北京交通大学 CWMT2011 评测技术报告[1]

蒋俊杰　徐金安　张玉洁
北京交通大学计算机与信息技术学院　北京　100044
E-mail: {10120469, jaxu, yjzhang}@bjtu.edu.cn

摘 要：本文介绍了北京交通大学自然语言处理研究组(BJTU-NLP)参加 CWMT2011 评测的情况。本次评测，我们一共参加了英汉新闻、汉英新闻、英汉科技等三个项目的机器翻译评测任务。文章主要介绍了我们参加各个评测任务的系统框架、模型及评测结果。

关键字：基于短语翻译模型，基于层次短语翻译模型，Moses，CWMT2011

BJTU Technical Report for CWMT2011 Evaluation

Junjie Jiang, Jinan Xu and Yujie Zhang
School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 100044, China
E-mail: {10120469, jaxu, yjzhang}@bjtu.edu.cn

Abstract: *This paper presents the overview of statistical machine translation systems that BJTU-NLP developed for participating CWMT2011.We have taken part in three tasks: English-Chinese News, Chinese-English News, and English-Chinese Science. The paper briefly describes the primary modules, the key techniques, and the evaluation results.*

Keywords: *phrase-based translation model, hierarchical phrase-based model, Moses, CWMT2011*

1 引言

本文对北京交通大学自然语言处理研究组（BJTU-NLP）参与 CWMT2011 机器翻译评测情况进行描述。本次评测我们参与的三个项目包括英汉新闻领域机器翻译、汉英新闻领域机器翻译、英汉科技领域机器翻译。其中，在英汉新闻领域和汉英新闻领域机器翻译评测项目中使用了基于短语的统计机器翻译系统，在英汉科技领域机器翻译评测项目中使用了基于层次短语的统计机器翻译系统。

本文第 2 节简单介绍了参评系统，第 3 节介绍实验与评测结果，第 4 节进行总结。

2 参评系统描述

BJTU-NLP 在本次评测中利用开源工具 Moses 实现了一个基于短语的统计机器翻译系统和一个基于层次短语的统计机器翻译系统。

在参与的三个评测项目中，BJTU-NLP 针对英汉新闻领域机器翻译评测任务和汉英新闻领域机器翻译评测任务，使用了基于短语的统计机器翻译系统。同时，针对英汉科技领域机器翻译评测任务，使用了基于层次短语的统计机器翻译系统。

下面我们将对各个系统进行简要介绍。

[1]基金项目：北京交通大学人才基金(2011RC034)；中央高校基本科研业务费专项资金(2009JBM027)；北京市重点学科共建项目（计算机应用技术）；中科院计算技术研究所智能信息处理重点实验室开放课题(IIP2010-4)

2.1 基于短语的统计机器翻译模型

本次评测中我们使用开源工具 Moses[2][Koehn et al., 2007]构建了基于短语的统计机器翻译系统。基于短语的统计机器翻译[Koehn et al., 2003]通过对数线性模型将句子的得分描述为若干特征的线性组合，如公式（1）所示。

$$\hat{e} = \arg\max_{e} \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e,f))}{\sum_{e'} \exp(\sum_{m=1}^{M} \lambda_m h_m(e',f))} \quad (1)$$

其中 e 为目标语言的句子，f 为源语言的句子，$h_m(e,f)$ 表示第 m 个特征函数，λ_m 表示第 m 个特征函数所对应的权重。解码算法采用逆向递归柱搜索 Beam-Search 进行单调搜索获取翻译结果 $\hat{e}$。

系统采用的基本特征为包括：正反向短语翻译概率、正反向词汇翻译概率[Koehn et al., 2004]、短语惩罚、基于距离的调序惩罚、词惩罚、语言模型以及 MSD(monotone, swap, discontinuous)词汇化调序模型特征。

系统利用 GIZA++[Och and Ney, 2003]训练词对齐模型，并抽取翻译短语对。对数线性模型的参数是利用最小错误率训练（MERT）方法[Och et al., 2003]通过在开发集上对 BLEU4 进行优化获得的。

2.2 基于层次短语的统计机器翻译模型

层次短语模型可以被认为是对基于短语模型的扩展，它可以抽取源语言句子中非连续的部分，并将其翻译成目标语言句子的非连续部分。基于层次短语的统计翻译系统是一个形式化语法的翻译系统，采用同步上下文无关文法(SCFG)建立翻译模型，其规则形式如公式(2)所示：

$$X \rightarrow <\gamma, \alpha, \sim> \quad (2)$$

其中，X 为非终结符，γ 和 α 为源语言端和目标语言端由终结符和非终结符组成的字符串，~为 γ 和 α 中非终结符的一一对应关系。

层次化短语模型使用了短语规则，与基于短语的方法类似，能够将连续的源语言词串翻译成为目标语言的词串；同时，还使用了层次化规则引入变量，能够实现短语调序功能。

层次化短语模型的翻译通常被看作是一个不断使用规则的推导过程。翻译模型同样采用对数线性模型。系统借鉴[Chiang, 2005]的方法，使用 7 个特征函数：翻译概率 $P(\alpha|\gamma)$ 和 $P(\gamma|\alpha)$，词汇化权重 $P_w(\alpha|\gamma)$ 和 $P_w(\gamma|\alpha)$，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大的那个推导生成翻译结果。

本次评测中我们使用 Moses 实现基于层次短语的统计机器翻译系统，利用 GIZA++训练词对齐模型，采用最小错误率训练方法得到优化的模型参数。

3 实验

3.1 系统硬件配置

在本次评测中使用的计算机配置与操作系统如表 1 所示。

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

表 1. 机器硬件配置与操作系统

CPU	内存	操作系统
Intel Xeon 2.40GHZ 四核	4G	Ubuntu desktop 10.04

下面我们对训练及测试过程进行描述。

3.2 数据处理方法及工具

对中文数据进行的预处理包括：常见 html 转义字符转化为对应的特殊字符，全角变半角，英文标点转为对应的中文标点，分词。对英文数据进行的预处理包括：常见 html 转义字符转化为对应的特殊字符，全角变半角，中文标点转为对应的英文标点，标点符号的分离，大写转小写。对语言模型训练数据去噪去重。对翻译模型训练数据去噪去重，去除长度大于 100 的句对。

对于中文的后处理主要是合并空格，对于英文的后处理是字母小写转大写和标点符号的合并。

中文分词使用了 ICTCLAS[3]，英文分词、小写、标点合并使用了 WMT08 Shared Task[4]提供的 tokenizer.perl、lowercase.perl 和 detokenizer.perl。英文的小写转大写使用 Moses 训练的 recase 模型。

3.3 数据使用

我们此次评测使用的语料全部是评测方提供的训练数据。在英汉新闻领域机器翻译中，汉语的语言模型训练数据为全部英汉新闻训练数据的汉语部分与搜狗全网新闻语料的合并数据。在汉英新闻领域机器翻译中，英语的语言模型训练数据为全部汉英新闻训练数据的英语部分与路透社 RCV1 新闻语料的合并数据。本次评测中所使用的开发集是评测方提供的开发集。预处理后的训练数据如表 2 所示。

表 2. 评测系统使用的预处理后的数据

评测项目	翻译模型训练数据	语言模型训练数据	开发集
英汉新闻领域	5,046,074 句对	12,732,503 句	1000 句，四个参考译文
汉英新闻领域	5,046,074 句对	11,554,133 句	1006 句，四个参考译文
英汉科技领域	892,438 句对	901,169 句	1116 句，四个参考译文

3.4 语言模型

在英汉科技领域机器翻译项目中语言模型训练数据为评测方提供的全部英汉科技训练数据的汉语部分，预处理后使用 SRILM[5][Stolcke, 2002]工具训练插值 Kneser-Ney 平滑[Chen and Goodman, 1998]的 5 元模型，具体的训练命令为：

ngram-count -order 5 -unk -interpolate -kndiscount -text train.txt -lm LM

在英汉新闻领域机器翻译项目中语言模型训练数据为评测方提供的全部英汉新闻训练数据的汉语部分与搜狗全网新闻语料合并后的数据，在汉英新闻领域机器翻译中语言模型训练数据为评测方提供的全部汉英新闻训练数据的英语部分与路透社 RCV1 新闻语料合并后的数据，预处理后使用 IRSTLM[6][Federico et al., 2008]工具训练 improved-kneser-ney 平滑的 5 元模型，具体的训练命令为：

build-lm.sh -n 5 -k 10 -s improved-kneser-ney -p -i train.txt -o train.ilm.gz

compile-lm train.ilm.gz --text yes LM

3.5 翻译模型

本次三个评测项目都使用了基于短语的统计机器翻译模型，其中英汉科技领域中还使用了基于层次短语的统计机器翻译模型。

---

[3] http://ictclas.org/
[4] http://www.statmt.org/wmt08/shared-task.html
[5] http://www.speech.sri.com/projects/srilm
[6] http://hlt.fbk.eu/en/irstlm

（1）基于短语的统计机器翻译模型的训练主要使用如下配置：

使用 GIZA++进行词对齐。双语语料完成双向词对齐后，采用启发式规则“grow-diag-final-and”[Koehn et al. 2003]合并两个方向的词对齐结果，作为最终的双语词对齐结果。

短语抽取，短语长度限制为 10。

使用基于 msd-bidirectional-fe 词汇化重排序模型。

具体地，使用 Moses 的训练脚本分步训练，训练命令为：

train-model.perl -root-dir . --corpus corpus/train -f fr -e en -alignment grow-diag-final-and -reordering msd-bidirectional-fe -parts 3 --parallel

（2）基于层次短语的统计机器翻译模型的训练主要使用如下配置：

使用 GIZA++进行词对齐，采用启发式规则“grow-diag-final-and”。

具体地，同样使用 Moses 的训练脚本分步训练，训练命令为：

train-model.perl -root-dir . --corpus corpus/train -f fr -e en -alignment grow-diag-final-and -hierarchical -glue-grammar -parts 3 --parallel

（3）此外，在汉英新闻领域机器翻译任务中，还训练了大小写转换模型，训练命令为：

train-recaser.perl --dir recaser --corpus corpus/tok.en -ngram-count $SRILM/bin/i686/ngram-count -train-script train-model.perl

3.6 实验结果与分析

3.6.1 英汉新闻领域机器翻译评测结果与分析

在对开发集进行最小错误率训练时，我们使用了两种目标语言模型进行对比：1）以全部英汉新闻语料的汉语部分作为训练集得到语言模型 Origin.5lm；2）以全部英汉新闻语料的汉语部分与搜狗全网新闻语料合并后作为训练集得到语言模型 Union.5lm。在开发集上的实验结果如表 3 所示：

表 3. 不同语言模型在开发集上的评分对比（英汉新闻）

语言模型	Bleu-4（基于词)
Origin.5lm	0.2386
Union.5lm	0.2467

从表 3 中看出，在加大语言模型规模后，开发集的 bleu 值上升了近 1 个百分点。因此在最终解码测试集时，我们选择语言模型 Union.5lm。最终的评测结果如表 4 所示：

表 4. 英汉新闻领域评测结果

测试集（基于字）	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
en_zh_news_progress	0.3296	0.3449	0.2791	9.6383	9.6475	0.7928	0.6418	0.3777	0.4253
en_zh_news_current	0.3146	0.3257	0.2648	9.2578	9.2682	0.7795	0.627	0.3792	0.4029

3.6.2 汉英新闻领域机器翻译评测结果与分析

类似地，在对开发集进行最小错误率训练时，使用两种目标语言模型进行对比：1）以全部汉英新闻语料的英语部分作为训练集得到的语言模型 Origin.5lm；2）以全部汉英新闻语料的英语部分与路透社 RCV1 新闻语料合并后作为训练集得到的语言模型 Union.5lm。在开发集上的实验结果如表 5 所示。从表 5 中看出，在加大语言模型规模后，开发集的 bleu 值也有所上升。因此在最终解码测试集时，我们选择语言模型 Union.5lm。

表 5. 不同语言模型在开发集上的评分对比（汉英新闻）

语言模型	Bleu-4(大小写不敏感)

Origin.5lm	0.2674
Union.5lm	0.2704

另外，我们观察到部分汉英平行语料库的训练数据集存在句对齐级别噪音问题。针对这个问题，我们在训练翻译模型时去除了部分噪音的干扰，训练了另一个系统，称为对比系统(constrast)。语言模型同样使用 Union.5lm。最终的评测结果如表 6 所示。

表 6. 汉英新闻领域评测结果

测试集（大小写敏感）	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT
zh_en_news_primary	0.2117	0.2288	7.6006	0.7095	0.7272	0.5094	0.3161
zh_en_news_contrast	0.2157	0.2325	7.5786	0.7068	0.7192	0.5066	0.322

从表 6 中看出，对比系统取得了更高的评分。这说明数据的噪音会在对齐处理中造成干扰，从而影响最终的翻译性能。由此可以得出结论，语料的去噪是一个很重要的步骤。

3.6.3 英汉科技领域机器翻译评测结果与分析

在英汉科技领域中我们除训练基于短语的翻译模型外，还训练了基于层次短语的翻译模型。我们使用全部英汉科技语料的汉语部分建立了相同的语言模型。开发集上的实验结果如表 7 所示。

表 7. 不同系统在开发集上的评分对比（英汉科技）

系统	Bleu-4(基于词)
基于短语的翻译模型	0.3724
基于层次短语的翻译模型	0.3829

从表 7 中看出，与基于短语的翻译模型比较，基于层次短语的翻译模型在开发集上的评分提高了 1 个百分点。在最终解码测试集时，我们同时使用了两个模型，最终的评测结果如表 8 所示。

表 8. 英汉科技领域评测结果

测试集（基于字）	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
en_zh_scie_phraseBased	0.3672	0.3848	0.3165	10.266	10.287	0.8354	0.6154	0.308	0.3958
en_zh_scie_hierarchical	0.3683	0.3819	0.3147	10.261	10.283	0.8365	0.6205	0.3118	0.3806

4 总结

本文主要介绍了北京交通大学计算机与信息技术学院计算机科学与技术系自然语言处理研究组 BJTU-NLP 参加 CWMT2011 评测的情况。这是 BJTU-NLP 小组首次参加全国机器翻译研讨会组织的评测。我们使用开源翻译工具 Moses 实现了基于短语、基于层次短语的翻译系统。在进行训练的过程中，加强了训练语料的预处理，降低了噪声干扰，提高了训练语料的质量，实验结果表明，我们的系统相对于基线系统有了一定的提高。

今后，在统计机器翻译的预处理和后处理、基于句法的翻译模型等方面，我们将进行深入的研究和探讨。

参考文献

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing, volume 2*, Pages 901-904.

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

Franz Joseph Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, July.

Koehn, Philipp, Franz J. Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*

M. Federico, N. Bertoldi, M. Cettolo. 2008 IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models, In *Proceedings of Interspeech, Brisbane*, Australia.

Philipp Koehn, Franz Josef Och, et al. 2003. Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, Edmonton, Alberta, Canada.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models [A]. In *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas*[C]. Pages 115-124.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL 2007 (demonstration session)*.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. *Technical Report TR-10-98*, Harvard University.