 训练、调参与解码

我们使用 Moses 中的程序进行训练。对于层次短语模型，我们选取 initial phrase length 为 7。对于短语模型，我们使用 Moses 默认的参数进行训练。

我们的解码器 infdecoder 中一共使用了 9 个特征，分别为：

(1) 主语言模型
(2) 次语言模型
(3) 翻译概率
(4) 词汇化概率
(5) 反向翻译概率
(6) 反向词汇化概率
(7) 规则惩罚
(8) glue 规则惩罚
(9) 词惩罚

我们利用最小错误率训练(Och, 2003)进行调参，使用 Moses 中的 mert 程序，以 BLEU(Papineni et al. 2002)作为指标。

在获得最优参数之后，我们进行解码。解码器 infdecoder 所使用的部分参数如下：

表 7 解码时参数

参数	值
beam size	500
beam threshold	-11.5
cube pruning pop limit	10000

在解码器 Moses 中，我们将 beam size 改为 500，其它均为默认参数。

4.4 后处理

对于汉英新闻译文处理，我们去除 kg、-year-old、n't 等前面的空格，并且去除未登录词。对多个连续的标点符号进行过滤，保留合法的情况。另外，我们对“万”翻译成 million 也进行了处理。

对于英汉新闻译文处理，我们将译文中出现“-year-old”的翻译成“岁”，将译文中出现“-year”的翻译成“年”。

4.5 实验结果

以下是我们这次评测的结果，评价指标均为 BLEU4-SBP。汉英新闻结果如下：

表 8 汉英新闻评测结果

评测系统	开发集	测试集 09
primary-a	**0.2559**	**0.2267**
contrast-b	0.2513	0.2256
contrast-c	0.2525	0.2213

其中 primary-a 为层次短语主系统。contrast-b 为短语对比系统，使用了经过过滤的训练语料 contrast-c 为短语系统 baseline。

英汉新闻结果如下：

表 9 英汉新闻评测结果

评测系统	开发集	测试集 09	测试集 11
primary-a	**0.3323**	0.3355	0.3253
contrast-b	0.3299	**0.3432**	**0.3296**
contrast-c	0.3305	0.3381	0.3270

其中 primary-a 为层次短语主系统。contrast-b 为短语对比系统，使用了经过过滤的训练语料。contrast-c 为短语系统 baseline。

藏汉政府文献结果如下：

表 10 藏汉政府文献评测结果

评测系统	开发集	测试集 09	测试集 11
primary-a	**0.5362**	0.5048	**0.2304**
contrast-b	0.5016	**0.5075**	**0.2304**
contrast-c	0.5280	0.5003	0.2293
contrast-d	0.5345	0.4675	0.2272

其中 primary-a 为层次短语主系统，使用计算所分词，对未登录词进行了一些替换。contrast-b 为层次短语系统 baseline，使用计算所分词。 contrast-c 为层次短语系统 baseline，使用实验室内部开发分词。contrast-d 为层次短语系统，使用实验室内部开发分词，对未登录词进行了一些替换。

4.6 实验结果分析

在英汉新闻评测中，我们提交的主系统比对比系统要高，这是预料之中的结果。

在英汉的新闻评测之中，我们的主系统在开发集上要比对比系统高，但是在测试集上却不如对比系统，有些出乎我们的预料。

在藏汉政府文献评测中，我们的主系统对未登录词进行了替换，在 09 年的测试集上比系统 contrast-b 要稍低，在 11 年与系统 contrast-b 相同。这个结果没有达到我们预期的效果，应该与我们对 BLEU-SBP 不熟悉而造成的。

5 总结

本次评测，由于时间和经验的问题，我们没能将一些改进运用在本次评测中，因此只取得了一个比较普通的成绩。

本次评测的主要不足有：

(1) 没有使用 BLEU-SBP 作为调参的指标

(2) 没有对模型进行改进

本次评测中我们使用了自己开发的解码器，并且对语料进行了许多处理，因此较 11 年评测结果有了很大的提高。

参考文献

[1] Philipp Koehn, Hieu Hoang, Alexandra Birch, et. al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation[C], *Annual Meeting of the As-sociation for Computational Linguistics*.

[2] David Chinag. 2007. Hierarchical phrase-based translation[J]. *Computational linguistics*.

[3] J.-C. Chappelier and M. Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG[C]. *In Proceedings of the First Workshop on Tabulation in Parsing and Deduction*, pages 133–137.

[4] Koehn, Philipp and Och, Franz Josef and Marcu, Daniel. 2003. Statistical phrase-based translation. *Proceedings of the 2003 Conference of the NAACL HLT*.

[5] Stolcke, Andreas and others . 2002. SRILM-an exten-sible language modeling toolkit [J]. *INTERSPEECH*.

[6] Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation [C]. *Proceedings of the 41st Annual Meeting on Association for Com-putational Linguistics*.

[7] Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation [C]. *Proceedings of the 40th annual meeting on association for computational linguistics*.